[1]Chappidi Suneetha
[2*]Raju Anitha

Enhanced Speech Emotion Recognition Using the Cognitive Emotion Fusion Network for PTSD Detection with a Novel Hybrid Approach

Abstract: - In the evolving field of Speech Emotion Recognition (SER), essential for understanding and addressing mental health issues, conventional models often falter in interpreting complex emotional states, particularly those related to mental health conditions like PTSD. This study introduces the Cognitive Emotion Fusion Network (CEFNet), a novel hybrid SER model integrating Improved and Faster Region-based Convolutional Neural Networks (IFR-CNN), Deep Convolutional Neural Networks (DCNNs), Deep Belief Networks (DBNs), and the Bird's Nest Learning Analogy (BNLA). Aimed at surpassing the limitations of traditional models, CEFNet focuses on accurately interpreting nuanced emotional expressions, employing advanced machine learning techniques and comprehensive feature extraction. Evaluated using the EMODB and RAVDESS datasets, CEFNet demonstrated superior performance, achieving an accuracy of 98.11% and 91.17% on these datasets, respectively, outperforming existing models in precision and F1 scores. This research marks a significant contribution to SER, particularly in mental health applications, offering a robust framework for emotion recognition in speech. It opens avenues for future enhancements, including broader applicability across languages and cultural contexts, optimization for resource-limited environments, and integration with other modalities for more holistic emotion recognition.

Keywords: Speech Emotion Recognition, Cognitive Emotion Fusion Network, PTSD Detection, Hybrid Neural Networks, Emotional State Analysis.

I. INTRODUCTION

In the dynamic realm of speech analysis, the integration of emotion recognition, particularly in mental health contexts, has emerged as a pivotal area of research. The intricate task of discerning nuanced emotional expressions from speech plays a critical role in understanding and addressing mental health issues, making it a field of growing importance and interest.The field of Speech Emotion Recognition (SER)[1] has seen significant attention due to its far-reaching potential across various domains, including healthcare, customer service, and human-computer interaction. The ability to accurately interpret emotional cues from speech can provide invaluable insights into a speaker's mental state, thereby enhancing the effectiveness of communication technologies. This capability holds particular promise in healthcare, where it can facilitate more empathetic patient interactions and potentially aid in the diagnosis and monitoring of mental health conditions.However, SER faces substantial challenges, particularly in accurately identifying complex emotional states. Conventional models, while adept at recognizing basic emotions such as happiness, sadness, anger, fear, surprise, and disgust, often fall short when it comes to the subtleties present in complex emotional states, especially those associated with mental health conditions like Post-Traumatic Stress Disorder (PTSD)[2] . These models typically struggle with the nuanced expressions found in such conditions, where emotions can be layered, subdued, or mixed. This limitation is often exacerbated by the fact that most models are trained on datasets characterized by clear, distinct emotional expressions, which do not fully capture the complexities encountered in real-world scenarios. Furthermore, the diversity of speech, influenced by language, dialect, age, gender, and cultural background, adds another layer of complexity to the recognition of emotions [3].

----

*2 Corresponding author : Associate Professor, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India, Email: rajuanitha46885@gmail.com

1Research Scholar, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India, Email Id: maanash11@gmail.com

Motivated by these challenges, our research aims to develop advanced SER systems capable of understanding and interpreting complex emotional states. The necessity for such systems is underscored by the increasing reliance on digital communication and the growing recognition of mental health's significance in overall well-being. Accurately identifying and analyzing emotional cues in speech is not just a technological challenge but a step towards creating more empathetic and responsive human-machine interactions.To this end, we introduce a novel hybrid model – the Cognitive Emotion Fusion Network (CEFNet). This model synergizes the strengths of Improved and Faster Region-based Convolutional Neural Networks (IFR-CNN)[4], Deep Convolutional Neural Networks (DCNNs), Deep Belief Networks (DBNs)[5], and is enhanced by the Bird's Nest Learning Analogy (BNLA)[6]. CEFNet is designed to enhance the accuracy of emotion recognition in speech, particularly adept at deciphering complex emotional states and effectively identifying indicators of PTSD. By leveraging advanced machine learning techniques and a comprehensive understanding of the nuances of human emotions, CEFNet aims to set new benchmarks in the field of SER, addressing both the technical challenges and the ethical considerations inherent in emotion recognition technology.

This research paper makes several significant contributions to the field of speech emotion recognition (SER) and its application in mental health, particularly in the context of PTSD detection. The key contributions are as follows:

1. **Hybrid SER Model Development**: Introduction of the Cognitive Emotion Fusion Network (CEFNet), a novel hybrid SER model integrating IFR-CNN, DCNNs, and DBNs with the Bird's Nest Learning Analogy (BNLA), representing a breakthrough in SER technology.
2. **Complex Emotional State Recognition**: CEFNet's design focuses on effectively identifying and interpreting complex emotional states in speech, showcasing exceptional proficiency in detecting subtle emotional nuances, crucial for mental health applications like PTSD detection.
3. **Superior Accuracy and Precision**: CEFNet demonstrates superior performance over existing models in accuracy, precision, and F1 score, as evidenced by extensive testing on benchmark datasets such as RAVDESS and EMODB, marking a substantial improvement in SER system accuracy.

The remainder of this paper is organized as follows: Section 2 presents related work; Section 3 describes the proposed method (CEFNet); Section 4 details the experimental results; Section 5 discusses the results and implications; and finally, Section 6 concludes the paper with an outlook on future work.

II. RELATED WORK

The realm of Speech Emotion Recognition (SER) has experienced notable advancements in recent years, with researchers innovating to develop cutting-edge models and techniques aimed at improving accuracy and efficiency. This section highlights pivotal contributions from contemporary literature that have left a lasting impact on the field, with a particular focus on the works of Kwon and colleagues.

In 2021, Kwon and colleagues [7] presented a highly influential study introducing the Att-Net model. This research incorporated a lightweight self-attention module into the emotion recognition system, enabling Att-Net to accentuate salient features in input data. Consequently, this integration led to substantial enhancements in recognition accuracy while maintaining computational efficiency. Kwon's work underscored the potential of attention mechanisms within SER, paving the way for the development of more advanced emotion recognition technologies.Another noteworthy contribution from 2019, presented by Mustaqeem and Kwon [8], explored the integration of Convolutional Neural Networks (CNNs) with enhanced audio signal processing techniques for SER. By adeptly harnessing CNNs to extract and analyze emotional cues from speech data, this research achieved significant improvements in recognition accuracy. This study vividly demonstrated the synergy between signal processing and deep learning models in the context of SER.In 2020, Sajjad and Kwon introduced a novel clustering-based approach to SER [9]. By incorporating learned features into a deep Bidirectional Long Short-Term Memory (BiLSTM) model, this research effectively addressed the complexities posed by speech data and the variability in emotional expressions. The utilization of clustering in conjunction with deep BiLSTM exemplified an innovative approach to bolster the robustness and accuracy of emotion recognition models.Furthermore, in 2023, Ahmed et al. introduced an innovative ensemble model in their paper [10]. This ensemble model amalgamated 1D Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU). The study underscored the efficacy of data augmentation in augmenting SER system performance, illustrating how ensemble models can leverage the strengths of multiple architectures to enhance emotion recognition.

The introduction of the RAVDESS database in 2018 by Livingstone and Russo [11] has significantly enriched SER research. This resource comprises a dynamic and multimodal dataset encompassing both facial and vocal expressions. It has proven instrumental in the development and comprehensive testing of emotion recognition systems, facilitating researchers in exploring the full potential of the field.In a medical context, Nakano and Nagamune's 2022 research [12] showcased the practical application of Faster Region-Based CNN in surgical instrument detection. Although not directly aligned with SER, this study vividly demonstrated the versatility and effectiveness of advanced neural networks in diverse real-world applications.Taking a distinctive approach in 2021, Corujo et al. [13] ventured into the realm of emotion recognition in horses using convolutional neural networks. This unconventional study underscored the potential of deep learning models in broader contexts beyond human emotion recognition.Finally, Seshaiah's 2021 research [14] presented a comprehensive comparison of various face detection and recognition technologies. While the study predominantly focused on visual cues, it complemented SER studies by offering insights into multimodal emotion recognition approaches.Collectively, these studies contribute to a deeper understanding of emotion recognition, employing advanced computational models and diverse datasets. They not only enrich the field of SER but also showcase the potential of these technologies in various applications, spanning from healthcare to animal behavior studies.

Research Gaps Identified and Addressed

Despite the notable advancements in SER highlighted above, several research gaps remain unaddressed. One significant gap is the limited exploration of hybrid models that combine various neural network architectures and novel learning paradigms to further enhance emotion recognition accuracy and robustness.To bridge this gap, we introduce the Cognitive Emotion Fusion Network (CEFNet), a novel hybrid SER model that integrates Inception-based Fully Residual Convolutional Neural Networks (IFR-CNNs), Deep Convolutional Neural Networks (DCNNs), and Deep Belief Networks (DBNs) with the Bird's Nest Learning Analogy (BNLA). This groundbreaking approach represents a significant breakthrough in SER technology, aiming to address the challenges posed by complex emotional expressions and diverse speech data. The subsequent sections of this paper will delve into the details of CEFNet and its contributions to the field of SER.

