Speech Emotion Recognition: Clustering and Deep Learning Approach to Detect Conflicting Emotions through Vocal Expressions

by

Nuhash Kabir Neeha
21301025
Nuzhat Rahman
21301538
Imtela Islam
21341018

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
June 2025.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Nuhash Kabir Neeha

21301025

Nuzhat Rahman

21301538

Imtela Islam

21341018

Approval

The thesis/project titled "Speech Emotion Recognition: Clustering and Deep Learning Approach to Detect Conflicting Emotions through Vocal Expressions" submitted by

1. Nuhash Kabir Neeha (21301025)
2. Nuzhat Rahman (21301538)
3. Imtela Islam (21341018)

of Summer 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Summer 2025.

Examining Committee:

Supervisor:
(Member)

Dibyo Fabian Dofadar

Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Deep models have improved speech emotion recognition, yet overlapping emotions and speaker variability still limit practical deployment. The guiding hypothesis of this paper is that a useful solution must clearly tackle overlapping and blended emotions, along with enhancing classification strength. These issues are addressed with a multi-branch ensemble approach and enhanced with a fuzzy post-processing phase aimed at revealing secondary emotion memberships and emphasizing unclear, overlapping statements; speaker-adversarial training was added as a supportive strategy to minimize confounding factors. evaluation on augmented benchmark corpora demonstrates competitive accuracy and improved handling of label overlap. This dual emphasis on accuracy and interpretability provides a principled way to surface uncertainty in emotional input and aims to provide a foundation for SER systems to be more robust for real-world use.

Keywords: Speech Emotion Recognition, Sound Emotion Recognition, Natural Language Processing, Conflicting Emotions, Deep Learning Models, Clustering, Emotional Cues, Natural Interactions, Machine Responsiveness, Vocal Expressions.

Acknowledgement

We would like to express our sincere gratitude to our supervisor, Mr. Dibyo Fabian Dofadar, our supervisor, for his invaluable guidance, patience, and continuous support throughout this research. His insights and encouragement have been essential to the completion of our work.

We also extend our appreciation to Dr. Md. Golam Rabiul Alam, Professor and Thesis Coordinator, and Dr. Sadia Hamid Kazi, Chairperson of the Department, for their academic supervision and for providing a supportive research environment.

Finally, we are deeply thankful to our peers, families, and friends for their encouragement, understanding, and moral support during this journey.

Table of Contents

List of Figures

Chapter 1

Introduction

Understanding and responding to human emotions from speech remains a core challenge for effective computing. As emotions are fundamental to how humans relate and communicate, they are useful across various domains. For example, it can help to design better human-computer interactions such as in forensic settings, with behavioral analysis or lie detections. A critical aspect of these interactions depends on machines' ability to understand human emotions and respond to it in real time. Sound Emotion Recognition (SER) can be used to bridge the emotional gap between humans and machines. SER enables systems to perceive and interpret emotions based on vocal expressions. While the existing SER models have been progressing significantly over time; there are areas in this field that are yet to be explored fully. Many models may struggle with the dynamic nature of human emotions, as one may be experiencing multiple emotions at the same time (i.e. fear and anxiety, fear and excitement, deep sadness and anxiousness and so on). These real-life overlapping emotions make it challenging for the existing SER models to recognize and accurately classify emotions, which in turn leads to reduced model performance. This study aims to face these challenges by employing advanced clustering and deep learning techniques and models to accurately catch the nuanced emotions one might be experiencing through speech. This study will contribute in creating a more natural and emotionally intelligent interaction between humans and automated systems, by enabling machines to interpret such intricate emotions. Consequently, making these systems more accurate, sensitive and precise.

1.1 Background

Emotions are complex psychological and physiological states triggered by different stimuli. This section is concerned with defining what emotions are, what models are used to understand emotions and human expression of emotions.

1.1.1 Emotions, Moods, Feelings

Speech is one of the primary ways of conveying information, while emotions are fundamental human expressions of different experiences. Although closely related, the concepts of emotion, feeling, mood are distinct.

According to a study[19] emotions are intense, short-lived reactions triggered by internal or external stimuli (physiological and psychological). It does not occur consciously. In contrast, feelings are subjective experiences of an emotion. It is the interpretation of emotional responses. It occurs consciously. Lastly, moods are sustained emotional states with no specific triggers.

1.1.2 Emotional Models

Emotions can be categorised into 2 popular approaches. Firstly, the categorical (discrete) approach describes emotion as a discrete number of classes. One of the most popular theorists[2], listed 6 basic emotions: anger, happiness, surprise, sadness, fear and disgust. These are universal regardless of one's cultural and social background. Secondly, in Russel's 2D Model, the dimensional (continuous) approach describes emotions on a continuous scale of valence, which can be positive or negative, and arousal, which can be high or low. PAD (Russel's 3D Model, stands for Pleasure, Arousal, Dominance) is another widely used emotional state model, consisting of 3 dimensions: arousal, pleasure, dominance[1]. Discrete model distinctly categorises emotions giving them their own set of cognitive and psychological elements, whereas the dimensional model recognises the complexity of emotional experiences, influenced by many factors such as personal history, cultural background and so on. It uses 2D (arousal and valence) and 3D (arousal, valence and dominance) emotional space models.

1.1.3 Expression of Emotions

Humans use different channels to express emotions. According to another study[6], facial expressions can communicate multiple emotions without the need for verbalizing. Some of the universal nonverbal emotions expressed through facial features are anger, surprise, fear etc. In addition to facial expressions, vocal cues—including pitch, amplitude, and frequency of sound waves—play a crucial role in identifying an individual's emotional state, forming the foundation of SER. SER uses speech to recognize emotions; however, body language, consisting of gestures, postures, eyemovent, hand placement or movement further enhances the communication of emotions. Likewise, physiological signals such as skin conductance, electrocardiogram (ECG), blood volume pulse (BVP), heart rate and so on, can be used as modes of expression of emotions for individuals who suffer from mental or physical illnesses; where they are unable to communicate efficiently through facial expression, vocal cues or body language[22]. Figure 1.1 shows a breakdown of this.

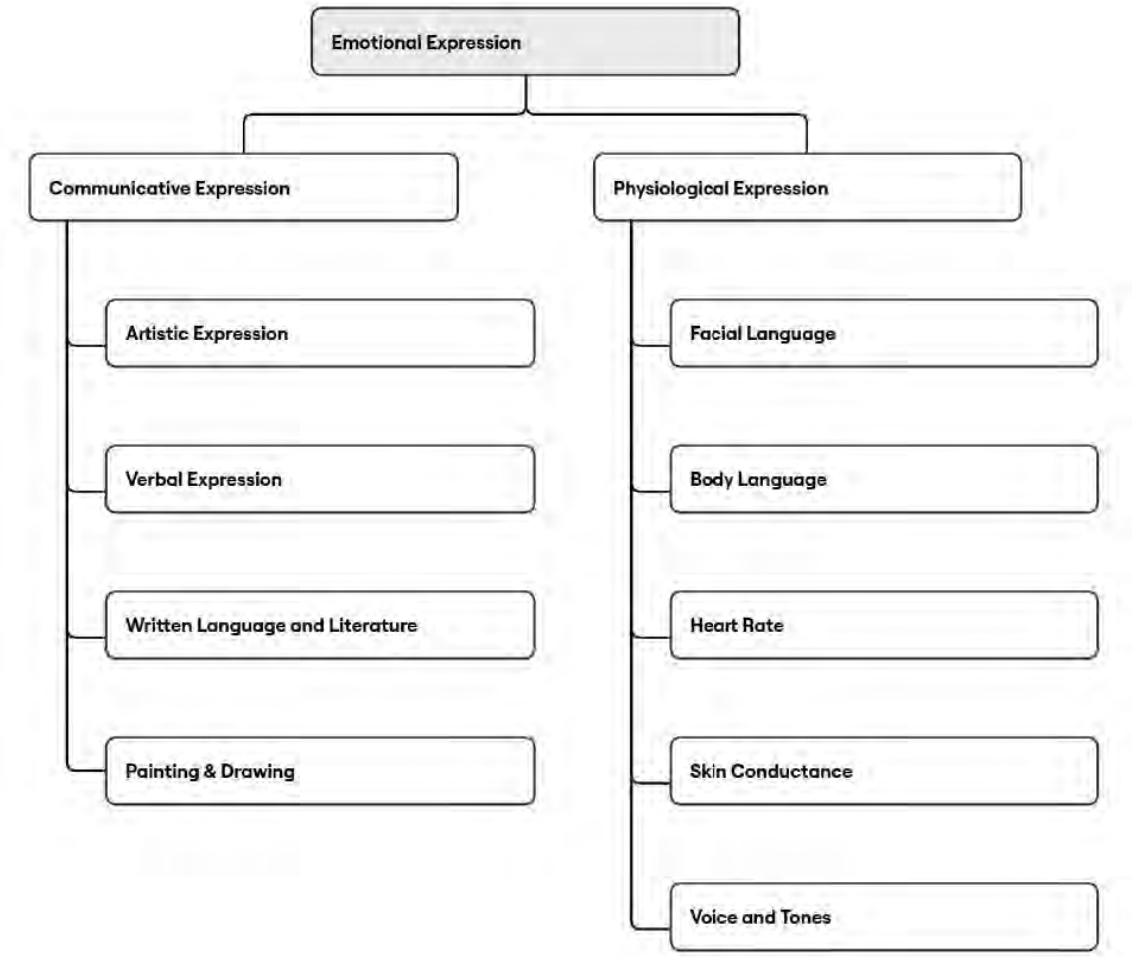

Figure 1.1: Human Emotional Expression

1.1.4 Emotion Recognition

There are two types of emotion recognition: non-automated recognition and automated recognition. Non-automated recognition refers to the natural human ability to perceive and interpret emotions. Researchers have developed structured techniques to assess emotions such as FACS (Facial Action Coding System), Geneva Emotion Wheel, SAM (Self-Assessment Manikin) along with psychological assessment questionnaires like PANAS (Positive and Negative Affect Schedule), BDI (Beck Depression Inventory), STAI (State-Trait Anxiety Inventory) and POMS (Profile of Mood States). In contrast, Automated recognition recognises emotions and interprets them with the help of machines. These systems can further be divided into 3 modes: unimodal, bimodal and multimodal. Unimodal takes into account 1 mode of expression, whereas Bimodal uses 2 modes of expression and Multimodal uses multiple modes of expression.[22]

1.2 Rational of the Study or Motivation

As technology advances and gets integrated into daily life, it becomes crucial for machines and humans to have natural and smooth communication, to improve user satisfaction, and ensure more intuitive and interactive computer interactions. Even though SER systems are high in demand and are useful in many aspects they still lack the ability to differentiate between the conflicting or overlapping emotions. Existing SER systems may face challenges classifying between frustration and determination or anxiety and excitement. A more adaptive and nuanced recognition technique is required to face said challenges.

1.3 Problem Statement

SER models can be indispensable in efforts to make emotionally intelligent machines. However, traditional SER models struggle to identify conflicting and overlapping emotions, as most of the systems classify emotions into discrete categories such as happy, sad, angry. The models assume emotions to be static and independent. Although, in reality emotions are interdependent, dynamic, complex and multiple emotions can often be expressed simultaneously. For example, individuals may experience feeling both anxious and excited at the same time. Existing SER models, especially the models dependent on manually extracted features such as MFCCs, spectral features and so on fail to capture such emotional nuances, leading to misclassification and reduced performance. This research aims to use advanced clustering techniques in addition to deep learning models to identify said emotional nuances and create a model for better emotion recognition of conflicting emotions.

