SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V11I12P122

Volume 11 Issue 12, 243-254, December 2024
© 2024 Seventh Sense Research Group®

Original Article

Performance Analysis of CBAM-Based Capsule Net for Neural Network for Speech Emotion Recognition

Nishant Barsainyan[1]*, Dileep Kumar Singh[1]

[1]*School of Engineering and Technology, Jagran Lakecity University, Chandanpura, Madhya Pradesh, India.*

Corresponding Author : nishant.barsainyan@gmail.com

Received: 12 October 2024 Revised: 13 November 2024 Accepted: 11 December 2024 Published: 31 December 2024

***Abstract** - Emotion recognition from audio signals is a challenging yet crucial task with applications in human-computer interaction, affective computing, and psychological research. This paper presents an innovative approach for audio emotion recognition, starting with a comprehensive pre-processing pipeline that integrates Infinite Impulse Response (IIR) filtering, MFCC, and Modified Emphasized Dynamic Com (MEDC). This combination surpasses conventional methods like MFCC and FFT by better isolating emotion-specific features and reducing noise. The paper introduces a novel architecture combining Capsule Networks (CapsNets) with a Convolutional Block Attention Module (CBAM). The CapsNet architecture, inspired by the human visual system, efficiently captures hierarchical spatial features and contextual dependencies, addressing the limitations of traditional CNN-based models. The integration of CBAM further refines the feature maps by emphasizing salient regions, improving emotion-related information extraction. The proposed system achieves an accuracy of 98.57% in recognizing emotions from audio data. Experimental results demonstrate the effectiveness of this approach on benchmark datasets, showing resilience to variations in voice quality, background noise, and speaker characteristics. A comparative analysis with traditional deep learning architectures and existing emotion recognition methods substantiates the CapsNet-CBAM model's accuracy and computational efficiency superiority*.

***Keywords** - Emotion recognition, Convolutional block attention module, CapsNet, Accuracy, Computational time*.

1. Introduction

In recent times, technological advancements have played a pivotal role in driving the progress of various systems dedicated to comprehending human emotions [1]. Among these, speech-based emotional recognition systems have taken the lead. The capability to interpret emotions conveyed through speech has far-reaching potential, spanning diverse domains such as healthcare, human-computer interaction, and beyond [2]. A key strategy employed to boost the precision and effectiveness of these systems is transfer learning, which is emerging as a promising technique [3]. This study explores the domain of speech-emotional recognition systems, honing in on applying transfer learning methods to elevate performance and analysis. Transfer learning, founded on harnessing knowledge acquired in one domain to address challenges in another, has drawn notice for its capacity to streamline training processes and enhance model generalization [4]. Speech Emotion Recognition (SER) stands at the intersection of linguistics, psychology, and technology, aiming to decode the intricate emotional cues embedded within the human speech. Every inflection, tone, and rhythm carries a wealth of emotional information, making speech a rich tapestry of feelings and intentions [5]. SER seeks to unravel this tapestry, employing advanced algorithms and machine learning techniques to discern and classify these emotions accurately. Emotions are the essence of human expression, and understanding them is pivotal in various domains, from human-computer interaction to mental health diagnostics [6]. SER delves into the subtle nuances of speech, examining parameters such as pitch, intensity, tempo, and spectral features to identify emotions like happiness, sadness, anger, fear, and more. By leveraging sophisticated models trained on vast datasets, SER endeavours to close the distance between raw audio signs and the intricate realm of human emotions. The applications are diverse, from enhancing customer service experiences to aiding in therapeutic interventions and designing more empathetic AI systems [7]. However, despite these advancements, several critical challenges persist in SER systems. Traditional deep learning architectures often fail to capture speech data's hierarchical features and contextual dependencies. Additionally, existing methods frequently struggle with resilience to variations in voice quality, background noise, and speaker diversity, limiting their practical applicability. Furthermore, while transfer learning has demonstrated its potential in addressing these issues, its application in SER systems remains underexplored, particularly concerning cross-linguistic applicability, emotional state diversity, and efficient feature extraction techniques. These limitations highlight the need for novel approaches to bridge these gaps and improve accuracy and robustness in real-world scenarios. This analysis aims to achieve three main objectives:

- Examine the basic ideas behind identifying emotions in speech signals.
- Examine the theories and practices of transfer learning for identifying speech emotions.
- Analyze the advantages and disadvantages of transfer learning in this field.