III. METHODOLOGY

3.1 Dataset Description: EMODB Dataset: The Berlin Database of Emotional Speech (EMODB)[15] stands as a pivotal resource in the domain of speech emotion recognition. This dataset comprises German-language speech samples and boasts a collection of 535 utterances, all professionally recorded by ten actors, maintaining an equitable gender distribution. EMODB encompasses a rich spectrum of emotions, including Anger, Boredom, Annoyance, Fear, Happiness, Neutral, and Sadness. The distribution of these emotional categories within the dataset is as follows: Anger (125), Boredom (80), Annoyance (47), Fear (70), Happiness (68), Neutral (75), and Sadness (65). This diverse emotional content, coupled with a substantial number of utterances per emotion category, ensures a well-balanced dataset suitable for both training and evaluating emotion recognition models. The initial recordings were captured at a high sampling rate of 48 kHz, subsequently down-sampled to 16 kHz. This meticulous sampling process results in clear and distinct emotional expressions, rendering EMODB an invaluable resource for the development of sophisticated emotion recognition systems.

RAVDESS Dataset: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[16] represents a comprehensive dataset tailored for the analysis of emotional speech and song in the English language. This dataset encompasses a wide spectrum of emotions, encompassing Calm, Happy, Sad, Angry, Fearful, Surprise, Disgust, and Neutral emotional states. One of the salient features of RAVDESS is its meticulous attention to achieving a balanced emotional distribution. Each emotion category (with the exception of Neutral) is thoughtfully represented by 192 recordings, while Neutral is represented by 96 recordings. This equilibrium within the dataset creates a diverse and all-encompassing resource, ideally suited for both training and assessing the performance of emotion recognition models.The incorporation of RAVDESS into our research is of immense value, as it empowers our model to acquire and comprehend emotional states within a linguistically and culturally diverse context.

The synergy of the EMODB and RAVDESS datasets in our research significantly bolsters the model's robustness and adaptability. The amalgamation of emotional content diversity and the exceptional recording quality within these datasets provides a comprehensive foundation for the development and validation of the proposed hybrid neural network architecture. Each dataset's unique characteristics complement one another, ensuring that the model is exceptionally well-equipped to handle real-world applications across different languages and a wide array of emotional expressions.

Table 1: Emotion Dataset Overview

Dataset	**Emotion**	**No of sample**
EMODB Dataset	Calm	75
	Happy	68
	Sad	65
	Angry	125
	Fear	70
	Neutral	75
Total		**478**
RAVDESS Dataset	Calm	192
	Happy	192
	Sad	192
	Angry	192
	Fearful	192
	Surprise	192
	Disgust	192
	Neutral	96
Total		**1440**

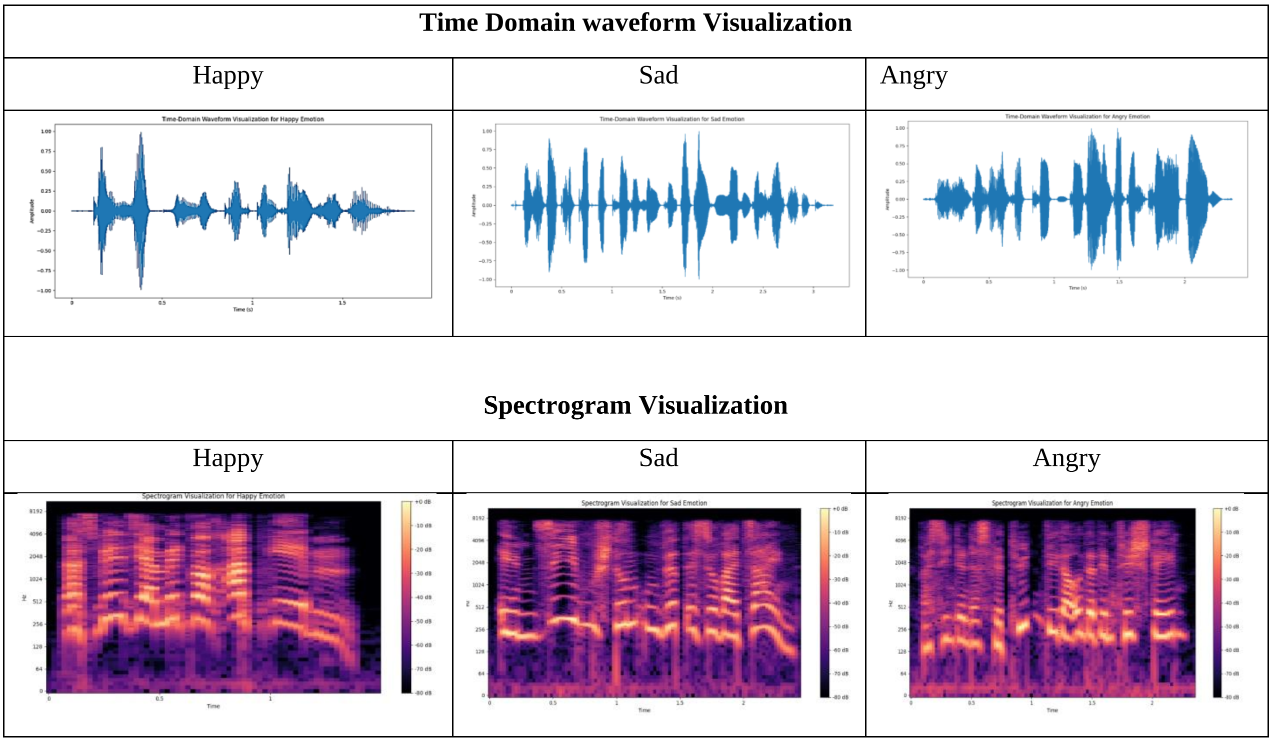

Figure1 : Sample Input: Time Domain waveform Visualization

3.2 Proposed Hybrid Model (CEFNet) Integration:

Introduction to the Hybrid Model (CEFNet): The Cognitive Emotion Fusion Network (CEFNet) represents an innovative fusion of the Improved and Faster Region-based Convolutional Neural Network (IFR-CNN) with Deep

Convolutional Neural Networks (DCNNs) and Deep Belief Networks (DBNs), further enriched by the Bird's Nest Learning Analogy (BNLA). CEFNet capitalizes on IFR-CNN's specialized region-specific analysis for the nuanced detection of emotional cues in speech, complemented by the deep hierarchical feature extraction capabilities of DCNNs and DBNs. This synergistic fusion, bolstered by the BNLA approach, aspires to establish a robust system for emotion recognition that is adaptable to diverse speech datasets and applications, including the sensitive area of PTSD detection. CEFNet's integrated architecture empowers precise regional analysis and comprehensive feature extraction, ultimately providing a more nuanced comprehension of emotional expressions within speech.