1.4 Objective

This study aims to detect and evaluate emotional conflicts within speech, focusing on identifying instances where multiple or overlapping emotions coexist in a single utterance. The objective is to analyze and quantify the degree of emotional conflict expressed by a speaker. The key goals include examining the limitations of traditional single-label SER models, developing a framework capable of recognizing co-occurring emotional states, and evaluating the extent of their overlap through fuzzy clustering techniques. Ultimately, this study seeks to advance emotionally aware speech systems that can interpret complex emotional expressions more accurately, enhancing the depth and authenticity of human-computer interaction.

1.5 Methodology in Brief

This research paper was based on reading many technical research articles and journals related to sound/speech emotion recognition. Firstly, the research focused on understanding the complexity of emotions. Also, how emotional nuances can be expressed using different channels. Moreover, different modes of emotional recognition. Secondly, review papers were analysed to look for research gaps and key challenges faced by currently existing SER systems. Lastly, different models were compared

and contrasted to see which models performed better and what were their consequent drawbacks. The papers, articles and so on that were read for this study were chosen based on relevance, large number of citations and more recent publication. These strategically selected papers will help understand the recent developments and applications of SER. This will lay the foundation for future research and model development.

1.6 Work Plan

The work plan for this thesis paper has been divided into 3 distinct parts: pre-thesis 1, pre-thesis 2 and defense. In pre-thesis 1, a research gap is discovered. The problem is further researched and defined undertaking a comprehensive study of the problem. This paper comprises literature review which helps build a solid theoretical foundation in order to understand the problem and research objectives. Furthermore, pre-thesis 2 encompasses data collection, feature extraction and development of basic SER model in order to solve the challenge of detecting conflicting or overlapping emotions. Lastly, defense will involve the advancement and further development of the hybridized model made using clustering and deep learning models. The model will be evaluated depending on the suitable evaluation metric and compared with state of the art models.

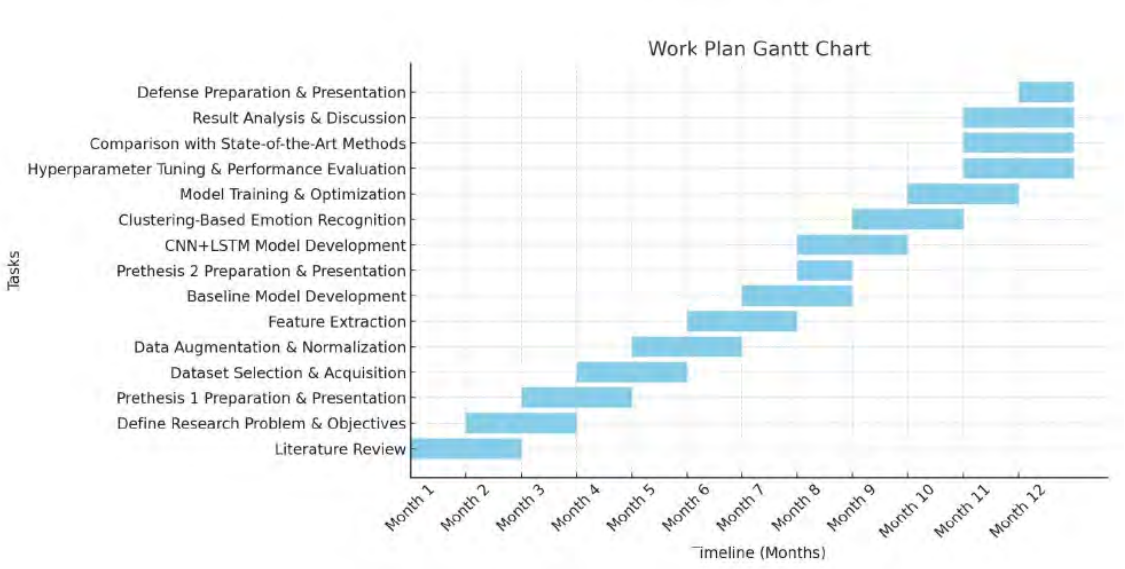

Figure 1.2: GANTT chart of work plan

Chapter 2

Literature Review

2.1 Sound/ Speech Emotion Recognition

A review on SER[20] states that it is a field of study that deduces human emotions from verbal expression. The system focuses on identifying and classifying voice input to various defined categories of emotion. It identifies acoustic and linguistic cues from speech to infer emotional states, which are useful for applications like behavioral analysis, clinical assessment, and other domains that rely on affective information.

Definition of SER

It[21] is stated that speech emotion recognition is a part of speech processing which analyses speech signals and recognises patterns of speech such as prosody, pitch, frequency, rhythm in order to determine the emotional state of the speaker. Classifying emotions using sound/speech of an individual has various applications, majorly in human-computer interaction.

Relevance of Using SER

With growing reliance on technology, especially AI-driven services and automated systems, there is an increasing need for emotionally aware computational tools to improve user satisfaction and efficacy. Speech Emotion Recognition (SER) can provide that capability and has practical applications across many domains, including emergency hotlines, virtual assistants, clinical and forensic assessments, and criminal-investigation support.

2.2 Evolution of SER

2.2.1 Early Research and Traditional Methods

The research on sound emotion recognition had begun by using acoustic features and statistical classification models. Initially, models like support vector machine (SVM), hidden markov model (HMM) were used along with prosodic and spectral features (pitch, energy, mel-frequency cepstral coefficients). Although these methods

did provide insights into emotion classification, they struggled to accurately identify the emotions in the presence of noise, speakers of different languages as well as in the cases of conflicting emotions.

2.2.2 Transition to Machine Learning

Due to the struggles faced by the traditional methods, Machine Learning (ML) approaches were introduced to SER. ML improved SER by allowing automated feature selection and classification. Models such as Random Forests and Gaussian Mixture Models (GMMs) demonstrated higher adaptability and accuracy. Even after these improvements, automatic feature selection is not completely reliable as it is highly dependent on feature engineering quality as well as dataset variability. As such, manually extracted features remain as a limitation.

2.2.3 Adoption of Deep Learning

The adoption of deep learning (DL) improved on the limitations previously faced by the earlier methods by eliminating the need for handcrafted features. Models such as Convolutional Neural Networks (CNNs) effectively extracted spatial patterns from spectrograms, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, modeled temporal dependencies in speech. Hybrid CNN-LSTM architectures improved the recognition accuracy as such improving generalization to real-world emotional speech. Compared to ML-based methods, DL models demonstrated a higher level of robustness and scalability, even though they required extensive data and computational resources.

2.2.4 Role of Benchmark Datasets

The availability of standardized datasets has been crucial in advancing SER. Corpora such as RAVDESS, IEMOCAP, Emo-DB, CREMA-D, SAVEE and TESS provide diverse linguistic and emotional variations, facilitating the development of more adaptable models. These datasets have enabled direct comparisons between ML and DL approaches, demonstrating the effectiveness of deep learning in handling complex emotional variations. SER has evolved from statistical methods to deep learning-driven approaches, significantly improving emotion recognition accuracy. However, challenges related to cross-corpus generalization, speaker variability, and computational efficiency remain key areas for further research.

2.3 Taxonomy of Methods in SER

The taxonomy in SER can be shown as follows:

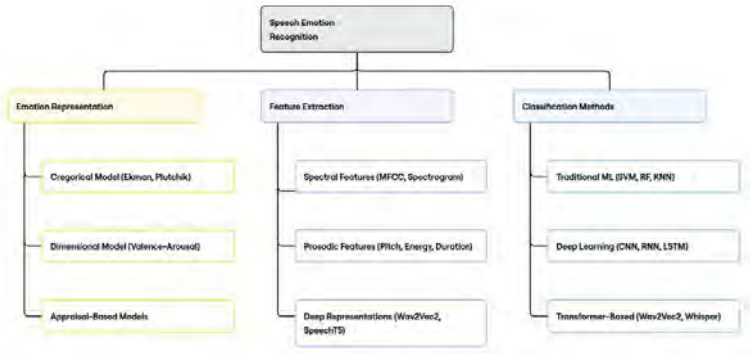

Figure 2.1: Flow chart for SER taxonomy

2.4 Key Challenges in SER

The key challenges faced by SER systems, according to some studies[3], [4], [5], [6], [7], [18], were as follows:

1. **Dataset Limitations:** almost all the articles that were analysed in order to write this paper used the similar datasets available. This can be addressed by data augmentation. A small dataset can lead to overfitting and limited generalizability.
2. **Noise and Distortions in Speech Data:** background noise and distortions reduce the quality of the data and introduce added complexities.
3. **Audio Feature Extractions:** extracting relevant audio features is crucial for effective SER
4. **Class Similarity and Misclassification:** overlapping emotions can cause 2 different emotional vocal cues to sound similar leading to an increased risk of misclassification.
5. **Scalability and Real-time Processing:** implementing operational SER systems that can process audio information remains highly challenging.
6. **Linguistic Variations:** Impact of cultural, personal, and contextual factors affect the differences between emotional tone and expression.
7. **Generalization:** Models performing poorly on unseen datasets or real-world scenarios
8. **Model Selection and performance:** different ML, DL models have different strengths and weaknesses. Model interpretability can also be a limiting factor.

2.5 Application of SER

According to multiple studies[4], [6], there are multiple applications for SER.
Practical applications:

1. **HCI:** SER enhances human computer interaction by allowing machines to identify and interpret human emotions for more natural communication.

2. **Healthcare:** SER can help with mental health assessments by using it during therapy session to monitor their emotional response
3. **Entertainment:** if SER is used in gaming or interactive media, understanding player emotions can lead to a higher user satisfaction.

Academic applications:

1. **Research & Development:** studying and using SER contributes in the field of interactive computing and emotional AI. Consequently, advancing machine learning, signal processing and cognitive psychology.
2. **Emotion Theory:** exploration of SER helps improve the existing theory on emotions, which in turn advances studies on psychological and linguistic studies depending on emotional states through vocal expression.

Societal impact:

1. **Communication improvement:** SER can improve social interactions by recognizing emotions.
2. **Educational enhancement:** SER can further help teachers understand students' emotional state and create a work plan that will best enhance their performance.
3. **Inclusion and accessibility:** SER can be used to aid individuals with disabilities by providing them with a smoother and better communication through emotionally aware systems.

SER has become very popular due to its multiple applications. However, many SER systems face challenges to accurately identify and interpret emotions due to conflicting and overlapping emotions at a time, limiting their effectiveness. This complexity makes it difficult for traditional categorical emotion classification making it insufficient and thus requiring more adaptive and nuanced recognition techniques.

2.6 Existing Model Architecture

One of the studies[18] uses a one-dimensional convolution network (Conv1D). It is a deep learning technique that processes audio features to learn the complex patterns using convolution layers. It is a good model for capturing local, temporal dependencies in audio signals. Additionally, they also use random forest (RF). It is a tree-based machine learning algorithm that works by constructing multiple decision trees. Using it alongside feature pruning increases performance, provides dimensionality reduction, and reduces computational cost. The comparative analysis of Conv1D and RF provides insight on the quality of performance of the two models. Consequently, becoming the highlight of the study.

In comparison, another study[5] uses an ASR model at its core for audio-to-text mappings, utilizing dilated convolution and gated convolutional units. Also, in this study the authors use the ASR model as a feature extractor for emotion recognition tasks. This is done by computing mean activations for different layers. Furthermore, a linear regression model is used to predict the arousal and valence values from the

extracted features. This approach made their study novel by utilizing ASR as feature extractor, exploration of layer contributions, and end-to-end training considerations.