Furthermore, this examination aims to investigate and consolidate existing research and methodologies, emphasizing the achievements and obstacles faced in implementing transfer learning for identifying feelings in speech. The thorough assessment will encompass various aspects, including dataset size, model architectures, methods for feature extraction, and the applicability of learned representations across a spectrum of emotional states and languages [8]. By merging theoretical insights with practical evidence and implications, this work aims to offer an extensive comprehension utilization of transfer learning in systems that recognize emotions in speech [8]. This exploration seeks to shed light on potential directions for future research, optimizations, and advancements within this rapidly evolving field. Ultimately, this study strives to contribute to the ongoing discourse, fostering progress that enhances the accuracy, resilience, and practicality of speech-emotional recognition systems utilizing transfer learning methodologies. To overcome these limitations, this paper proposes a novel architecture combining CapsNets and the CBAM for robust emotion recognition from audio signals. Additionally, the proposed work introduces advanced pre-processing techniques, such as Infinite Impulse Response (IIR) filtering, alongside detailed feature extraction methods like MFCC, Discrete Cosine Transform (DCT), and Modified Emphasized Dynamic Compression (MEDC) to improve input audio quality and highlight emotion-relevant features. The main contribution of the proposed work is as follows.

- A novel architecture combining Capsule Networks (CapsNets) and Convolutional Block Attention Module (CBAM) is proposed for emotion recognition from audio signals.
- Conventional methods like MFCC or FFT often struggle with noise reduction and isolating emotion-specific features. Our approach combines IIR filtering, MFCC, and MEDC to enhance feature quality, with MEDC introducing dynamic emphasis for better emotion representation, surpassing prior MFCC-based methods.
- Existing SER systems often utilize Convolutional Neural Networks (CNNs), which struggle with hierarchical spatial features and contextual dependencies. The proposed CapsNet-CBAM architecture bridges this gap by combining CapsNets' hierarchical feature extraction capabilities with CBAM's attention-driven refinement, outperforming state-of-the-art methods.
- CapsNets effectively capture hierarchical and spatial features, while CBAM enhances feature maps by emphasizing salient regions, improving the extraction of emotion-relevant information. An efficient feature extraction pipeline, combining MFCC, DCT, FFT, and Modified Emphasized Dynamic Compression (MEDC), enhances emotion-specific features before input to the CapsNet-CBAM model.
- The integrated CapsNet-CBAM model leverages attention mechanisms to refine feature maps and captures hierarchical spatial features, improving emotion classification accuracy and robustness in noisy environments.

This work is structured with an initial Introduction, preceded by a literature survey. A comprehensive proposed methodology is outlined, accompanied by an explanation of the used dataset subsequent sections present outcomes and discussion, culminating the work with a conclusive summary.

2. Literature Survey

Sandeep Kumar et al. [9] introduced a voice-activated intelligent assistant that recognizes emotions. It analyses human emotions using a biometric system and manages electronic accessories to trigger alarms. The intelligent assistant can identify the seven emotions: worry, surprise, neutral, melancholy, happiness, hatred, and love. The system uses voice processing to identify emotions and automatically identify actions utilizing buzzers and LEDs. After training and testing the system on multiple datasets, it outperforms previous technologies in terms of error rate and execution time. Compared to the GMM and SVM models, the accuracy of the suggested model is 81.02%, 84.23%, and 85.12%, respectively, for the RAVDEES, TIMIT, and Emo-DB datasets. Christy et al. [10] say emotion recognition is crucial in understanding human interactions and forming social relationships. One can determine an individual's emotions by listening to the tone and pitch of their voice. This research classifies and predicts human emotions such as neutral, calm, happy, sad, afraid, disgusted, and surprised using methods such as CNN, decision trees, random forests, SVM, and linear regression. Compared to decision trees, random forests, and SVM, CNN demonstrated 78.20% accuracy in identifying emotions when evaluated on the RAVDEES dataset.