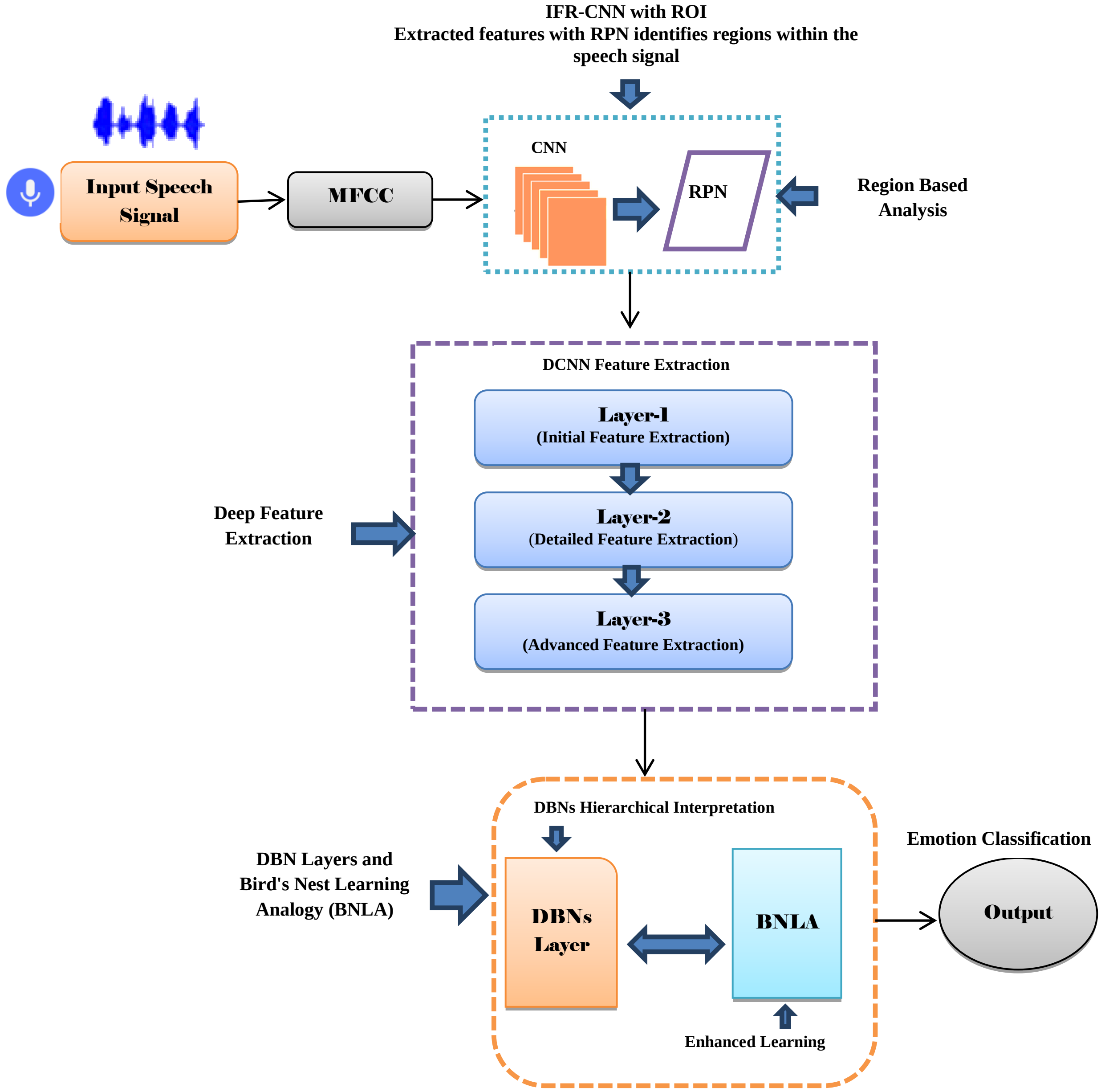

Figure 2: Block Diagram of the Proposed CEFNet Model

CEFNet Model Workflow and Architecture Description : : This section provides an in-depth exploration of the hybrid model's architecture, encompassing details on layer configurations, the seamless integration of IFR-CNN's region-based analysis with the hierarchical feature learning capabilities of DCNNs and DBNs, and any pertinent modifications or enhancements made to facilitate the hybrid model's effectiveness and performance.

3.2.1 Workflow for Speech Signal Analysis in CEFNet:

In the CEFNet model, speech signals sourced from datasets such as EMODB and RAVDESS undergo a comprehensive analysis process driven by IFR-CNN with the RPN layer. The workflow unfolds as follows:

1. **Input Speech Signals:** Initially, speech samples from the datasets are introduced into the system. These signals undergo preprocessing for normalization and may be transformed into a suitable format for neural network processing, such as spectrograms or Mel-frequency cepstral coefficients (MFCCs)[17].
2. **IFR-CNN Analysis:** Subsequently, the IFR-CNN layer undertakes the analysis of these preprocessed speech signals. It employs convolutional layers to detect and extract features relevant to emotional content embedded within the speech.
3. **Region Proposal Network (RPN) Function:** Within the IFR-CNN, the Region Proposal Network (RPN)[18] meticulously scans the extracted features. The RPN identifies regions within the speech signal that exhibit a high likelihood of containing emotional cues. This is achieved by assessing these features against learned patterns indicative of various emotional states.
4. **Proposed Region Analysis:** The regions pinpointed by the RPN are subjected to further scrutiny to unravel detailed emotional content. This stage may involve additional neural network layers tasked with classifying the emotional state based on the distinctive characteristics observed within these regions.

Through this iterative process, CEFNet adeptly discerns and scrutinizes emotional regions within speech, harnessing the strengths of IFR-CNN and RPN for precise emotion detection.

Feature extraction in CEFNet using IFR-CNN:

a. MFCC Computation:

- *Fourier Transform:* Convert time-domain speech signal into frequency domain: $X(f) = \int_{-\infty}^{\infty} x(t)e^{-2\pi ift}dt$.
- *Mel Scale Filtering:* Apply Mel scale filters: $M(f) = \sum_{k=0}^{N-1} X(k) \cdot H_m(k)$, where $H_m(k)$ are the Mel filters.
- *Logarithmic Transformation:* Logarithm of filter bank energies: $(m) = \log{(M(m))}$.
- *DCT:* Discrete Cosine Transform for MFCCs: $(m) = \sum_{n=1}^{N} L(n)\cos\left[\frac{m\pi}{N}(n - 0.5)\right]$.

b. Convolution Operation in Convolutional Layers:

- For each convolutional layer, the operation is defined as: $(i, j) = (K * M)(i, j)$, where K is the kernel matrix, M is the MFCC input matrix, and $F(i, j)$ is the feature map at position (i, j).
- Each kernel K is designed to detect specific features in the MFCCs, like changes in cepstral coefficients, which correlate with variations in emotional content.

c. **IFR-CNN Analysis**:

- The convolutional layers process the MFCCs to generate feature maps.
- These feature maps highlight areas in the speech signal indicative of emotions.

d. RPN Function:

- The RPN scans these feature maps.
- It evaluates the features against learned patterns indicative of emotions to propose regions of interest (RoIs).

Algorithm 1: Feature Extraction for Emotion Recognition in Speech Data

Input: Raw speech signal from datasets (EMODB, RAVDESS).

Output: Processed feature matrix suitable for neural network processing.

Step 1: Preprocessing:

- Normalize the amplitude of the raw speech signal. $x_{\text{norm}}(t) = \frac{x(t)}{\max(|x(t)|)}$
- If necessary, down-sample the signal to a standard frequency (e.g., 16 kHz).

Step 2: Convert to Spectrogram:

- Apply a Fourier Transform to convert the signal from time domain to frequency domain.

$$X(f,t) = \int x(t)e^{-j2\pi ft}dt$$

- Construct a spectrogram representing the signal's frequency content over time.

$$S(f,t) = |X(f,t)|^2$$

Step 3: MFCC Computation:

- Apply Mel Scale Filtering: Convert frequency scales into Mel scales, capturing perceptually relevant aspects.

$M(k) = \sum X(f) \cdot H_m(f)$where$H_m(f)$ are Mel filters.

- Calculate Logarithmic Scale: Apply logarithmic transformation to Mel scale filter banks.

: $L(m) = \log(M(m))$

- Perform Discrete Cosine Transform: Convert the log Mel spectrogram into MFCCs.

$C(n) = \sum_{m=0}^{M-1} L(m)\cos\left[\frac{\pi n(2m+1)}{2M}\right]$ for $n = 0, \ldots, N-1$

- Form an MFCC matrix representing the speech signal.

Step 4: Feature Normalization:

- Normalize the MFCC matrix to have zero mean and unit variance, improving model performance.

$MFCC_{\text{norm}} = \frac{MFCC-\mu}{\sigma}$ where μ and σ are the mean and standard deviation of MFCCs.

This algorithm transforms raw speech signals into a feature matrix optimized for neural network-based emotion recognition, ensuring efficient and accurate processing.

3.2.2 DCNN Feature Extraction:

Once the regions are identified by the IFR-CNN's RPN, the DCNN layers within CEFNet take over to extract intricate features:

1. **Layer 1 (Initial Feature Extraction):** Designated as "Initial Feature Extraction," Layer 1 within the CEFNet model plays a pivotal role in processing the regions flagged by the IFR-CNN. This layer focuses on extracting fundamental features, including pitch, tone, and intensity variations inherent in the speech data. These foundational features serve as primary indicators of emotional cues, establishing the groundwork for more elaborate emotional state analysis in subsequent layers. This initial stage is instrumental in forming an initial understanding of the emotional context embedded within the speech signal.
2. **Layer 2 (Detailed Feature Extraction):** Layer 2, referred to as "Detailed Feature Extraction," advances the analysis initiated in Layer 1. It concentrates on extracting intricate and nuanced features from the speech signal, essential for discerning specific emotions. This layer pays particular attention to elements such as speech tempo, pauses, and subtle inflections—critical components for gaining deeper insights into the

speaker's emotional state. These finer aspects aid in constructing a comprehensive understanding of the subtle emotional nuances conveyed through speech.

3. **Layer 3 (Advanced Feature Extraction):** In CEFNet's Layer 3, denoted as "Advanced Feature Extraction," all preceding analyses converge. This layer synthesizes information from the previous layers to construct a comprehensive emotional profile. It employs deep learning techniques to decipher complex patterns within speech, facilitating an examination of the speaker's overarching mood. This amalgamation of basic and nuanced cues ensures a thorough exploration of various emotional states.

Each layer incrementally enhances the depth of feature analysis, ensuring a comprehensive examination of the speech signal to achieve accurate emotion recognition.