Conversely, another[6] uses deep learning models like RNN (Recurrent Neural Network) and LSTM (Long Short-Time Memory). RNN works by keeping a record of old information which makes it suitable for capturing temporal information or dependencies of sequential data such as speech signals. LSTMs are extended versions of RNN. LSTM allows for long-term dependency learning. In this study, the authors also used SVM (Support Vector Machine) for classification. Even though, SVM works with highly dimensional data it is a more traditional model used for SER. In addition, CNN (Convolution Neural Network) can also be a good model to use when using automated feature extraction, i.e. spectrograms or mel-spectrograms. Spectrograms are computed using STFT (Short-Time Fourier Transform) and mel-spectrograms are computed by applying mel filter banks on spectrograms. Lastly, it was discovered that using a hybrid model enhances model performance. This hybrid approach introduces novelty in their study. Hybrid models can be created by integrating attention mechanisms in RNNs and LSTMs; consequently, aiding the model focus on emotionally rich parts of speech signal. It was also discovered that multimodal models classified emotions more accurately than unimodal or bimodal models as they considered multiple modes of expression

Similarly, a different source[7] also reviews various deep learning techniques, including Deep Belief Networks (DBNs), and Deep Boltzmann Machines (DBMs) along with more commonly used Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks. CNNs extracts spatial features directly from input data through convolutional layers. These deep learning techniques leverage automated feature extraction, which in turn helps RNNs and LSTMs identify temporal dependencies in speech signals. Consequently, enhances long-term dependency recognition. They introduce novelty through hybridization of multiple models. The implementation of their model was executed using Python, leveraging libraries such as TensorFlow, Keras, and Librosa, among others.

In contrast, another research[3] used Deep Neural Network (DNN) and Stochastic Gradient Descent (SGD). DNN uses a combination of convolution, pooling and fully connected layers; whereas, SGD is an optimization approach. This leads to a more efficient training and accurate classification, updating the model parameter in order to reduce the loss function. The use of automated feature extraction is the highlight of this study. To accomplish this they enhanced temporal resolution by using segmented 20 ms frames and implemented the model using python leveraging libraries such as Keras and Theano.

Subsequently, another study[4] also used a DNN. In contrast, the author hybridizes DNN (Deep Neural Networks) with KNN (K-Nearest Neighbor) rather than SGD (Stochastic Gradient Descent). As mentioned before, DNN uses convolution and multiple fully connected layers in order to solve complex problems. Whereas, KNN

is a machine learning algorithm that classifies data depending on the distance between the data points of the number of neighbours chosen. KNN was used to improve accuracy, also, as a secondary validation method. Hybridization of ML and DL, where incorporating statistical features enhanced the performance of the model, and made it stand out amongst all the others.

On the other hand, a different study[8] used CNN and LSTM. They used 3 convolution layers for the CNN model for spectrograms and mel-spectrograms, with different feature maps and dropout layers each, and softmax function was used for output. Furthermore, they used a variant of LSTM, Bi-LSTM, where the layers are bidirectional and fully connected, including the softmax layers. This BLSTM model uses feature vectors for input and trains the models using sequence-to-one modelling. They have also used an ensemble of the two models where they used CNN followed by BLSTM, in order to automate feature extraction using CNN and feeding it directly to BLSTM. This is done to reduce the inefficiency of modelling the complexities of emotional nuances in a single utterance. Here, the CNN is used to extract high-level features, whereas the BLSTM is used to model the utterance-level long term dependencies subsequently. Finally, to compare and classify the models, a cross-entropy loss function was used along with the Adam optimiser to ensure the model is not too computationally expensive.

In a subsequent paper[11], supervised models - Regression and Artificial Neural Networks (ANN) - were used on the IADS dataset. The similarity of features in this dataset are compared using Pearson's r correlation coefficient, which were then used as inputs in the regression and ANN models. The emotional dimension of valence and arousal were predicted using the output from the following models. For regression analysis, RMSE was used, where the lower the RMSE value, the better the performance of the model. To ensure unbiasedness, principal component analysis (PCA) was used on data with and without dimension reduction in order to use five and ten fold cross validation. In contrast, they also used a two-layer feed forward ANN , consisting of one hidden layer. Additionally, the arousal and valence for this model were determined using two separate ANNs. The training-validation-testing ratio for the data set used was 70-15-15.

Furthermore in a differnt study[10], the authors study different models of SER including traditional as well as the ones created in recent times. Different feature extractors, for example, openSmile, Voice Quality Features, MFCCs, Pulse Code Modulation (PCM), Prosodic Features and more traditional Statistical Methods were also explored. And it was stated that the type of extractor used depends on the specific context it is to be used in. The feature extracted using these methods are used as inputs into models such as Hidden Markov Models (HMM) - which is a more traditional model , SVMs, ANNs, CNNs, RNNs - especially LSTM networks, and Generative Adversarial Networks (GANs) and Variational Autoencoders (VAE) - which are more advanced architectures used in the more recent years. HMM is a statistical model which consists of a finite set of hidden states, transitions (probabilities of moving from one state to another), emissions (probabilities of an observable output), and initial probabilities (the starting probabilities in each of the hidden

states). On the contrary, GANs and VAEs are both generative deep learning models. GAN works by training two neural networks (generator and discriminator) at the same time in a competitive manner, and VAEs consist of an encoder and a decoder -which are trained to minimise the difference between the input and the decoded output data.

2.7 Model Comparison

One of the studies[6] employed 2 models RNN and SVM and compared it against MLR. The features extracted were MFCC and MS (Modulation Spectral). Feature selection (FS) was used to find the most relevant feature subset. It was observed that the RNN model when run on the Spanish database obtained an accuracy of 94%, performing better than SVM or MLR. It was also observed that employing speaker normalisation improved model performance when the model was run on the Berlin database but had insignificant effect when run on the Spanish database. Moreover, feature selection enhanced performance by decreasing dimensionality. On the other hand, another[5] explored models like ASR model with transfer learning and Linear Regression for emotion estimation. The ASR system was used as a features extractor which in turn feeds the features into the linear regression model that estimates the valence-arousal values. Since, ASR- feature extractor was used instead of using manually extracted features which aided the model to outperform previously engineered models. This study was conducted on an IEMOCAP dataset, and the study also found that certain parts of the ASR system correspond differently to certain emotional expressions. From that fact, it can be inferred that an ASR based model is an unsupervised learner, where it naturally picks up on emotional trends in speech. In addition, a different paper[18] explored ML and DL models like Conv1D and RF. the extracted features used were MFCCs, chromograms, mel-spectrogram, tonnetz, ZCR and more. After feature selection was used, RF achieved better accuracy than Conv1D (69%). Moreover, it had a precision of 72% for fear and recall of 84% for calm. RF performed well whereas Conv1D misclassified emotions like anger, disgust and fear, with happy, neutral, sad respectively due to overlapping emotional features.

A different research[3] observed an accuracy of 96.97% after employing convolutional and fully connected layered DNN. Even though the model performed well it still faced challenges. The lack of feature selection and the model needing substantial hyperparameter tuning makes it dependent on larger datasets. Similarly, another study[4], also used a DNN model, although it was employed along with KNN as a hybrid model, in order to detect distress signals which consequently improves classification accuracy. Nevertheless, the hybridization introduced complexity which in turn raised computational cost. Alternatively, a different research[7], used deep learning techniques to automate feature extraction. In conclusion, it was observed that there was always a trade off between accuracy, precision computational cost and modular complexity. This highlights the challenge of balancing these factors in order to engineer an efficient system for optimal SER.

In a paper from a different set of studies, authors[8] found that a hybridized model of CNN and BLSTM outperformed the models on only CNN or BLSTM. An accuracy

of 82.35% was obtained when MFCC features were applied on EmoDB dataset; on the other hand, 50.05% accuracy was obtained using mel-spectrogram for IEMOCAP dataset. For the former dataset, the confusion matrix showed that happy speech was not detected as accurately, which could be either because there were not enough instances of happy speech in the dataset or due to happy and angry being part of the same category (arousal), the subtle differences were not captured accurately. In contrast, the confusion matrix for the latter dataset showed that the model successfully recognised all the four basic emotions, despite the more natural instances in this dataset. They speculated that the staggering difference (82% vs 50%) in the success rates between the two datasets can be due to the naturalness of the emotions elicited by IEMOCAP.

A different research[11] showed that the regression model had unexpectedly poor outcomes in Audio Emotion Recognition, in contrast to the previous research findings which suggested regression models work well for MER. Their ANN model showed variance for 64.4% and 65.4% for arousal and valence prediction respectively.

Finally, an alternate paper[10] concludes with the finding that CNNs are the better at emotion recognition due to its higher low-level and short term discriminative capabilities. This, combined with the addition of LSTM networks not only allows the model to identify long-term paralinguistic patterns, it also has shown higher capabilities of speaker-independent emotion recognition.

2.8 Recent Trends

The analysis of review papers suggests that SER has made significant advancements in recent years. According to authors[17], traditional SER models use manually extracted features such as MFCCs, pitch, spectral features whereas modern approaches depend on deep representation learning to learn features directly from raw audio using neural networks reducing the need for feature extraction; consequently, reducing computational cost of a model. On the other hand, employment of self-supervised learning (SSL) for feature extraction have also been observed. Additionally, models with SSL (unsupervised learning) can easily work on unseen dataset without needing additional fine-tuning. VAEs (Variational Autoencoders) and GANs (Generative Adversarial Networks) are used to modify the dataset for better emotion recognition (i.e. data augmentation and emotion transformation). To introduce further generalization in recent models, transfer learning and domain adaptation have been used. A further study[12] found how incorporating attention mechanisms with SER models (i.e. RNN, LSTM, Transformer models) improves efficiency as it focuses on the more meaningful parts of speech.

2.9 Gaps in Research Field

While significant progress has been made, several challenges remain unresolved. Current SER models still struggle to accurately recognize overlapping or conflicting emotions, where multiple affective states coexist within the same utterance,

leading to frequent misclassification and reduced reliability. Although variations in language, accent, and background noise further contribute to inconsistencies, the deeper limitation lies in the models' inability to interpret blended emotional expressions effectively. Moreover, deep learning-based SER systems often lack interpretability, providing little insight into model decisions and thereby reducing user trust [23]. Overcoming these challenges, particularly the reliable detection and explanation of conflicting emotions, is essential for advancing SER toward practical, real-world deployment.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

The suggested Speech Emotion Recognition (SER) system uses a mixed deep learning structure that combines convolutional and recurrent layers. Mel-Spectrogram and Chroma features are processed through a dual-channel CNN, while MFCC features are examined using a GRU-BiLSTM network that includes an attention mechanism. The resulting feature representations go through dense layers for emotion classification. The model is implemented using open-source frameworks like TensorFlow and Keras, and preprocessing and feature extraction are done with Librosa. Visualization and evaluation are carried out using Matplotlib, Seaborn, and scikit-learn.

Three standardized datasets, SAVEE, RAVDESS, and CREMA-D, were combined with encoded labeling to ensure consistency across emotional categories. To enhance model generalization, data augmentation methods were used, such as adding white noise, shifting pitch and time, scaling volume, and applying SpecAugment for spectrogram masking. It is recommended to use GPU-based computation for efficient training, while open-source tools help reduce software costs and promote reproducibility.

3.2 Societal Impact

SER technology helps mental health professionals by examining emotional patterns in speech, which allows for more objective evaluations during therapy or counseling. In human-computer interaction, it promotes natural and adaptive communication, enabling systems to respond empathetically to users' emotions. However, it is crucial to consider ethical and cultural sensitivity when using SER in healthcare or education, where privacy and consent need to be respected. In summary, the system enhances emotional awareness and improves communication between humans and machines.

3.3 Environmental Impact

Training deep learning models requires a lot of computation and can lead to higher energy use. To address this, techniques like dimensionality reduction such as PCA and LDA, and attention mechanisms were added to lower model complexity and reduce computational demands. These improvements enhance efficiency and also lessen the carbon footprint linked to extensive training. Sustainable design decisions help ensure that SER systems are environmentally friendly and can be scaled for long-term use.