Ala Saleh Alluhaidan et al. [11] say that a critical procedure in many domains, MFCCs with time-domain features MFCCT, has improved SER. A CNN was employed in the study to develop a SER model. The Emo-DB, SAVEE, and RAVDESS datasets yielded 97%, 93%, and 92% accuracy values for the hybrid MFCCT features. CNN outperformed SER's machine learning classifiers, suggesting that different SER datasets may be able to identify emotions. This strategy has much potential for several applications. Rashid Jahangir et al. [12] say SER is a demanding arena of study with gaming applications, mobile services, person-computer interfaces, and psychological testing. Previous research has employed handcrafted characteristics, but their accuracy may suffer in complicated circumstances. This work introduces a brand-new SER framework that extracts seven informative feature sets from each speech through data augmentation techniques. The restored feature vector is fed into a 1D CNN, utilizing the EMO-DB, RAVDESS, and SAVEE databases for emotion recognition. With a precision of 96.7%, the suggested framework performs better than the current SER frameworks. Kudakwashe Zvarevashe et al. [13] explained that for applications involving human-computer interaction, voice emotion recognition is essential. To this end, researchers have employed machine learning and hand-crafted feature sets. These techniques lack robustness and require a lot of computing power. DL techniques have accomplished robust feature extraction from datasets. For feature extraction and vocal utterance categorization, a bespoke 2D-convolution neural network was created for this investigation. The network was evaluated against deep multilayer perceptron and deep radial basis function neural networks using the Surrey audio-visual expressed emotion corpus, the Ryerson audio-visual emotional speech database, and the Berlin emotional speech database. When measured against other current algorithms, the deep learning algorithm produced the top outcomes regarding F1 scores, recall, and precision. Depending on the application domain, customized solutions could be required for various language settings.

Shibani Hamsa et al. [14] aim to develop an artificial emotional intelligence system for detecting unidentified emotions in a speaker. It presents a unique paradigm for emotion detection in interference and noise by considering energy, temporal, and spectral factors. Instead of a Gammatone Filterbank and a short-time Fourier transform, the system employs a cochlear filter bank based on wavelet packet transform. When paired with a random forest classifier, the system outperforms existing methods on three unique voice corpora in two languages under stressful and loud conversation settings. Chawki Barhoumi et al. [15] described a speech emotion recognition system that uses deep learning and two data augmentation techniques. The system employs MFCC, ZCR, Mel spectrograms, RMS, and Chroma to identify voice features indicating speech emotions from three datasets: RAVDESS, TESS, and EmoDB. The deep learning models employed are MLP, CNN, and a hybrid model that combines CNN and Bi-LSTM. The best model for accurately identifying emotional states from speech cues in real-time scenarios is found. The work demonstrates the effectiveness of two data augmentation techniques and the proposed CNN + BiLSTM deep learning model for real-time speech emotion recognition. Khan et al. [16] propose a novel Multimodal Speech Emotion Recognition (MSER) model using a multi-headed cross-attention mechanism with a deep feature fusion technique. The model combines audio and text cues to predict emotional labels, processing raw speech signals and text through CNNs for feature extraction. The proposed architecture enhances the interaction between text and audio by applying cross-attention to both feature sets, followed by a deep fusion strategy to optimize the final emotion recognition decision. The authors evaluate their model on the IEMOCAP and MELD datasets, achieving a 4.5% improvement in accuracy, setting a new state-of-the-art in the MSER field. Despite the model's promising results, its performance may be limited by the complexity of the fusion strategy, which could affect its scalability for real-time applications. Wang et al. [17] proposed a Speech Emotion Recognition (SER) model tailored for smart home systems using an ensemble deep learning approach, TF-Mix, which merges time and frequency domain features for enhanced emotion representation. The model employs data augmentation to address dataset limitations and integrates CNNs, BiLSTM-FCNs, and Transformer-based architectures. An ensemble model (D) of these architectures achieved superior accuracy, with results of 87.513% on RAVDESS, 86.233% on SAVEE, 99.857% on TESS, 82.295% on CREMA-D, and 97.546% on the TOTAL dataset. Despite its success, the model's reliance on feature fusion increases computational complexity, which may limit real-time applicability in smart home systems.