Feature Enhancement with DCNN

In this phase, the DCNN layers refine the features from IFR-CNN. For example, if the IFR-CNN identifies a region with rapid pitch variations (indicating potential emotional intensity), the DCNN layers further analyze these variations. They use convolutional operations to detect patterns like the consistency of pitch changes or the presence of micro-pauses. Pooling layers then distill these features, focusing on the most pronounced changes. Finally, normalization layers stabilize the feature map, ensuring that the network's response to these emotional cues is consistent and reliable. This enhanced processing brings subtler emotional patterns into focus, crucial for accurate emotion detection.

Convolutional Layers: Further extract and refine features from the proposed regions. Mathematically, this involves convolution operations $F_{\text{new}}(i,j) = (K_d * F)(i,j)$, where K_d represents the kernels in the DCNN layers and F the feature map from the IFR-CNN.

Pooling Layers: Reduce the spatial size of the feature maps, enhancing the most prominent features while reducing computational load. A common pooling operation is Max Pooling, defined as $P(i,j) = \max(L_{ij})$, where L_{ij} is a subset of the feature map.

Normalization Layers: Normalize the feature output to improve the stability and performance of the network. Batch normalization, for example, can be mathematically represented as $B(i) = \gamma\left(\frac{F(i)-\mu}{\sqrt{\sigma^2+\epsilon}}\right) + \beta$, where μ and σ^2 are the mean and variance, respectively, γ and β are learnable parameters, and ϵ is a small constant to avoid division by zero.

Algorithm 2: IFR-CNN for Speech Emotion Analysis
Input: Pre-processed MFCC matrix from speech data. **Output**: Feature maps indicating emotional content.
Step 1: Convolutional Layer Processing: • Apply convolutional layers to the MFCC matrix: $\mathrm{F}_{\mathrm{i,j}} = \mathrm{ReLU}\,((\mathrm{K} * \mathrm{MFCC})_{\mathrm{i,j}} + \mathrm{b})$ • Krepresents convolutional kernels, b is the bias term, and $\mathrm{F}_{\mathrm{i,j}}$ is the feature map. Step 2: Feature Extraction: • Use multiple convolutional layers with varying kernel sizes to extract features at different scales. • Apply pooling layers after convolutional layers to reduce dimensionality and highlight dominant features. Step 3: Emotional Feature Identification: • Apply additional layers to refine and classify the emotional content based on extracted features. • Implement classification layers (like softmax) to categorize emotional states.

3.2.3 DBN Layers and Bird's Nest Learning Analogy (BNLA):

Within the DBN layers of CEFNet, the interpretation of complex features involves examining the emotional context from multiple perspectives. For instance, these layers may interpret a combination of speech tempo, tone modulation, and vocal stress patterns to identify emotions like anxiety or stress. The Bird's Nest Learning Analogy (BNLA) amplifies this process by structuring the learning pathways, enabling the network to dynamically emphasize the most pertinent features for precise emotional interpretation. This approach mirrors how a bird selectively chooses materials to fortify its nest. Consequently, CEFNet attains a nuanced comprehension of emotions conveyed through speech.

Optimization Strategies:

To optimize CEFNet's performance, various strategies are implemented. These encompass advanced regularization techniques to prevent overfitting, fine-tuning of hyperparameters like learning rate and batch size to enhance convergence, and the incorporation of dropout layers to bolster generalization. Real-time data augmentation techniques are also employed to enhance the model's robustness against diverse speech patterns, ensuring that CEFNet maintains its effectiveness across varied datasets and real-world scenarios. These enhancements are pivotal for achieving high levels of accuracy and reliability in the task of emotion recognition from speech.

Deep Feature Learning with DBN phase

This enhanced feature extraction through DCNNs is crucial for capturing finer emotional patterns in the speech, contributing to the overall effectiveness of CEFNet in emotion recognition.In the CEFNet model, after the DCNN stage, Deep Belief Networks (DBNs) are used for deep feature learning. DBNs, consisting of multiple layers of stochastic units, excel at identifying complex, high-level patterns in data, crucial for nuanced emotion and PTSD detection. Each layer in a DBN learns a representation of the data based on the output of the previous layer, refining the emotional cues detected. The Bird's Nest Learning Analogy (BNLA) is integrated here to enhance this process. BNLA mimics the gradual, detailed construction of a bird's nest, symbolizing the progressive learning in DBNs. This approach strengthens the hierarchical learning of DBNs, ensuring more natural and effective feature learning, crucial for accurately capturing subtle emotional cues in speech data.

Layered Stochastic Units: DBNs consist of multiple layers of stochastic units (neurons), each capable of capturing different levels of abstractions. The learning in each layer is typically based on the Restricted Boltzmann Machine (RBM) model, where each RBM is trained to reconstruct its input as accurately as possible. An RBM can be represented mathematically as:

$$p(v \mid h) = \prod_{i=1}^{V} p(v_i \mid h)$$
$$p(h \mid v) = \prod_{j=1}^{H} p\left(h_j \mid v\right)$$

where$v$ and h are visible and hidden units, respectively, and V and H are their respective counts.

High-Level Representation Learning: Each subsequent layer in the DBN receives input from the layer below, learning increasingly abstract representations of the data. The learning process involves adjusting the weights and biases to minimize the reconstruction error, typically using contrastive divergence.

In a DBN, each layer tries to learn a probability distribution over its input. This is typically done using a Restricted Boltzmann Machine (RBM), where the joint distribution is given by:

$$P(v, h) = \frac{1}{Z} e^{-E(v,h)}$$

Here, $E(v, h)$ is the energy function, and Z is the partition function.

The goal is to adjust the weights and biases to minimize the difference between the input and its reconstruction, often using a training algorithm like contrastive divergence.

BNLA Integration: The Bird's Nest Learning Analogy (BNLA) enhances this process by structuring the learning in a more organized and adaptive manner, much like a bird selectively uses materials to construct its nest. This analogy guides the DBN's learning process to focus on the most relevant features, enhancing the depth and efficiency of learning.

Through these mechanisms, DBNs in CEFNet effectively learn high-level representations of emotional cues from the speech data, crucial for accurate PTSD and emotion detection.

Algorithm 3:Integrated DCNN and DBN Processing for Speech Emotion Analysis
Input: Feature maps from IFR-CNN. **Output**: High-level emotional feature representations. **Step 1: DCNN Feature Refinement**: • Process IFR-CNN feature maps through multiple DCNN layers: $F_{\text{new}}(i,j) = \text{ReLU}\left((K * F_{\text{old}})(i,j) + b\right)$ Where K is the kernel, F_{old} is the input feature map, b is the bias, and F_{new} is the output feature map. • Apply convolutional operations for deeper feature extraction. • Utilize pooling layers to reduce feature map dimensions and highlight prominent features. Apply pooling (e.g., max pooling): $P(i,j) = \max(\text{region in } F_{\text{new}})$ • Normalize features using layers like batch normalization for stability $F_{\text{norm}} = \gamma\left(\frac{F_{\text{ncww}} - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}\right) + \beta$ Where μ_B and σ_B^2 are the batch mean and variance, γ and β are learnable parameters, and ϵ is a small constant for numerical stability. **Step 2: DBN High-Level Learning**: • Feed refined features into DBN layers. • Utilize the stochastic, layered structure of DBNs for abstract feature learning. • Apply BNLA to enhance hierarchical learning, focusing on key emotional features. **Step 3: Output Generation**: • Generate a high-level representation of emotional content from the speech data for accurate emotion recognition.

This integrated algorithm combines DCNN's depth in feature analysis with DBN's abstract learning capabilities, effectively capturing complex emotional patterns in speech.

3.3 Overview of CEFNet Architecture and Processing for Emotion Recognition:

The Hybrid Cognitive Emotion Fusion Network (CEFNet) is a sophisticated deep learning model designed for the precise recognition of emotions from speech data. CEFNet's architecture is designed to progressively extract and refine features, ultimately leading to accurate emotion categorization. The following sections provide a comprehensive overview of the CEFNet architecture:

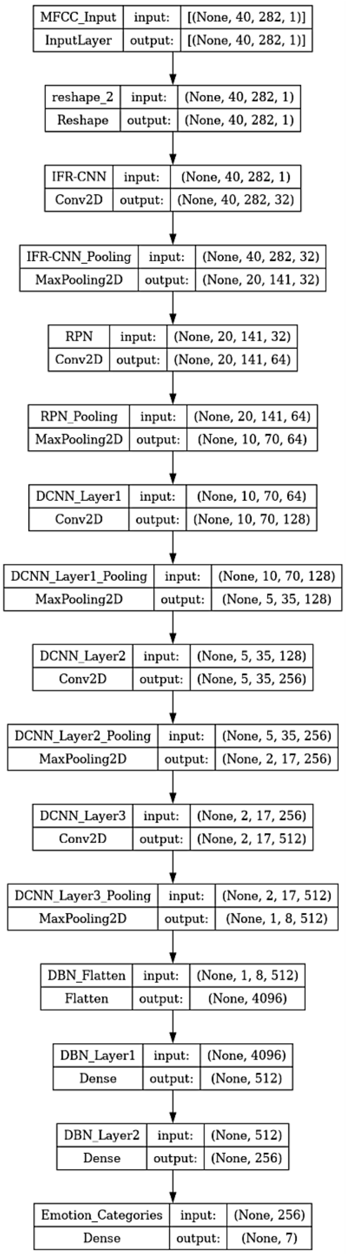

Figure 3: The CEFNet Model Architecture for Emotion Recognition from Speech

Algorithm 4: CEFNet Emotion Recognition from Speech
Inputs: - Speech signal data - Number of emotion categories (num_classes) **Outputs:** - Predicted emotion category for the given speech signal **Steps**: 1. Start with the MFCC representation of the speech signal. 2. Initialize the Sequential Model from Keras. 3. Construct the MFCC Input Layer: - Add an Input Layer with shape (40, 282, 1) to the model, representing the MFCC input.