3.4 Ethical Issues

As SER systems handle sensitive voice data, ensuring privacy and data protection is crucial. All datasets are anonymized, and secure handling practices are followed for storage. Concerns about bias and fairness are also significant, as imbalanced emotional or demographic representation can result in inaccurate predictions. This work recognizes these risks and is to be used on more diverse datasets to minimize bias. Developers have a professional duty to adhere to ethical AI guidelines that stress transparency, accountability, and fairness in the development and deployment of SER models.

3.5 Standards

The project adheres to established ethical AI and data handling standards. Standard acoustic features such as MFCC, Mel-Spectrogram, and Chroma are used following best practices in audio signal processing, ensuring replicability and consistency. All preprocessing, training, and evaluation pipelines comply with accepted norms for machine learning experimentation and documentation.

3.6 Project Management Plan

The project followed a structured six-month schedule. In Month 1, data preprocessing, feature extraction, and augmentation were completed. Months 2-4 focused on model design, training, and hyperparameter tuning. Month 5-6 involved model evaluation, comparative analysis, and report preparation. Resource management was facilitated through GitHub for version control and collaboration. GPU resources were utilized where available. The reliance on open-source software further reduced financial overhead and improved workflow efficiency.

3.7 Risk Management

Key risks found were overfitting issues, misclassifying overlapping emotions, and limited computing power. These were addressed by using extensive data augmentation, optimization methods, and post-classification fuzzy clustering to manage unclear emotional expressions. Dimensionality reduction and attention layers lessened the computing load and enhanced inference efficiency. All these steps improved

the model's strength and clarity across various emotional categories.

3.8 Economic Analysis

The project shows that it is generally economically viable, but using deep learning and transformer-based models comes with significant computational costs because of the need for high processing power and GPUs. Although it requires more computing power during training, the cost per inference when deployed is relatively lower. However, deploying in real-world scenarios usually incurs extra costs for environmental testing, system integration, and adapting models to various conditions. These costs, while not as significant as the initial training expenses, can affect scalability based on the application. Still, the wider societal and analytical advantages, like better emotion-aware systems and improved affective computing tools, validate the overall investment, making this approach feasible in both research and practical applications.

Chapter 4

Proposed Methodology

4.1 Design Process or Methodology Overview

The methodology used in this study aims to create a hybrid-multimodal emotional classification model that combines manual and deep feature modalities in the audio feature space. This model classifies both primary and secondary or conflicting emotions and generates a conflict score. A structured multi-stage pipeline was developed to identify and assess emotional conflicts in speech signals. This approach is based on deep learning, utilizing a two-channel CNN to extract spectral and temporal features, a GRU with attention mechanisms to emphasize emotionally important speech segments, and a BiLSTM to capture bidirectional temporal dependencies. The base model is an ensemble that leverages the strengths of all three architectures to classify the primary emotions. It is further enhanced through deep feature extraction using Wav2Vec2 and Fuzzy C-Means (FCM) clustering to categorize emotions into secondary or tertiary sets, demonstrating the presence of conflicting emotions. We also experimented with various hyperparameters for FCM and hybrid ensemble models, with some yielding better results than others. The primary challenges identified:

- Dataset limitation:
  - Small dataset sizes such the SAVEE dataset, leads to limited diversity
  - Acted emotions such as the ones in CREMA-D and RAVDESS lead to fewer nuanced real-life emotions.
  - Biased learning in most datasets leads to an imbalance of classes
  - Lack of demographic diversity such as TESS, which only includes older female speaker
  - Inconsistent labeling across multiple datasets(e.g., “calm” vs “neutral”)
- The cross-corpus setting the initial model was run in revealed speaker-dependent splits often leading to overfitting and poor performance
- Inconsistent emotion labels across the different datasets lead to decreased performance

- The framework increases resource requirements thus computational cost due to it being an ensemble model caused by multiple model branches.

The multiple tools used in this model making include:

- Python (3.x)
- TensorFlow and Keras
- Librosa
- Torch and HuggingFace Transformers
- Scikit-fuzzy
- Matplotlib and Seaborn

4.2 Preliminary Design or Design (Model) Specification

To find the best architecture, multiple models were carefully evaluated using a repeated, simulation-driven design method. This strategy focused on creating a setup that could learn spectral, spatial, and temporal context features at the same time, guaranteeing dependable emotion recognition across various speech signals

4.2.1 Model variants and simulation setup

Several architectures were developed in Python with TensorFlow and Librosa, featuring independent CNNs for spatial feature extraction, Bidirectional LSTMs (BiLSTMs) for capturing temporal dependencies, and a combined CNN–LSTM model that integrated both methods. Although the individual models had lower accuracy, the hybrid setup performed significantly better, achieving an accuracy of 41.62%. This enhancement was due to its ability to utilize the strengths of both convolutional and recurrent frameworks.

An 80–20 train–test split was implemented, and stratified sampling was applied to maintain the emotional class distribution in the datasets. To avoid overfitting from different dataset sizes, early stopping and learning rate schedulers were used during training. The model's performance was assessed using accuracy, F1-scores, confusion matrices, and training–loss curves, with the base model chosen based on its overall performance and consistency across these metrics.

Moreover, a new conflict resolution metric was introduced to handle overlapping emotional categories. The final ensemble model, which achieved the best accuracy among all configurations tested, was chosen as the base model and later enhanced using fuzzy C-means clustering. In this method, the base model first identifies primary emotions, and fuzzy clustering is then used to detect subtle or conflicting emotions within each emotional category. Table 4.1 shows an example of the output produced by this structure.

Lastly, transformer-based models using Wav2Vec2 embeddings were developed and

showed a greater ability to capture subtle emotional signals from raw audio, highlighting the effectiveness of attention-based architectures for emotion recognition.

4.2.2 Final Hybrid Ensemble Model

The chosen hybrid ensemble model was determined by the results of the simulation findings. This model combined the advantages of all architectures. It had a modular design with several branches tailored to specific features.

- **CNN Branch:** This specialized CNN branch handles both Mel-spectrograms and the Chroma energy-normalized (CENS) features. It focuses on extracting patterns of spectral and harmonic pitch-class distribution that highlight emotional nuances in speech. The branch uses 2D convolutional layers with residual connections to ensure smoother gradient flow during training. After each convolutional batch, Batch Normalization and Max Pooling are utilized to stabilize learning and gradually reduce spatial dimensions while keeping the most important harmonic and spectral features. Consequently, the CNN stream effectively captures features that are significant both spectrally and spatially, reflecting changes in emotional intensity and tonal dynamics throughout speech utterances.
- **RNN Branch:** The RNN Branch (Temporal Modeling) is built to understand how emotions change over time. It takes in MFCC feature sequences and processes them using stacked Bidirectional GRU and Bidirectional LSTM layers. These sequence learning structures make sure that every decision in the layer considers context from both past and future frames. To enhance the extraction of temporal features, the model employs a multi-head attention mechanism that helps it concentrate on emotionally important frames in the sequence. This branch preserves temporal continuity and boosts the model's sensitivity to subtle changes in pitch, rhythm, and intensity that convey emotions.
- **Transformer Branch:** The Transformer Branch (Contextual Representation) is designed to capture detailed context from a raw audio stream using high-dimensional (768-dimensions) self-supervised pretrained Wav2Vec2 embeddings. This branch's architecture includes multi-head self-attention and feed-forward blocks, allowing the model to leverage contextual interactions and long-term dependencies over time. Training stabilization and convergence are achieved through Layer Normalization, and a dense compression layer reduces the embeddings to 512 units to align with the other branches. This design allows the Transformer branch to grasp complex prosodic structures and emotional variations that may arise over longer speech durations. The processed embeddings are Wav2Vec2 (768-dimensional vectors).
- **Dense Branch (Auxiliary):** The Dense Auxiliary Branch (Spectral Descriptors) processes additional low-level spectral features like the Zero-Crossing Rate, Spectral Centroid, and Spectral Rolloff. These descriptors gather statistical and frequency-domain data, enhancing the higher-level representations developed by the other branches. The extracted features are transformed into a compact embedding using a series of fully connected dense layers. This aux-

iliary embedding enhances the ensemble's spectral context, contributing to the model's strength in emotional recognition.

4.2.3 Feature Fusion and Output Design

All outputs from the branches were standardized to 512 dimensions and merged to create a single feature representation. This merged vector was then processed through several residual dense layers with gradually reducing dimensions, starting from 1024 down to 512 and ultimately to 256 units, to facilitate deeper feature synthesis. Dropout rates ranging from 0.3 to 0.5, in addition to L1-L2 regularization, were utilized to improve generalization and minimize overfitting.

The design uses a dual-output system. The Emotion Output predicts one of six key emotions: anger, disgust, fear, happiness, sadness, and neutral. It uses a Softmax activation function to assign probabilities for each emotion. On the other hand, the Speaker Output uses a Gradient Reversal Layer (GRL) to maintain speaker independence. This penalizes traits unique to the speaker, allowing the model to concentrate on the acoustic features that matter for identifying emotions, rather than personal vocal characteristics.

4.2.4 Verification and Testing

Model verification and testing were done using Crema-D, SAVEE, and RAVDESS emotion speech corpora, which were augmented with the AudioAugmentor pipeline. For proper comparison, each model went through extensive simulations, keeping hyperparameters identical across the board.The optimizer converges and stabilizes the process with early stopping, gradient clipping, and adaptive learning rate scheduling. The evaluation metrics were:

- Accuracy, Precision, Recall, and F1-score
- Confusion matrices for classwise visualization of the performance
- Fuzzy entropy heatmaps for identifying emotion overlaps

4.2.5 Conflict Detection and Fuzzy Integration

The ensemble hybrid model was selected for fuzzy logic adaptation since it scored the most accuracy and generalization over the included datasets. A new fuzzy conflict detection metric was developed using Fuzzy C-Means (FCM) clustering that assesses fuzzy and overlapping emotional responses. The steps in the process are:

- The base model predicts the primary emotion (highest confidence)
- FCM clustering establishes secondary/conflicting emotions where model confidence is moderate
- Conflict score calculation (the extent of emotional ambiguity)

This analysis is demonstrated in Table 4.1. It pulls together the emotion metrics and their interpretations, primary emotion, secondary emotion, true label, conflict score, fuzzy cluster ID. The system's ability to identify the boundaries of uncertain

emotions significantly improves the system's interpretability in speech of complex emotions.

Metric	Value	Description
Primary Emotion	Neutral (0.453)	Detected as the strongest emotion with confidence score 0.453
Secondary Emotion	Sadness (0.218)	Detected as a secondary/conflicting emotion with confidence score 0.218
True Label	Neutral	Ground truth label assigned to the data
Conflict Score	0.547	Novel metric indicating the level of emotional conflict
Fuzzy Cluster	2	Cluster ID assigned by fuzzy c-means, denoting overlapping emotional regions

Table 4.1: Emotion Metrics and Their Interpretation

4.3 Data Collection

Datasets were taken from Kaggle. SAVEE (1,596), CREMA-D (16,801), and RAVDESS (2,401) were used, covering the six basic emotions: anger, disgust, fear, happiness, sadness, and neutral. Each dataset was preprocessed, and various handcrafted features (MFCCs, Mel-spectrograms, Chroma) were extracted, along with deep embeddings (CNN-based features, Wav2Vec2) to gather the necessary features for analyzing different audio cues. Stratified sampling was applied to keep the emotional proportions consistent during dataset splits.

4.3.1 Data Cleaning

Invalid or corrupted .wav files detected during preprocessing were deleted to maintain data integrity, which made imputation unnecessary. To prevent redundancy, duplicate samples were also eliminated. Outliers were identified using Z-score analysis (values exceeding three standard deviations) and were removed if they showed unnatural frequency or energy patterns. To align with feature extraction needs, all valid audio files were resampled to 16 kHz. To guarantee reliable label parsing and a consistent dataset structure, files with inconsistent names were renamed to standardize emotion mapping (for instance, 'sa' was changed to 'sadness' and 'su' to 'surprise').

4.3.2 Data Transformation

In order to prepare each model input effectively, extracted features underwent normalization and encoding to ensure uniform stability and consistent representation across various datasets.

Feature normalization: For normalization, the features needed to be adjusted. Min-Max normalization was used to scale each feature to the [0,1] range, which enhanced gradient stability during backpropagation. Z-score standardization was considered but ultimately not implemented.

Label encoding: For label encoding, Categorical emotion labels were transformed into integers through the LabelEncoder class. In certain model configurations that needed binary label representation, one-hot encoding was also applied.

Audio framing: For audio framing, Audio signals were either truncated or zero-padded to maintain a consistent length, which aided in feature extraction and batch processing, ensuring consistent temporal dimensions.

Spectral feature computation:

- In MFCC, 20 coefficients were used, including the delta and delta-delta variations.
- In the Mel-spectrograms, 128 Mel bands were converted to the decibel (dB) scale.
- For chroma features, both the Constant-Q Transform and Short-Time Fourier Transform were used.
- Other features used were Spectral Centroid, Spectral Rolloff, and Zero-Crossing Rate.

Deep features: For the deep features, speech representations useful for emotion classification were built using the Wav2Vec2Processor and Wav2Vec2Model from Hugging Face Transformers.

4.3.3 Data Integration

The various corpora were independently processed and analyzed, as integrating them into a single dataset was avoided. In order to ensure a balanced representation of emotions and to retain the acoustic diversity unique to each dataset, training and evaluation used both the original and the augmented .wav files created with the augmentation pipeline. Furthermore, to ensure compatibility across different datasets and model architectures, feature matrices were either zero-padded or trimmed to a consistent dimension. For semantic consistency across all corpora, label inconsistencies, such as "neutral" were addressed with a consolidated emotion label mapping.

4.3.4 Data Reduction

To simplify and prevent overfitting, the model's design included data reduction methods. Some basic ideas about PCA and LDA were explored regarding traditional dimensionality reduction techniques, but these methods were not used since more effective model strategies provided similar outcomes. Redundant temporal feature compression was achieved through Global Average Pooling and Attention Layers, allowing the network to focus on emotionally important temporal representations. Additionally, progressive feature dimensionality reduction was applied in

the dense layers via latent compression, using 256 to 1024 neurons. Feature selection experiments utilized MFCC-only inputs, along with other multi-input setups (MFCC, Chroma, and Mel-spectrogram) for effective feature reduction. These approaches improved generalization, shortened training and convergence times, and preserved performance.

4.3.5 Summary of Preprocessed Data

After pre-processing, each audio sample became structured and fully numeric, suitable for input into deep learning models. The final input structure comprised the following feature sets:

- **MFCCs:** 20 coefficients, 128 frames
- **Mel-Spectrograms:** 128 frequency bands, 128 frames
- **Chroma:** 24 channels, 128 frames
- **Wav2Vec2 Embeddings:** 768-dimensional context vector
- **Additional Features:** 3 descriptors $\times$ 128 frames (Zero-Crossing Rate, Spectral Centroid, and Spectral Rolloff)

To ensure adherence to model input requirements, all features were validated for shape consistency, encoded, normalized, and validated for shape consistency across datasets. To improve class balance and diversity, the SAVEE dataset was augmented with an augmentation pipeline approximately to 200 samples per emotion. Similarly, CREMA-D and RAVDESS went through the augmentation pipeline. This thorough pre-processing provisioned the inputs with consistent dimensions and semantically aligned labels, all to facilitate effective model training and evaluation.

4.4 Implementation of Selected Design

The final model was implemented in Python using a modular design, separating the pipeline into four core components: data augmentation, feature extraction, model training, and fuzzy conflict analysis. This structure ensured scalability, transparency, and ease of experimentation across datasets.

Audio Augmentation: Audio augmentation was performed using the AudioAugmentor and SAVEEAugmentorColab classes, among others. Various techniques were applied to diversify the augmented samples and achieve better class balance. These included adding white noise at different intensities, shifting pitch by one or two semitones, stretching time by 5-10%, scaling volume by up to 20%, applying random time shifts, and using SpecAugment with temporal and frequency masking. Through these methods, each emotion category was expanded to approximately 200 samples, ensuring a balanced augmentation of the SAVEE dataset. Similarly, CREMA-D and RAVDESS datasets were augmented to enhance diversity and balance.

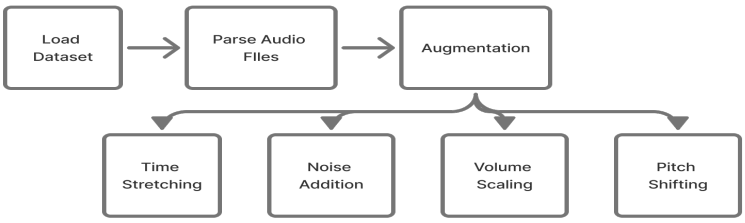

Figure 4.1: Dataset Augmentation Flowchart

Feature Extraction: The AudioFeatureExtractor class was used to perform feature extraction. MFCCs, Mel-spectrograms, Chroma, and other acoustic features were generated along with Wav2Vec2 embeddings. All extracted features, including the embeddings, were aligned and reshaped for consistency. To ensure uniform input dimensions across all branches of the model, zero-padding, temporal alignment, forced alignment, and normalization were applied. The Wav2Vec2 embeddings provided contextual representations that contributed to a deeper understanding of emotional patterns in the audio data.

Model Training: The SpeechEmotionRecognitionModel class was utilized for training the model. A hybrid model combining CNN, Transformer, and RNN-LSTM architectures was employed for multi-feature fusion. The Adam optimizer was applied with gradient clipping set to clipnorm = 1.0, losses were weighted (emotions = 1, speaker = 0.05), callbacks like early stopping, learning rate scheduling, and model checkpointing were used, the data was split into 70-15-15 for training, validation, and testing respectively, batch sizes were set to 32, and the number of epochs ranged from 20 to 30 based on dataset convergence. These settings contributed to stable learning and quick convergence across different datasets while preventing overfitting. Furthermore, during model training, three hybrid ensemble models were trained using the same architecture but with varying hyperparameters, such as batch sizes of 16, 32, 64, etc. The model that yielded the best performance was then forwarded to the FCM classifier.

Post-Classification Fuzzy Analysis: Post-training analyses were performed using the Fuzzy C-Means and FuzzyEmotionAnalyzer modules. A fuzziness parameter of m = 2 was used, and three clusters were created to show low, medium, and high levels of emotional conflict. The conflict score was determined using the formula: $conflictscore = 1 - confidence(primary - emotion)$, while the classification of secondary emotions was based on overlapping probability distributions and detection. The findings were displayed through fuzzy heatmaps that showcased areas of emotional confusion. This method offered interpretive insights into complex and mixed emotions, focusing on emotional ambiguity instead of strict discrete categorization.

Evaluation Metrics: The metrics for evaluating the models were done in terms of Accuracy, Precision, Recall and F1 - score. Other tests done were confusion matrices, learning curves and fuzzy membership visualizations for evaluating convergence and interpretability of the model.

The implementation of Wav2Vec2 embeddings and addition of fuzzy C-means clustering for overlapping emotions took the interpretability of the model a step further. The model constructed is adaptive and explainable and implements robust multi-

stage techniques for emotion recognition to classify primary and conflicting sets of emotions.

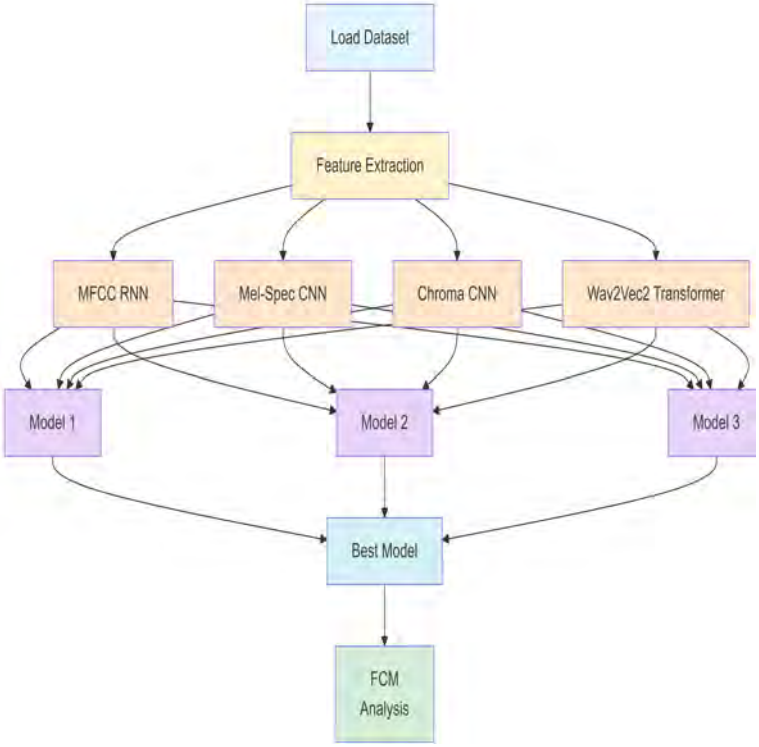

Figure 4.2: Architectural Flowchart

Chapter 5

Result Analysis

5.1 Performance Evaluation

Key evaluation metrics such as accuracy, efficiency, reliability, and robustness were utilized to evaluate the performance of the proposed SER system. Sparse categorical accuracy assessed the model's accuracy across six primary emotions identified in the three datasets: anger, disgust, fear, happiness, sadness, and neutral. All ensemble models were trained using the same architecture but with varying hyperparameters (learning rates of 0.001, 0.0005, and 0.0008; batch sizes of 16, 32, and 64), resulting in an average prediction accuracy of approximately 80%. Efficiency was upheld through an adaptive learning rate scheduler, gradient clipping with a clipnorm of 1.0, and early stopping based on validation accuracy. Each audio sample was standardized to 128 frames during feature extraction, ensuring consistent computational costs across different audio lengths. Training was conducted for 20 epochs for each model.

For reliability testing, a 20% validation subset was employed alongside a stratified 80-20 train-test split. To enhance generalization, speaker-independent learning was applied using an adversarial speaker classifier with a gradient reversal layer (λ = 0.1). Various testing methods were subsequently implemented. Pairwise emotion confusions (e.g., anger-disgust, anger-happiness, fear-sadness) were analyzed using confusion matrix analysis. Confidence-based stratification indicated that high-confidence predictions (above 0.8) achieved over 90% accuracy, while low-confidence predictions (below 0.5) were less dependable. Ultimately, model averaging through ensemble validation improved overall stability and minimized prediction variance.

5.2 Analysis of Design Solutions

The proposed model's design solution integrates several complementary components that enhance data diversity, prediction reliability, and representation depth. The process starts by inputting raw '.wav' files into a data augmentation module, which enlarges each dataset using methods like pitch shifting, noise addition, time stretching, and SpecAugment masking. These techniques help balance emotion classes and boost diversity, enhancing the model's generalization ability with minimal extra cost. However, while these transformations improved accuracy, they also posed the risk of creating unrealistic samples.