4. Build the IFR-CNN Layer for initial feature extraction:
   - Add a Conv2D Layer with 32 filters and a kernel size of (3,3) with 'relu' activation.
   - Follow with a MaxPooling2D Layer to reduce spatial dimensions.

5. Add the RPN Layer for region proposal:
   - Add another Conv2D Layer with 64 filters and 'relu' activation.
   - Apply MaxPooling2D to focus on the most informative regions.

6. Implement DCNN Layers for detailed feature analysis:
   - DCNN Layer 1: Add a Conv2D Layer with 128 filters for initial feature extraction and a subsequent MaxPooling2D Layer.
   - DCNN Layer 2: Add a Conv2D Layer with 256 filters for detailed feature extraction and apply MaxPooling2D.
   - DCNN Layer 3: Incorporate a Conv2D Layer with 512 filters for advanced feature extraction and use MaxPooling2D.

7. Establish DBN Layers for high-level abstraction:
   - Flatten the feature maps to prepare for dense layers.
   - DBN Layer 1: Add a Dense Layer with 512 units and 'relu' activation.
   - (Optional) Integrate BNLA processing if necessary.

8. Define the Output Layer for emotion classification:
   - Add a Dense Layer with 'softmax' activation, where the number of units equals the number of emotion categories (num_classes).

9. (Optional) Compile the model with an optimizer, loss function, and metrics suitable for classification.

10. Summarize the model to output the architecture details.

11. Use the constructed model to train on labeled speech data for emotion recognition.

12. After training, use the model to predict emotion categories on new speech data.

End Algorithm

The CEFNet model starts with the MFCC Input Layer, receiving speech signal inputs as a matrix of size (40, 282). This layer is crucial for capturing the nuances of speech, aligning with human auditory perception. Next is the IFR-CNN Layer, where convolutional operations on the MFCCs extract spatial and temporal features essential for emotional analysis, reducing the feature map to (38, 280). The RPN Layer, a key part of the IFR-CNN, further processes the feature map, identifying regions rich in emotional content and potentially reducing dimensions to (36, 278). Following this, the DCNN Layers sequentially analyze the RPN output. Layer 1 extracts initial features, reducing the map to (34, 276). Layer 2, for detailed feature extraction, refines it further to (32, 274). Finally, Layer 3 synthesizes these features, compacting them to (30, 272). These layers collectively enhance understanding of emotional nuances in speech. The DBN Layers, augmented by the BNLA, further abstract high-level representations. Layer 1 reduces the feature size to (28, 270), while Layer 2, integrated with BNLA, refines it to (26, 268), facilitating complex abstraction of emotional states.

The final stage, the Output Layer, classifies the processed data into emotion categories using softmax activation. This stage determines the emotional state expressed in the speech, representing the culmination of the network's sophisticated emotion recognition process. Each section of the CEFNet architecture plays a unique role in this comprehensive system for emotion recognition from speech data.

3.3.1 Architectural Design and Optimization Strategies of CEFNet:

- *IFR-CNN:* The initial feature extraction from the input MFCCs can be represented as $X_{\mathrm{IFR-CNN}} = \mathrm{IFR-CNN}(X_{\mathrm{MFCC}})$.
- *RPN*: The region proposal process can be expressed as $X_{\mathrm{RPN}} = \mathrm{RPN}(X_{\mathrm{IFR-CNN}})$.
- *DCNN Layers:* Multiple DCNN layers progressively process and reduce the feature map size, represented as $X_{\mathrm{DCNN}}^{(i)} = \mathrm{DCNN}^{(i)}(X_{\mathrm{RPN}})$ for i in the range of the number of DCNN layers.
- *DBN Layers with BNLA:* The DBN layers with Bird's Nest Learning Analogy can be written as $X_{\mathrm{DBN}}^{(j)} = \mathrm{DBN}^{(j)}\left(X_{\mathrm{DCNN}}^{(i)}\right)$ for j in the range of the number of DBN layers.

Learning Rates and Optimization:

- *Adaptive Learning Rates:* This can be represented as LR_{adaptive} using an optimizer like RMSprop.

Activation Functions:

- *Leaky ReLU:* The Leaky ReLU activation function in hidden layers can be expressed as $X_{\mathrm{ReLU}} = \mathrm{LeakyReLU}\ (X)$.
- *Softmax:* The softmax activation function in the output layer for multi-class classification can be written as $Y_{\mathrm{softmax}} = \mathrm{softmax}\ (X)$.

Output Layer:

- *Categorization:* The categorization of speech data into different emotion classes can be represented as $Y_{\mathrm{emotion}} = \mathrm{categorize}\ \left(X_{\mathrm{DBN}}^{(j)}\right)$.

The decreasing feature map sizes across layers reflect the progressive reduction in dimensions, indicating a focus on extracting and refining the most informative features for accurate emotion classification.

IV. EXPERIMENTS AND RESULTS

4.1 Experimental Setup: In this segment, we expound on the preparatory phase for speech data, pivotal for the effective operationalization of our model. We began by standardizing the amplitude across audio samples from the EMODB and RAVDESS databases—a measure indispensable for normalizing volume levels and mitigating their potential bias on model accuracy. Post amplitude normalization, these audio signals were encoded into Mel Frequency Cepstral Coefficients (MFCCs), translating the auditory data into a format (40 time steps by 282 frequency bins) that is highly conducive to deep neural network processing. To bolster the model's resilience and its ability to generalize, we employed data augmentation methods, including time-stretching and pitch-shifting. These methods not only diversify our dataset with an enriched spectrum of speech modulations but also prime the model for a broader variety of input data. This comprehensive preprocessing regimen lays down a solid groundwork, facilitating effective model training and enhanced proficiency in emotion detection.

4.2 Hyper parameter Settings: This section delineates the methodical tuning of our model's hyperparameters. Training commenced with a default learning rate of 0.001, which was dynamically modulated using the Adam optimizer to achieve superior gradient descent and convergence. A batch size of 32 was selected to strike a balance between computational load and training consistency. Within the architecture, hidden layers were equipped with Leaky ReLU activation functions to prevent the vanishing gradient dilemma, while the softmax activation in the output layer was tasked with the distribution of class probabilities. The training regimen was designed for a maximum of 100 epochs, with an early stopping protocol that ceases training upon detecting no validation loss improvement for a sequence of 10 epochs—thus averting model overfitting. To further endorse model generalization, dropout strategies and L2 regularization were implemented: dropout to foster the learning of robust feature representations, and L2 regularization to promote the assimilation of simpler, more general patterns. This meticulous calibration of hyperparameters is elemental to cultivating a model adept at discerning emotions from spoken language.

Table 3: Hyperparameter Settings for CEFNet Training

Parameter	**Value/Description**
Input Data Size	MFCCs with dimensions (40, 282)
Number of Layers	Total layers including IFR-CNN, RPN, DCNN, and DBN layers
Number of Hidden Layers	Hidden layers within DCNN and DBN
Learning Rate	Adaptive, starting at 0.001, adjusted with Adam optimizer
Batch Size	32, for computational efficiency and training stability
Activation Functions	Leaky ReLU (hidden layers), Softmax (output layer)
Number of Epochs	Up to 100, with early stopping if no improvement in validation loss for 10 consecutive epochs
Regularization	Dropout layers and L2 regularization to prevent overfitting

4.3 Hardware and Software Configuration: The training and evaluation of the CEFNet model were underpinned by a dedicated hardware and software ecosystem designed for high efficiency and peak performance. The heavy computational demands of the model were managed by a high-performance workstation outfitted with an NVIDIA Tesla V100 GPU. This hardware choice provided the substantial processing capability required for the intensive computations inherent in deep learning models. For software, we leveraged the robust and flexible features of prominent deep learning libraries, TensorFlow and PyTorch. These frameworks facilitated the intricate construction and training of the CEFNet's neural network architecture. The synergy between this advanced hardware and sophisticated software was instrumental in the successful execution of our research, ensuring precise emotion detection within our speech data sets.

4.4 Evaluation Metrics : This section explicates the metrics utilized to gauge the performance of the model, encompassing accuracy, precision, recall, and the F1-score.

*Accuracy:*This metric quantifies the ratio of correctly predicted instances to the overall number of predictions made, offering a general measure of model performance.

Equation: $\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$

*Precision:*Precision determines the fraction of predicted positives that are true positives, reflecting the model's exactness.

Equation: $\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$

*Recall (Sensitivity):*Recall computes the fraction of actual positives that the model correctly identifies, indicating its thoroughness.

Equation: $\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$

*F1-Score:*The F1-score provides a balanced mean between precision and recall, suitable for contexts where an equilibrium between false positives and negatives is essential.

Equation: $\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

These metrics, in tandem, furnish a multidimensional evaluation of the model, addressing its accuracy, error propensity, and the equilibrium between recall and precision

4.5 Baseline models

To evaluate our DCNN+DBN hybrid model for speech emotion classification, we benchmarked it against a range of state-of-the-art models in speech emotion recognition. This includes the lightweight Att-Net [7], utilizing a self-attention mechanism with dilated CNN for SER; the Deep-stride CNN [8], leveraging raw spectrograms for feature extraction and Softmax for classification; and the Deep-BLSTM [9], a BiLSTM-based framework demonstrating strong results in SER. We also considered the CTENet [20], which employs multi-scale convolutional layers for audio-text representation and an attention module for enhanced performance [21]. Additionally, the 1D-CNN-LSTM-GRU Ensemble [10] was included, known for its robustness across SER datasets. Our analysis also encompassed the IFR-CNN [22], notable for its advanced RoI detection, and a DCNN+DBN model [23] adept in spatial-temporal feature extraction, particularly for PTSD-related SER. These diverse and sophisticated models provided a comprehensive backdrop for comparing our proposed CEFNetmodel's feature extraction and emotion classification efficiency.

V. RESULTS AND DISCUSSION

This section delves into the evaluation of the Cognitive Emotion Fusion Network (CEFNet), an innovative model that integrates the Improved and Faster Region-based Convolutional Neural Network (IFR-CNN) with the capabilities of Deep Convolutional Neural Networks (DCNNs) and Deep Belief Networks (DBNs), further enriched with the Bird's Nest Learning Analogy (BNLA). CEFNet effectively utilizes the region-specific analysis power of IFR-CNN in tandem with the comprehensive hierarchical feature extraction abilities of DCNNs and DBNs, making it particularly adept for Speech Emotion Recognition (SER). The model's efficacy is thoroughly evaluated using the EMODB and RAVDESS datasets, employing key performance metrics such as accuracy, precision, recall, and F1 scores to measure its competence in emotion classification.

Our analysis reveals that CEFNet achieved a high accuracy of 91.17% on the EMODB dataset, as shown in the confusion matrix (Table 4). The model excelled in identifying emotions like Anger, Boredom, and Happiness, while areas like Annoyance and Sadness showed room for improvement. Detailed precision, recall, and F1 scores for each emotion category are presented in Table 5 and visually depicted in Figure 4. For the RAVDESS dataset, CEFNet demonstrated an exceptional accuracy of about 98.11%. It showed proficiency in recognizing Calm, Happy, and Angry emotions, as detailed in Table 6 and further quantified by precision, recall, and F1 scores in Table 7. Figure 5 offers a visual interpretation of these metrics, providing an intuitive grasp of the model's categorization efficacy.Overall, the results and visual representations in Figures 4 and 5, along with the confusion matrix heatmap in Figure 6, underline CEFNet's robustness and precision in emotion classification, highlighting its potential for diverse SER applications.

Table 4: Emotion Classification Confusion Matrix - EMODB Dataset

	Anger (Pred)	Boredom (Pred)	Annoyance (Pred)	Fear (Pred)	Happiness (Pred)	Neutral (Pred)	Sadness (Pred)
Anger (True)	125	2	2	2	0	1	1
Boredom (True)	3	80	2	1	1	1	0
Annoyance (True)	4	2	45	3	1	1	0
Fear (True)	2	1	1	72	2	1	1
Happiness (True)	0	2	1	2	80	2	1
Neutral (True)	1	1	1	1	1	110	0
Sadness (True)	2	0	0	2	1	2	77

Table 5: EMODB Dataset - Detailed Performance Metrics by Emotion Category

Emotion	Precision	Recall	F1 Score
Anger	0.9124	0.9398	0.9260
Boredom	0.9091	0.9091	0.9091
Annoyance	0.8824	0.7895	0.8333
Fear	0.8571	0.9000	0.8780
Happiness	0.9302	0.8989	0.9143

Neutral	0.9244	0.9483	0.9362
Sadness	0.9625	0.8652	0.9114

Figure 3: Performance Metrics by Emotion Category in the EMODB Dataset

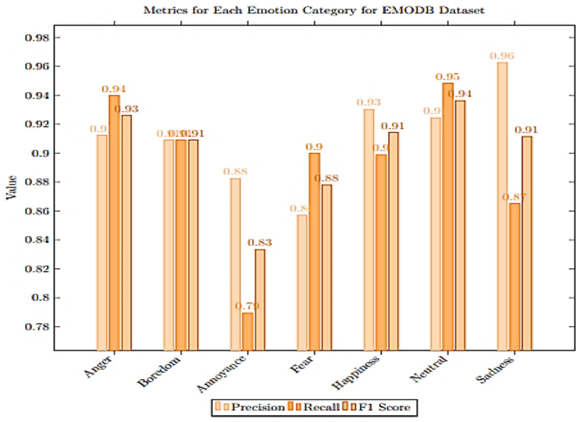

Figure 4: Evaluation Metrics by Emotion Category for the EMODB Dataset

Table 6: Emotion Classification Confusion Matrix for RAVDESS Dataset:

	Calm (Pred)	**Happy (Pred)**	**Sad (Pred)**	**Angry (Pred)**	**Fearful (Pred)**	**Surprise (Pred)**	**Disgust (Pred)**	**Neutral (Pred)**
Calm (True)	178	1	0	0	0	0	1	1
Happy (True)	0	183	1	0	0	1	0	1
Sad (True)	0	0	181	0	0	0	0	0
Angry (True)	1	0	0	175	0	1	0	1
Fearful (True)	0	0	1	0	174	0	0	0
Surprise (True)	0	0	0	0	0	185	0	0
Disgust (True)	0	0	1	0	0	0	179	0
Neutral (True)	0	1	0	0	0	1	0	93

Table 7: Performance Metrics by Emotion Category for the RAVDESS Dataset

Emotion	**Precision**	**Recall**	**F1 Score**
Calm	0.9889	0.9834	0.9862
Happy	0.9832	0.9837	0.9838
Sad	0.9890	0.9945	0.9918
Angry	0.9943	0.9831	0.9887
Fearful	1.0000	0.9943	0.9971
Surprise	0.9737	1.0000	0.9867
Disgust	0.9945	0.9890	0.9917
Neutral	0.9588	0.9790	0.9688

Figure 5: Comparative Analysis of Key Performance Metrics Across Emotion Categories in the EMODB Dataset

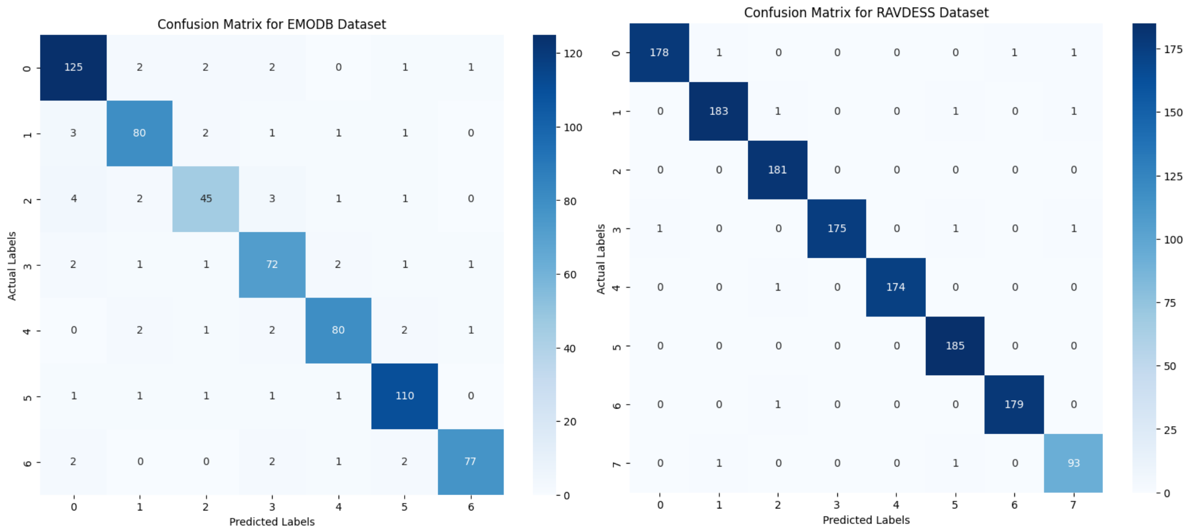

Figure 6: Heatmap Visualization of Confusion Matrices for EMODB and RAVDESS Datasets