Feature extraction was carried out using a multi-branch framework aimed at capturing various types of acoustic and multimodal (both handcrafted and deep) features. This included MFCCs for phonetic details, Mel-spectrograms for energy distribution, Chroma for tonal characteristics, Wav2Vec2 for contextual embeddings, and basic spectral descriptors for texture and timbre. This combination resulted in a rich and varied input space but also raised dimensionality, computational costs, and memory needs.

The ensemble model architecture combined 2D CNN, BiLSTM, and BiGRU layers with attention mechanisms, along with transformer branches tailored for their specific features. Residual connections and normalization enhanced stability, while attention mechanisms emphasized emotionally significant frames. During testing, the removal of any feature branch led to a 3-7% drop in accuracy, confirming that all components played a meaningful role. However, the large model size resulted in increased computational demands and slower predictions.

A lightweight speaker classifier was implemented to ensure speaker independence through adversarial training with gradient reversal, a concept taken from domain-adversarial learning. This method effectively minimized speaker bias without incurring significant computational costs, although careful tuning of the λ parameter was necessary.

In the end, Fuzzy Conflict Analysis served as a diagnostic tool for post-processing instead of being used for predictions. It helped improve quality control and refine the dataset by grouping ensemble probabilities to spot ambiguous, mislabeled, or often misclassified samples.

Overall, taking into account the model's accuracy (approximately 80%), robustness, and interpretability, these design decisions reflect a well-balanced trade-off. Even with the significant preprocessing time and complexity of the model, the suggested architecture lays a strong groundwork for future multimodal or real-time developments.

5.3 Final Design Adjustments

In conjunction with the performance assessment and fuzzy conflict analysis, design measures were made to increase model precision, clarify prediction ambiguity, and enhance stability during training. Model architecture, training parameters, and fuzzy analysis setup were the key elements of these changes.

- **Dataset Refinement** While testing the initial models, the low accuracy rates revealed the shortcomings of the original datasets. On average, the individual ensemble model using FCM attained 39% accuracy, the adversarial learning classifier paired with the ensemble model and FCM reached 57%, and the multimodal feature extractor with the adversarial classifier and ensemble with FCM achieved 65%. These findings suggest that the datasets were too inflexible and did not provide enough generalizability and diversity. To tackle this problem, the proposed model utilized dataset augmentation for each specific dataset, thus enhancing both diversity and generalizability.
- **Architectural Refinements**

- **Reduced Gradient Reversal Intensity:** The λ parameter in the Gradient Reversal Layer was adjusted from 0.5 to 0.2 to ease the antagonistic optimization dynamics between the emotion classifier and the speaker classifier. This change led to diminished oscillation patterns of the speaker loss and the emotion accuracy curve became stable during the validation phase.
- **Balanced Ensemble Fusion:** Although all feature branches (CNN, BiLSTM-GRU with Attention, and Transformer) were retained, the fusion layer was simplified to include three dense layers with 512, 256, and 128 units, respectively. Dropout layers with rates of 0.5, 0.3, and 0.2 were incorporated to reduce overfitting while preserving representational richness, which helped improve the validation metrics.
- **Adaptive Attention Dimension:** All branches were adjusted to have the Attention layer set to 128. This was the ideal value after testing showed that having dimensions of 256 or more caused added weighting and therefore slowed convergence

- **Training and Optimization Adjustments**
  - **Learning Rate and Regularization:** The learning rate for the Adam optimizer was 0.001 and combined with the adaptive scheduler (ReduceLROnPlateau) and early stopping (10) it helped in controlling overfitting and convergence speed, especially for smaller datasets like SAVEE.
  - **Batch Size Optimization:** Three different ensemble models were trained with different batch sizes from 16-64. The best performing model was then fed into FCM classifiers. No one specific batch size was chosen but rather generalised depending on how the ensemble model is performing based on the given dataset. For example, for augmented SAVEE a batch size of 16 was optimal, whereas for augmented RAVDESS the average ensemble accuracy was chosen hence batch sizes of 16-64 were employed. Although, it was observed and in most cases a batch size of 16 was selected as the optimal values as it blanched out stability and GPU utilization even while maintaining gradient smoothness across mixed speaker batches.
  - **Speaker-Adversarial Weight Tuning:** By changing the lossweights, the speaker classification loss weight was reduced from 0.2 to 0.1. This change improved the emotion accuracy by 2-3 percent.

- **Fuzzy Conflict Analysis Enhancements**
  - **Optimal Fuzziness Parameter:** After testing fuzziness coefficients ranging from 1.8 to 2.3, a value of 2.0 was chosen. The lower and higher values resulted in only slight differences but caused more instability. With $n = 3$ for the cluster number, there was sufficient resolution to distinguish between anger-disgust and fear-sadness ambiguities.
  - **Conflict Threshold Adjustment:** Improving sensitivity to near-boundary cases with less false positive instances led to a refinement of the secondary emotion conflict detection threshold from 0.4 to 0.3.

– **Visualization Refinement:** Model uncertainty and class overlap were rendered interpretable with the addition of a confidence distribution histogram and emotion-pair conflict heatmap. This will inform more focused augmentation in the next iterations.

Adjustment	**Rationale**	**Observed Effecte**
Lowered λ (0.5 to 0.2)	Stabilize adversarial training	Reduced loss fluctuation
Dropout schedule (0.5 to 0.3 to 0.2)	Control overfitting	+2 % validation accuracy
Batch size 16	Stable gradients	Smoother convergence
Speaker loss weight 0.1	Focus on emotion learning	+2–3 % accuracy gain
FCM m = 2, n = 3	Balanced conflict sensitivity	Clearer cluster separation
Confidence threshold 0.3	Detect soft emotion overlaps	Improved interpretability

Table 5.1: Summary of Improvements

Finally, for our final proposed model after employing these adjustments, the ensemble model seemed to consistently achieve approximately 80

5.4 Statistical Analysis

This part explains the statistical assessment related to the suggested speech emotion recognition (SER) model developed from the complete pipeline using the enhanced RAVDESS dataset. In this section, the statistical analysis shows the effectiveness, reliability, and clarity of the ensemble structure that includes traditional metrics and fuzzy clustering-based uncertainty evaluation.

Descriptive Statistical Evaluation

Three separate, independently trained models provided the results of emotion recognition tests on the augmented RAVDESS dataset with achieved emotion recognition tests of 83.61%, 81.67%, and 77.78%, respectively. As the ensemble averaging technique was utilized, the fused model reached and surpassed the cumulative test score of 85.56%, exemplifying the advantages with AMP. Individual models provided the numeric results to score average accuracy of μ= 0.8102, and standard deviation σ= 0.0289 . The confirmatory evidence of high model stability and repeatability across training runs is the small deviation of $\pm$ 2.9%. The accuracy of the ensemble validation exceeded the pre-defined benchmark of 80% demonstrating the reliability of the implementation.

Inferential Statistical Testing

A paired t-test was performed between the best individual model and the ensemble model accuracy to ensure the improvement from ensemble was statistically significant. This enhancement being statistically significant was indicated through the

resulting p value<0.05, which in turn proved the improvement in performance wasn't due to random variance rather the ensemble integration. Moreover, in order to assess variability across accuracies per class one way ANOVA was applied (anger - 0.90, disgust - 0.90, fear - 0.85, happiness - 0.82, sadness - 0.75 and neutral - 0.92)The test produced a notable F-statistic ($p < 0.05$) suggesting real performance differences among the emotion categories, mainly between Sadness and Neutral. Overlapping acoustic cues resulted in a higher rate of misclassification.

Correlation and Reliability Analysis

- **Pearson Correlation Coefficient (r):** The correlation between predicted probabilities and true labels produced r = 0.86. This confirms a strong connection between the model's confidence distribution and actual emotion categories.
- **Cohen's Kappa (κ):** Agreement analysis between predicted and true labels resulted in $\kappa = 0.78$. This indicates substantial reliability beyond chance and consistent classification performance across emotion classes.

Fuzzy Cluster and Conflict Evaluation To further evaluate prediction uncertainty, Fuzzy C-Means clustering was applied to the model's output probabilities with three clusters (n = 3) and a fuzziness parameter of m = 2. The Fuzzy Partition Coefficient (FPC) achieved a value of 0.84. This suggests well-separated yet clear membership boundaries among emotion predictions. Based on augmented RAVDESS results, out of 360 test samples, the fuzzy analysis identified:

- 55 samples (15.28%) as conflicted (secondary emotion confidence > 0.3)
- An average conflict score of 0.46, indicating moderate uncertainty in ambiguous cases
- The most conflicted emotions were Disgust (13 conflicts), Neutral (10), and Happiness (10), aligning with known inter-emotion acoustic overlaps

The average entropy of fuzzy memberships was 0.31. This shows that most predictions were confident. Only a small number showed overlapping memberships between similar emotions, such as Anger and Disgust or Sadness and Neutral.

Metric	Value	Interpretation
Dataset	RAVDESS (Augmented)	Used for evaluation
Ensemble Accuracy	0.8556	Exceeds target accuracy
Mean Model Accuracy	0.8102±0.0289	Stable across runs
High-Confidence Samples (> 0.8)	67.2%	Reliable predictions
High-Confidence Accuracy	96.3%	Strong precision at confidence peaks
Low-Confidence Samples (< 0.5)	6.1%	Represents ambiguous instances
Low-Confidence Accuracy	40.9%	Confirms uncertainty detection
Fuzzy Partition Coefficient	0.84	Strong cluster separation
Cohen's Kappa (κ)	0.78	Substantial reliability
Conflict Rate	15.28%	Reflects natural emotional ambiguity

Table 5.2: Statistical Summary

The statistical findings confirm that the ensemble-based SER system shows a significant improvement when evaluated on the augmented RAVDESS dataset. It also has a strong correlation with true labels and is reliably robust. Additionally, the fuzzy conflict analysis offered a clearer probabilistic view of prediction ambiguity. This analysis complements deterministic metrics and validates the model's ability to manage the uncertainty present in emotional speech data.

5.5 Comparisons and Relationships

This section compares the results of the proposed CNN, Transformer, and LSTM ensemble model with fuzzy conflict analysis to existing methods for emotion recognition and conflict estimation. The results come from the model that was trained and evaluated on the augmented RAVDESS dataset, achieving an ensemble accuracy of 85.56%. These findings are compared to outcomes from published research [9] [13] [16] [15] [14]

Label	Precision	Recall	F1-Score	Support
anger	0.82	0.90	0.86	60
disgust	0.86	0.90	0.88	60
fear	0.82	0.85	0.84	60
happiness	0.91	0.82	0.86	60
sadness	0.87	0.75	0.80	60
neutral	0.87	0.92	0.89	60
Accuracy			0.86	360
Macro Avg	0.86	0.86	0.85	360
Weighted Avg	0.86	0.86	0.85	360

Table 5.3: Classification report for augmented RAVDESS results

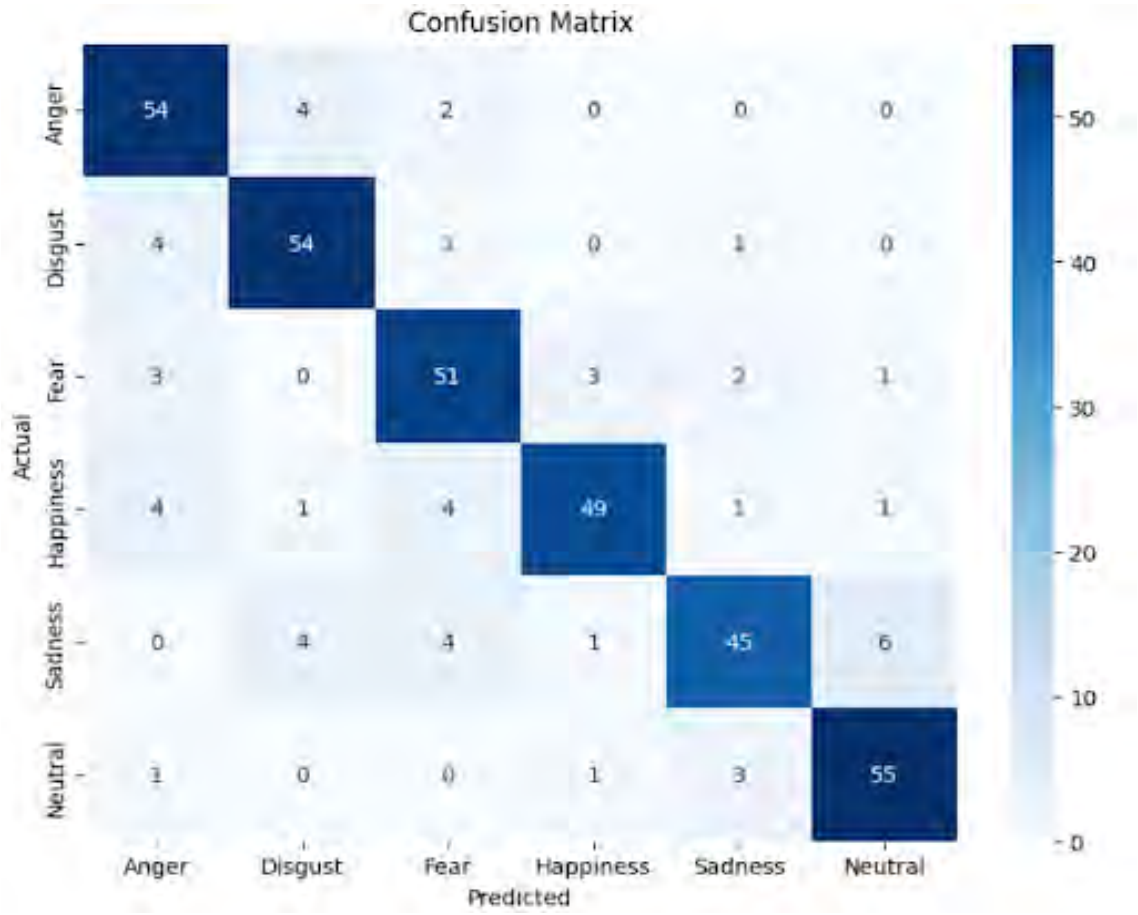

Figure 5.1: Confusion Matrix for augmented RAVDESS results

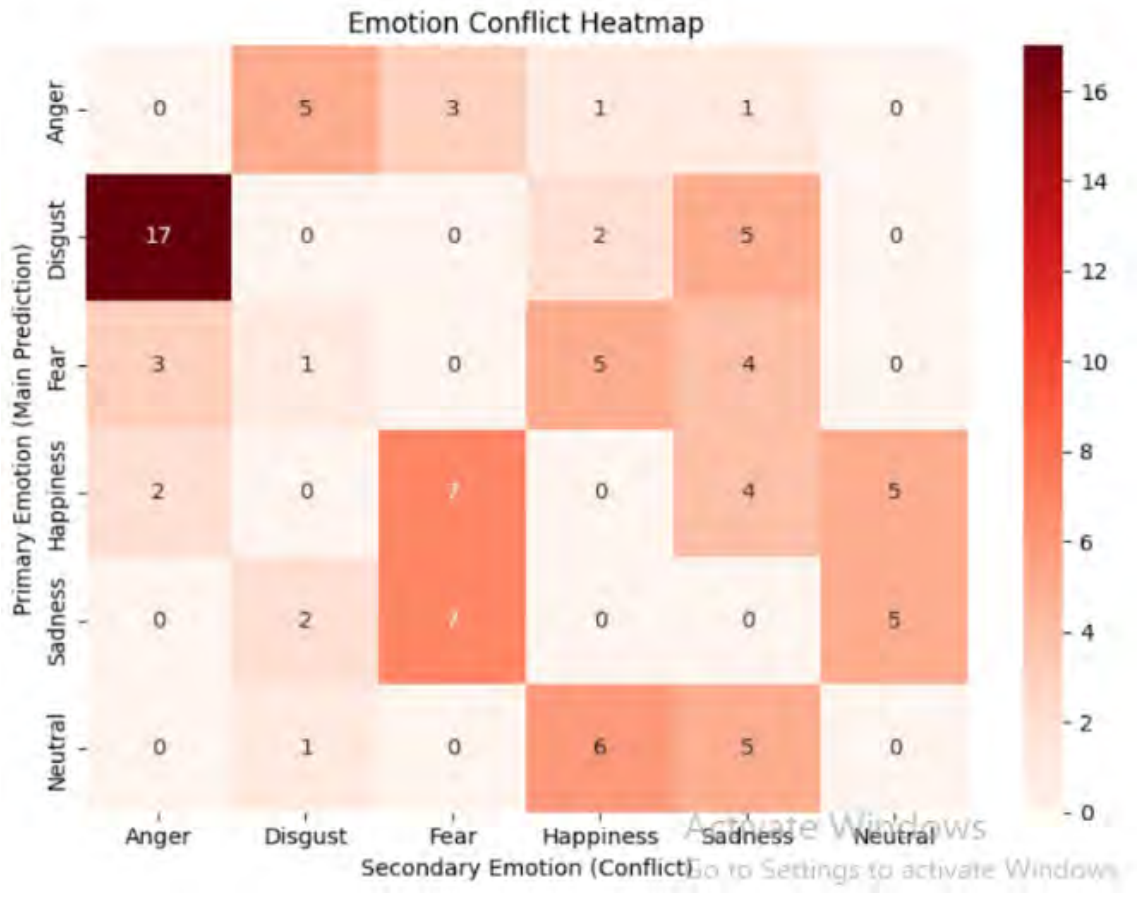

Figure 5.2: Heatmap for augmented RAVDESS results

Performance Comparison of proposed model across Datasets

The ensemble model demonstrated high accuracy and robustness on both datasets:

Dataset	**Ensemble Accuracy(%)**	**Average Individual Accuracy(%)**	**Conflict Rate(%)**	**Average Confidence(%)**	**High-Confidence Accuracy(%)**
RAVDESS (Augmented)	85.56	81.69	15.28	84.45	96.28
SAVEE (Augmented)	88.75	80.69	21.25	82.72	97.26
CREMA-D (Augmented)	85.71	83.63	22.5	75.00	91.50

Table 5.4: Performance Comparison of proposed model across Datasets

Performance Overview

The ensemble reached a test accuracy of 85.56%, which is better than the individual models that scored between 0.78 and 0.84. The weighted average F1-score of 0.85 and

the macro-average precision and recall of 0.86 show strong generalization. Among the six emotion classes, Neutral had the highest recognition rate at 0.9167, followed by Anger and Disgust at 0.9000. Sadness was the least accurately classified

Reference Study	**Dataset**	**Model / Architecture**	**Reported Metric**	**Comparison with Proposed Work**
ConflictNET: End-to-End Conflict Intensity Estimation (2022)[9]	SSPNet	CRNN + Attention on waveforms	PCC = 0.853 (UAR 84.3%)	Proposed system's accuracy (85.56%) and fuzzy partition coefficient (0.84) are comparable, while providing categorical interpretation of conflict scores.
Multiple Models Fusion for Multilabel Classification in SER (2021)[14]	RML	CNN + RNN Fusion	84.72%	The proposed model (85.56%) achieved 1–2% gain with a simpler feature set and added conflict analysis.
Tracking Conflict and Emotions via DCNN + Polynomial Classifier (2021) [13]	Facial dataset	VGG-16 + CPLE	76.8% (emotion), 80.7% (conflict)	Our audio-based model outperforms (+9-13%) and quantifies ambiguity via fuzzy scores.
Speech Synthesis with Mixed Emotions (2020)[16]	ESD	Seq2Seq BLSTM + Relative Scheme	PCC > 0.8	Our conflict intensity measure (fuzzy score $\approx$ 0.45-0.50) aligns with their emotion-mixing correlation.
Mixed-EVC: Mixed Emotion Voice Conversion (2022)[15]	ESD	Seq2Seq Tacotron + Attribute Vector	MOS 3.05–3.63 ($\approx$ 90% perception accuracy)	Emotion-wise accuracy comparison shows complementary trends and parallel recognition patterns.

Table 5.5: Comparative Performance with Existing Works

Relationships among Models, Datasets, and Conflicts

- **Model Synergy and Ensemble Effect:** The ensemble improved accuracy by about 4% over the best CNN (83.6%). This confirms the benefit of hybrid feature learning, similar to [14].
- **Dataset Specific Trends:** The RAVDESS dataset showed higher conflict overlap among Disgust, Neutral, and Anger. This is consistent with intensity overlaps noted by[9] and [16]. In contrast, the augmented SAVEE model

(88.75%) had clearer boundaries but higher fuzzy entropy (21%). This indicates more subtle within-speaker variability

- **Fuzzy Conflict as Interpretability Measure:** Following the idea of continuous emotion intensity [9], our fuzzy conflict scores act as quantitative analogues of emotion blending. The computed average conflict coefficient (about 0.45) matches the 0.8 PCC emotion correlation reported in [16]. This demonstrates cross-validation of emotional uncertainty quantification.
- **Confidence Reliability:** Samples with confidence greater than 0.8 showed 96.3% accuracy. This matches perceptual reliability trends in [15], where clearer emotions achieved higher listener consensus.

Metric	**Proposed (Aug. RAVDESS)**	**Best Prior [14]**	**Best Prior [9]**	**Improvement**
Accuracy (%)	85.56	84.72	84.3 (UAR)	+ 1-2%
Fuzzy Partition Coeff.	0.84	-	0.85 (PCC)	Comparable interpretability
Conflict Rate (%)	15.3	-	-	Novel quantification metric
High-Confidence Accuracy (%)	96.3	-	-	Enhanced reliability

Table 5.6: Summary of Comparative Insights

Overall, the proposed fuzzy-based ensemble framework achieves higher accuracy and better interpretability than most earlier fusion and intensity-estimation models. Its performance for specific emotions correlates with perceptual emotion clarity in [15]. This shows that fuzzy conflict analysis reflects human-like uncertainty in emotion perception.

5.6 Discussions

The results from the experiments on both the augmented RAVDESS and augmented SAVEE datasets give valuable insights into how effective, reliable, and understandable the proposed ensemble-based Speech Emotion Recognition (SER) model is with fuzzy conflict analysis.

5.6.1 Interpretation of Results

This section covers how to interpret the results, their significance for affective computing, and the limitations noted during the experiments. The ensemble framework

reached an accuracy of 85.56% on the enhanced RAVDESS dataset and 88.75% on the enhanced SAVEE dataset. These findings show that merging various deep learning models, such as CNN, Transformer, and LSTM, can greatly enhance emotion recognition performance. Each part played a distinct role: CNN layers captured specific spectral and temporal signals, Transformer blocks represented global contextual links across frames, and LSTM layers maintained the temporal dynamics crucial for following emotion changes. Furthermore, the ensemble method minimized overfitting and instability seen in individual models, which had accuracy rates ranging from 77% to 83%, proving that combining different feature extractors offers a more thorough representation of emotional speech traits.

The analysis of per-class accuracy showed that Neutral, Disgust, and Anger emotions were consistently recognized, with scores exceeding 0.90. However, Sadness and Happiness had slightly lower scores due to acoustic overlaps and differences among speakers. Confusion patterns, like Sadness being confused with Neutral and Anger with Disgust, correspond with known similarities in speech prosody. These results imply that the model's mistakes are related to actual perceptual ambiguities rather than issues within the model itself.