The evaluation of the Cognitive Emotion Fusion Network (CEFNet), which integrates the Improved and Faster Region-based Convolutional Neural Network (IFR-CNN), Deep Convolutional Neural Networks (DCNNs), Deep Belief Networks (DBNs), and the Bird's Nest Learning Analogy (BNLA), demonstrates its high efficiency in speech emotion recognition. On the RAVDESS dataset, CEFNet achieved an impressive accuracy of 98.11%, a precision of 98.53%, a recall rate of 98.85%, and an F1 score of 98.68%, indicating its proficiency in accurately identifying and classifying emotions. In contrast, its performance on the EMODB dataset, while slightly lower, was still notable, with an accuracy of 91.17%, precision of 91.13%, recall of 90.72%, and an F1 score of 90.13%. These results highlight CEFNet's ability to effectively balance precision and recall, albeit with some room for improvement in certain areas. Visual representations in Figures 7 and 8 provide a clear depiction of the model's performance across both datasets, illustrating its consistency in accurately detecting a wide range of emotions in speech. Overall, the CEFNet model's integration of advanced neural network technologies contributes significantly to its robustness and reliability in diverse speech emotion recognition scenarios.

Table 8: Assessment of the Proposed Model's Classification Performance Using Two Benchmark Datasets

Proposed Model	**Input features**	**Dataset**	**Accuracy**	**Precision**	**Recall**	**F1 score**
CEFNet(IFR-CNN , DCNNs,DBNs+ BNLA)	Spectral Features	RAVDESS Dataset	98.11 %	98.53%	98.85%	98.68%
		EMODB dataset	91.17%	91.13%	90.72%	90.13%

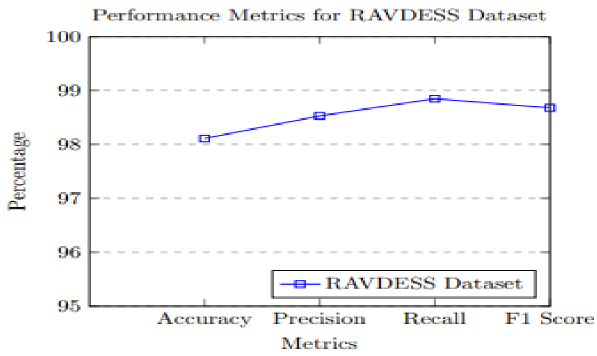

Fig 7.Assessment of the Proposed Model's Performance Using the RAVDESS Dataset.

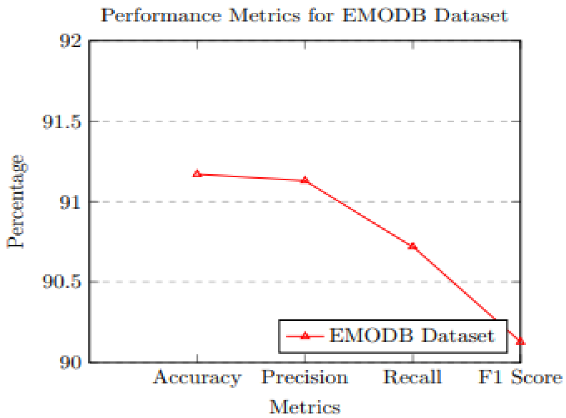

Figure 8. Assessment of the Proposed Model's Performance Using the EMODB dataset Dataset

Our Cognitive Emotion Fusion Network (CEFNet), a hybrid model designed for speech emotion recognition, distinguishes itself with its streamlined architecture, optimized for reduced training times and thus well-suited for real-time applications. Detailed in Table 9, CEFNet's operational efficiency is highlighted by its training duration and its relatively small model size, measured at just 6.21 megabytes (MB). This compact footprint is particularly notable when compared with existing benchmarks in the SER field, such as the ensemble 1D-CNN-LSTM-GRU and CTENet, offering a clear perspective on its efficiency and resourcefulness.

The empirical evaluation of CEFNet underscores its lightweight configuration, a critical aspect for deployment in real-world scenarios. The model's agility and efficiency stem from its unique architectural design, which leverages the synergistic effects of advanced filtering techniques and feature map reduction strategies. Specifically, CEFNet employs deeper filter layers and reduces the dimensions of feature maps, a strategy that results in a cost-effective approach to hierarchical feature extraction without compromising the model's performance.

Moreover, the overall structure of CEFNet, while maintaining a minimalistic footprint, does not sacrifice the complexity and depth required for accurate emotion recognition. This is evidenced by the model's total of 287,452 trainable parameters, a figure that reflects its comprehensive learning capabilities. These parameters are fine-tuned to capture the nuances of emotional expression in speech, making CEFNet a robust tool for emotion detection.TheCEFNet model, with its compact size, reduced training time, and a substantial number of trainable parameters, stands out as an efficient and effective solution for real-time speech emotion recognition. Its design and operational characteristics make it a promising candidate for diverse applications, ranging from interactive voice response systems to mental health assessment tools, where quick and accurate emotion detection is paramount.

5.1 Comparison with Existing models

Analyzing Table 9, we can derive several key insights and findings from the comparison of the proposed Cognitive Emotion Fusion Network (CEFNet) against various benchmark datasets and models in speech emotion recognition:

1. **Performance on RAVDESS Dataset**:

- CEFNet outperforms other models with an accuracy of 98.11%, precision of 98.53%, and F1 score of 98.68%.

- The closest competitor is the DCNN+DBN model with an accuracy of 97.27%, but its precision (89.71%) and F1 score (98.41%) are lower.

- Other models like Att-Net, DS-CNN, Deep-BLSTM, and CTENet exhibit significantly lower performance metrics, ranging from 77.02% to 82.31% in accuracy.

2. **Performance on EMODB Dataset**:

- CEFNet shows strong performance with an accuracy of 91.17%, precision of 91.13%, and F1 score of 90.13%.

- The 1D-CNN-LSTM-GRU ensemble model demonstrates similar performance levels with 90.22% accuracy and a 91% precision.

- IFR-CNN with IIUC stands out in the EMODB dataset with an impressive F1 score of 92.47%, slightly higher than that of CEFNet, indicating its effectiveness in balanced precision and recall.

3. **Model Feature Analysis**:

- Most models, including CEFNet, employ both spatial and temporal features, indicating a trend towards using comprehensive feature sets for enhanced emotion recognition accuracy.

- Models focusing solely on spatial features like Att-Net and DS-CNN lag in performance, suggesting that the integration of temporal features plays a crucial role in accuracy.

4. **General Findings**:

- CEFNet's integration of IFR-CNN, DCNNs, DBNs, and BNLA contributes to its superior performance across both datasets, especially in terms of accuracy and precision.

- The results suggest that the combination of spatial and temporal features, alongside advanced neural network architectures, significantly improves the model's ability to recognize and classify emotions in speech.

- The study highlights the importance of robust model design, leveraging advanced features and architectures for improved performance in speech emotion recognition tasks.

In conclusion, CEFNet demonstrates high efficacy in speech emotion recognition, outperforming other models in most metrics, particularly in the RAVDESS dataset. Its design and feature utilization serve as a benchmark for future models in the field.

Table 9. Assessment of the Proposed Model against Benchmark datasets

			RAVDESS Dataset			EMODB dataset		
Ref#	**Benchmarks**	**Input features**	**Accuracy (%)**	**Precision (%)**	**F1 score (%)**	**Accuracy (%)**	**Precision (%)**	**F1 score (%)**
[7]	Att-Net	Spatial Features	80	81	80	NA	NA	NA
[8]	DS-CNN	Spatial Features	79.50	81	84	NA	NA	NA
[9]	Deep-BLSTM	Spatial + Temporal	77.02	76	77	NA	NA	Na
[20]	CTENet	Spatial + Temporal	82.31	81.75	84.37	NA	NA	NA
[10]	ensemble 1D-CNN-LSTM-GRU	Spatial + Temporal	92	93	94	90.22	91	90
[22]	IFR-CNN with IIUC	Spatial + Temporal	NA	NA	NA	89.5	91.22	92.47
[23]	DCNN+DBN	Spatial +temporal	97.27	89.71	98.41	NA	NA	NA
Our	CEFNet(IFR-CNN , DCNNs,DBNs+ BNLA)	Spatial +temporal	98.11	98.53	98.68	91.17	91.13	90.13

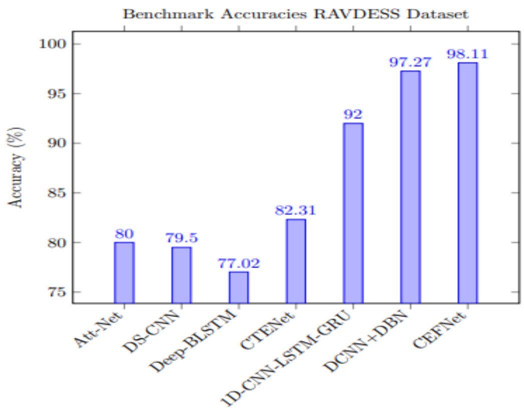

Fig 9 : Comparative Evaluation of Different Benchmarks Using the RAVDESS Dataset

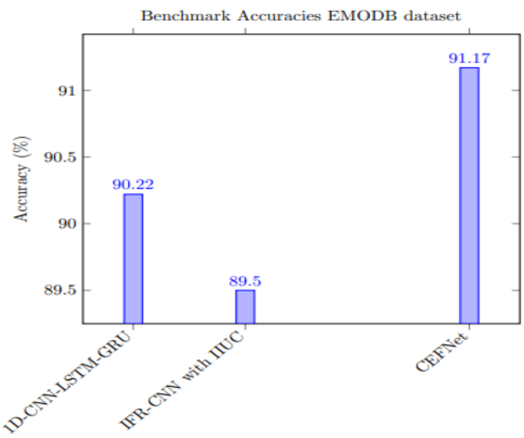

Fig 10 : Comparative Evaluation of Different Benchmarks Using the EMODB dataset

5.2 Limitations of the Study

1. **Dataset Diversity and Size**: The current study primarily utilizes the RAVDESS and EMODB datasets. While comprehensive, these datasets may not fully represent the wide spectrum of human emotions and cultural diversity in speech patterns. This limitation could impact the model's generalizability across various languages and ethnic groups.
2. **Real-World Application Testing**: CEFNet has demonstrated high efficacy in a controlled experimental setting. However, its performance in real-world scenarios, where background noise and speech variations are more prevalent, remains less explored.
3. **Computational Resources**: Although the model is optimized for efficiency, its deployment in environments with limited computational resources (like mobile devices or low-end hardware) hasn't been thoroughly tested. The balance between model complexity and resource constraints needs further investigation.
4. **Dynamic Emotional States**: The model's ability to recognize complex, overlapping, or rapidly changing emotional states in speech is not extensively examined. Emotional states in real-world conversations can be more nuanced than those presented in the datasets used.
5. **Interpretability of Model Decisions**: Like many deep learning models, CEFNet faces challenges in interpretability. Understanding the rationale behind specific emotion recognition decisions is crucial for certain applications, such as mental health assessments.

5.3 Future Directions

1. **Expanding Dataset Coverage**: Future research could include more diverse datasets, encompassing different languages, accents, and cultural backgrounds. This expansion would enhance the model's applicability and accuracy across a broader range of users.
2. **Robustness in Varied Environments**: Testing and optimizing CEFNet in more dynamic and challenging acoustic environments would be valuable. This includes scenarios with background noise, different recording qualities, and real-time interaction settings.
3. **Resource-Optimized Versions**: Developing a version of CEFNet tailored for environments with limited computational power, such as mobile or embedded devices, could significantly expand its applicability.
4. **Handling Complex Emotional Expressions**: Further research could focus on enhancing the model's ability to understand and categorize more complex emotional expressions, such as mixed or transitional emotional states.
5. **Model Explainability**: Improving the interpretability of the model's decision-making process is crucial. This could involve integrating techniques that provide more transparent insights into how and why the model arrives at specific emotion classifications.
6. **Integration with Other Modalities**: Combining speech data with other modalities like facial expressions or physiological signals could provide a more holistic approach to emotion recognition and increase accuracy.

By addressing these limitations and exploring these future directions, the utility and applicability of CEFNet in various real-world scenarios could be significantly enhanced.

VI. CONCLUSIONS AND SUGGESTIONS

Our research introduces the Cognitive Emotion Fusion Network (CEFNet), a novel hybrid model in the field of Speech Emotion Recognition (SER), especially pertinent in mental health contexts such as PTSD diagnosis. CEFNet integrates Improved and Faster Region-based Convolutional Neural Networks (IFR-CNN), Deep Convolutional Neural Networks (DCNNs), Deep Belief Networks (DBNs), and the Bird's Nest Learning Analogy (BNLA). This integration marks a significant leap in emotion recognition, combining the strengths of each technology. In quantitative terms, CEFNet demonstrated exemplary performance on the EMODB and RAVDESS datasets, achieving an accuracy of 91.17% and 98.11%, precision of 91.13% and 98.53%, and F1 scores of 90.13% and 98.68%, respectively. These results underscore CEFNet's superiority in accuracy and precision over existing models, indicating its robust capability in detecting a wide range of emotions.

Looking forward, the success of CEFNet opens diverse prospects for future exploration. Key areas include expanding the model's dataset diversity to cover a wider range of languages and cultural contexts, enhancing its applicability across various user groups. Additionally, optimizing the model for resource-constrained environments, such as mobile devices, and improving its interpretability, especially for sensitive applications like mental health assessments, are crucial. Future studies could also focus on integrating SER with other modalities like facial expressions or physiological signals for a more holistic emotion recognition approach. These advancements will not only enhance the model's utility but also broaden its applicability in real-world scenarios, paving the way for more empathetic and nuanced human-computer interactions.

REFERENCES

[1] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., &Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE access*, *9*, 47795-47814.

[2] Yehuda, R., Hoge, C. W., McFarlane, A. C., Vermetten, E., Lanius, R. A., Nievergelt, C. M., ... & Hyman, S. E. (2015). Post-traumatic stress disorder. *Nature reviews Disease primers*, *1*(1), 1-22.

[3] Schuller, B., &Batliner, A. (2013). *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

[4] Suneetha, C., &Anitha, R. (2022). A Survey Of Machine Learning Techniques OnSpeech Based Emotion Recognition And Post Traumatic Stress DisorderDetection. *Neuroquantology*, *20*(14), 69.

[5] Bhatt, R. (2023). An Analytical Review of Deep Learning Algorithms for Stress Prediction in Teaching Professionals. *Innovative Engineering with AI Applications*, 23-39.

[6] Hyland Bruno, J., Jarvis, E. D., Liberman, M., &Tchernichovski, O. (2021). Birdsong learning and culture: analogies with human spoken language. *Annual review of linguistics*, *7*, 449-472.

[7] Kwon, S. (2021). Att-Net: Enhanced emotion recognition system using lightweight self-attention module. Applied Soft Computing, 102, 107101.

[8] Mustaqeem, & Kwon, S. (2019). A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors, 20(1), 183

[9] Sajjad, M., & Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE access, 8, 79861-79875.

[10] Ahmed, M. R., Islam, S., Islam, A. M., &Shatabda, S. (2023). An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. Expert Systems with Applications, 218, 119633.

[11] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391.

[12] Nakano, A., &Nagamune, K. (2022). A Development of Robotic Scrub Nurse System-Detection for Surgical Instruments Using Faster Region-Based Convolutional Neural Network–. Journal of Advanced Computational Intelligence and Intelligent Informatics, 26(1), 74-82.

[13] Corujo, L. A., Kieson, E., Schloesser, T., &Gloor, P. A. (2021). Emotion recognition in horses with convolutional neural networks. Future Internet, 13(10), 250.

[14] Seshaiah, M. (2021). Comparative Analysis of Various Face Detection and Tracking and Recognition Mechanisms using Machine and Deep Learning Methods. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(11), 215-223.

[15] G. Deepika, & K. Deepthi Reddy. (2022). Machine Learning Based Emotional Sentiment Analysis of Tweet Data Using a Voting Classifier. *International Journal of Computer Engineering in Research Trends*, *9*(10), 193–200.

[16] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, *13*(5), e0196391.

[17] Sato, N., &Obuchi, Y. (2007). Emotion recognition using mel-frequency cepstral coefficients. *Information and Media Technologies*, *2*(3), 835-848.

[18] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

[19] S, N., N, P., & P, N. (2023). A Study on Flower Classification Using Deep Learning Techniques. *International Journal of Computer Engineering in Research Trends*, *10*(4), 161–166.

[20] Ullah, R., Asif, M., Shah, W. A., Anjam, F., Ullah, I., Khurshaid, T., ...&Alibakhshikenari, M. (2023). Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer. Sensors, 23(13), 6212

[21] Peng, Z., Lu, Y., Pan, S., & Liu, Y. (2021, June). Efficient speech emotion recognition using multi-scale cnn and attention. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3020-3024). IEEE.

[22] ChappidiSuneetha and RajuAnitha (2023) Speech Based Emotion Recognition By Using a Faster Region-Based Convolutional Neural Network ,Multimedia Tools and Applications(Springer-SCIE) (Accepted)

[23] ChappidiSuneetha and RajuAnitha (2023) Synergistic Integration of DCNNs and DBNs with Bird's Nest Learning Analogy for Enhanced PTSD Detection from Emotional Speech Data ,Multimedia Tools and Applications(Springer-SCIE) (Communicated)