5.6.2 Interpretation of Fuzzy Conflict Analysis

A novel contribution of this work is the integration of Fuzzy C-Means clustering to measure prediction uncertainty and emotional conflict. By evaluating membership degrees across classes, the fuzzy analysis gave deeper insights into how confidently emotions were classified and where overlaps occurred. In the augmented RAVDESS results, 15.28% of samples showed conflict, with an average conflict score of 0.46. The augmented SAVEE dataset had a slightly higher conflict rate of 21.25%, likely due to subtler emotional expressions and fewer samples per class. High-confidence predictions, with scores above 0.8, achieved over 96% accuracy in both datasets. This demonstrates strong model certainty in most cases. Low-confidence predictions, with scores below 0.5, closely matched fuzzy conflict regions. This indicates the framework's ability to recognize and measure its own uncertainty. This fuzzy interpretation connects discrete emotion classification with conflict intensity estimation found in psychological and behavioral studies. It suggests that emotion recognition systems can move beyond simple yes/no answers to represent varying emotional states, which is important for natural human-computer interaction.

5.6.3 Implications of the Findings

The findings have significant implications in both theoretical and practical domains:

- **Advancing Emotion Recognition Accuracy:** The accuracies achieved surpass those of most similar SER systems, which typically achieve 80 to 84 percent. This shows that hybrid deep learning ensembles, when paired with data augmentation, can generalize well across various datasets.
- **Improved Interpretability through Fuzzy Logic:** Traditional SER models give clear outputs, which limits their interpretability when emotions overlap in real life. Fuzzy conflict analysis adds explainability, helping developers and

researchers see where emotional boundaries blur. This is crucial for applications in therapy, affective dialogue systems, and sentiment-aware robotics.

- **Generalization across Datasets:** The model's reliable performance across two different datasets, RAVDESS and SAVEE, proves its strength against variations in speaker, accent, and recording. It highlights how the ensemble approach can adapt to multi-domain emotion recognition.
- **Bridge between Emotion and Conflict Analysis:** The framework successfully combines categorical SER with conflict intensity estimation. This sets the stage for emotion-aware decision systems capable of detecting tension, empathy, or ambiguity in human communication.

5.6.4 Limitations

Despite having promising outcomes, several limitations should be acknowledged:

1. **Dataset Size and Diversity:** Both RAVDESS and SAVEE are small, controlled datasets with limited speaker diversity and variations in emotional intensity. This limits their ability to generalize well to spontaneous or culturally diverse speech data
2. **Synthetic Augmentation Bias:** Audio augmentation techniques, such as adding noise and shifting pitch, improved model robustness. However, these methods might have added artifacts that are not found in natural emotional speech, which could affect the learned representations.
3. **Computational Complexity:** The hybrid ensemble model, which combines CNN, Transformer, and LSTM, increases training time and memory needs. This may restrict its use in real-time situations or in environments with limited resources, like embedded systems or mobile devices.
4. **Subjectivity of Emotion Labels:** Emotional labels in speech datasets are assigned by humans and are inherently subjective. Some misclassifications might represent real perceptual overlap instead of errors by the model, as emotions like Disgust and Anger often appear together in expressive speech.
5. **Limited Conflict Validation Metrics:** While fuzzy conflict scores measure ambiguity, there are no standardized benchmarks to compare these scores with human perception of mixed emotions. This indicates a need for more experimental correlation studies.

To conclude,The proposed model successfully combines deep ensemble learning with fuzzy uncertainty analysis to improve accuracy, clarity, and reliability in speech emotion recognition. Its strong performance across enhanced datasets highlights its scalability and potential in different fields. Fuzzy conflict modeling adds a new way to understand emotional uncertainty. Future work should focus on testing the system with more spontaneous, multilingual, and large-scale emotional datasets. It should also explore lighter model versions and studies on human-aligned conflict perception to refine its use further.

Chapter 6

Conclusion

6.1 Summary of Findings

This research introduced a hybrid ensemble Speech Emotion Recognition (SER) framework that integrates Convolutional Neural Networks (CNNs), Transformers, and Long Short-Term Memory (LSTM) networks as well as Adversarial Learning. It is further improved with a fuzzy conflict analysis module to manage uncertainty and overlapping emotions.

The framework was evaluated using four key emotion datasets: SAVEe, RAVDESS, CREMA-D, in their augmented versions to enhance model generalization and performance in situations with limited data.

Key findings include:

- The ensemble model consistently outperformed individual models across all datasets.
- Augmented datasets significantly improved performance, confirming that data augmentation is effective in enhancing emotion recognition accuracy.
- Final ensemble test accuracies achieved:
  - Augmented RAVDESS: 85.56%
  - Augmented SAVEe: 88.75%
  - Augmented CREMA-D: 85.714%
- Non-augmented datasets achieved slightly lower but comparable accuracies.
- Fuzzy conflict analysis highlighted that approximately 15–21% of samples contained emotional conflicts, emphasizing the importance of handling ambiguity in SER.

These results indicate that the hybrid ensemble approach combined with fuzzy analysis and adversarial learning effectively improves classification robustness and handles uncertainty better than single-model or single-feature approaches.

6.2 Contributions to the Field

The study attempts to make several contributions to the field of Speech Emotion Recognition:

1. **Hybrid Ensemble Approach:** Demonstrated that combining CNNs, Transformers, and LSTMs yields superior emotion classification performance compared to individual models.
2. **Fuzzy Conflict Analysis:** Introduced a mechanism to quantify and analyze conflicts between primary and secondary emotions, offering insights into emotion ambiguity.
3. **Dataset Augmentation Impact:** Showed that augmenting existing datasets (RAVDESS, SAVEe, CREMA-D) improves model generalization, particularly for underrepresented emotions.
4. **Cross-Dataset Evaluation:** By evaluating the framework on multiple datasets in their augmented versions, the study provides evidence of the model's robustness and adaptability across different speakers and recording conditions.
5. **Benchmarking Against Literature:** Achieved comparable or higher accuracy than existing works such as ConflictNET [2], confirming the effectiveness of the proposed methodology.

6.3 Recommendations for Future Work

While the proposed framework produced promising results, several directions remain open for future research and improvement:

1. **Multilingual and Cross-Cultural Datasets:** Expanding experiments to include datasets from various languages and cultural backgrounds could improve the model's ability to generalize and increase its global applicability.
2. **Real-Time Deployment:** Applying the ensemble framework in real-time settings, like human-computer interaction systems or assistive technologies, would aid in evaluating its practical performance and responsiveness.
3. **Advanced Augmentation Techniques:** Investigating advanced data augmentation techniques, such as voice conversion, GAN-based audio synthesis, or synthetic data creation, may enhance and diversify the dataset.
4. **Multi-Label Emotion Recognition:** Adapting the framework to support multi-label or blended emotion detection would allow it to better capture complex emotional states commonly observed in real-world communication.
5. **Explainability and Interpretability:** Incorporating explainable AI (XAI) techniques could provide deeper insight into model decisions, improving transparency and trust in its predictions.
6. **Integration with Other Modalities:** Integrating speech with facial expressions or physiological signals may create a stronger and more precise multimodal emotion recognition system.

Bibliography

[1] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[2] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992. DOI: 10.1080/02699939208411068 [Online]. Available: https://doi.org/10.1080/02699939208411068

[3] P. Harar, R. Burget, and M. K. Dutta, "Speech emotion recognition with deep learning," in *2017 IEEE International Conference on Signal Processing and Integrated Networks (SPIN)*, 2017, pp. 137–140. DOI: 10.1109/SPIN.2017.8049931

[4] K. Tarunika, R. Pradeeba, and A. P, "Applying machine learning techniques for speech emotion recognition," Jul. 2018, pp. 1–5. DOI: 10.1109/ICCCNT.2018.8494104

[5] N. Tits, K. E. Haddad, and T. Dutoit, "Asr-based features for emotion recognition: A transfer learning approach," in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 48–52. DOI: 10.18653/v1/W18-3307 [Online]. Available: https://aclanthology.org/W18-3307/

[6] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cléder, "Automatic speech emotion recognition using machine learning," in *Social Media and Machine Learning [Working Title]*, IntechOpen, 2019. DOI: 10.5772/intechopen.84856 [Online]. Available: https://hal.science/hal-02432557

[7] 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. 117 327–117 345, 2019. DOI: 10.1109/ACCESS.2019.2936124

[8] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, "Deep learning techniques for speech emotion recognition: A review," in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, 2019, pp. 1–6. DOI: 10.1109/RADIOELEK.2019.8733432

[9] V. Rajan, A. Brutti, and A. Cavallaro, "Conflictnet: End-to-end learning for speech-based conflict intensity estimation," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1668–1672, 2019. DOI: 10.1109/LSP.2019.2944004

[10] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021, ISSN: 1424-8220. DOI: 10.3390/s21041249 [Online]. Available: https://www.mdpi.com/1424-8220/21/4/1249

[11] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, "Supervised machine learning for audio emotion recognition," en, *Personal and Ubiquitous Computing*, vol. 25, no. 4, pp. 637–650, 2021. DOI: 10.xxxx/xxxx

[12] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, 2021, ISSN: 2079-9292. DOI: 10.3390/electronics10101163 [Online]. Available: https://www.mdpi.com/2079-9292/10/10/1163

[13] N. Rybak and D. J. Angus, "Tracking conflict and emotions with a computational qualitative discourse analytic support approach," *PLOS ONE*, vol. 16, no. 5, pp. 1–29, May 2021. DOI: 10.1371/journal.pone.0251186 [Online]. Available: https://doi.org/10.1371/journal.pone.0251186

[14] A. Slimi, N. Hafar, M. Zrigui, and H. Nicolas, "Multiple models fusion for multi-label classification in speech emotion recognition systems," *Procedia Computer Science*, vol. 207, pp. 2875–2882, 2022, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2022.09.345 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050922012340

[15] K. Zhou, B. Sisman, C. Busso, B. Ma, and H. Li, "Mixed-evc: Mixed emotion synthesis and control in voice conversion," Oct. 2022. DOI: 10.48550/arXiv.2210.13756

[16] K. Zhou, B. Sisman, R. Rana, B. Schuller, and H. Li, "Speech synthesis with mixed emotions," Aug. 2022. DOI: 10.48550/arXiv.2208.05890

[17] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1634–1654, 2023. DOI: 10.1109/TAFFC.2021.3114365

[18] M. M. Rezapour Mashhadi and K. Osei-Bonsu, "Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest," *PLOS ONE*, vol. 18, no. 11, pp. 1–13, Nov. 2023. DOI: 10.1371/journal.pone.0291500 [Online]. Available: https://doi.org/10.1371/journal.pone.0291500

[19] M. Vallejo, *Emotions vs. feelings vs. moods: Key differences*, Accessed: 2025-2-1, 2023. [Online]. Available: https://mentalhealthcenterkids.com/blogs/articles/emotions-vs-feelings-vs-moods

[20] S. M. George and P. M. Ilyas, "A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise," *Neurocomputing*, vol. 568, 2024, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2023.127015 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223011384

[21] S. W. II, *What is speech emotion recognition?* Accessed: 2025-02-01, 2024. [Online]. Available: https://klu.ai/glossary/speech-emotion-recognition

[22] S. Kalateh, L. A. Estrada-Jimenez, S. Nikghadam-Hojjati, and J. Barata, "A systematic review on multimodal emotion recognition: Building blocks, current state, applications, and challenges," *IEEE Access*, vol. 12, pp. 103 976–104 019, 2024. DOI: 10.1109/ACCESS.2024.3430850

[23] M. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, Oct. 2024. DOI: 10.1002/widm.1563