Paul et al. [18] proposed a speech emotion recognition model utilizing feature fusion techniques to enhance prediction accuracy. The model extracts MFCC, LPC, energy, ZCR, and pitch features and applies them to SVM, LDA, D-Tree, and KNN classifiers. Experimental results show high recognition rates: 96.90% on SUBESCO (Bengali), 99.82% on TESS (English), 95% on RAVDESS (English), and 95.33% on EMO-DB (Berlin). The fusion approach significantly improves the model's accuracy and applicability. However, the method may face limitations in real-time deployment due to computational overhead during feature extraction and fusion processes. Existing Speech Emotion Recognition (SER) systems face several limitations. Many models, particularly those that use deep learning or feature fusion techniques, require substantial computational resources. Additionally, these systems are often trained on specific datasets, making it difficult to generalize to diverse speech patterns or languages. Hand-crafted feature extraction techniques can struggle to capture nuanced emotions, especially in complex environments, leading to decreased accuracy. Overfitting is another concern, as models may perform well on training data but fail to generalize effectively. Finally, the complexity of multimodal or feature fusion strategies can hinder the scalability of these models, making them less suitable for larger or more dynamic settings.

3. Proposed Methodology

This research paper focused on the challenging task of recognizing emotions from audio signals, a crucial aspect with applications in various fields such as human-computer interaction, affective computing, and psychological research. The proposed approach combines Capsule Networks (CapsNets) with a CBAM to enhance Extracting features and capture contextual information, aiming for an accuracy of 98.57% in emotion recognition from audio data. Inspired by the human visual system, Capsule Networks provide a unique way of representing hierarchical features. This enables the model to efficiently capture spatial hierarchies in audio data, contributing to more effective emotion recognition. The integration of CBAM further refines feature maps by highlighting salient regions, thereby promoting the extraction of robust emotion-related information. CapsNets offer a distinctive way to represent hierarchical features in audio data, allowing for efficient capture of spatial hierarchies related to emotions. The integration of CBAM further enhances the model's performance by refining feature maps, emphasizing salient regions, and promoting the extraction of robust emotion-related information. This unique combination demonstrates an accuracy of 98.57%, showcasing the model's capability to discern and classify emotions accurately. Notably, the system exhibits resilience to variations in voice quality, background noise, and speaker characteristics, making it practical for real-world scenarios. Comparative analysis substantiates the superiority of the proposed CapsNet-CBAM fusion model over traditional deep learning architectures and existing emotion recognition techniques in terms of precision and computational effectiveness.

The results of the proposed approach are promising, demonstrating its effectiveness on benchmark emotion recognition datasets. The achieved accuracy of 98.57% indicates—the model's capacity to accurately discern and classify a range of emotions. Notably, the system shows resilience to variations in voice quality, background noise, and speaker characteristics, enhancing its practical viability in real-world scenarios. Additionally, the paper includes a comparative analysis, where the proposed CapsNet-CBAM fusion model is compared with traditional deep learning architectures and existing emotion recognition methods. The results of this comparison highlight the superiority of the suggested model in terms of computational efficiency and accuracy. This indicates that the CapsNet-CBAM fusion model offers a more effective and efficient solution for emotion recognition from audio signals than existing approaches. Figure 1 shows the architecture Diagram of the Proposed Method.

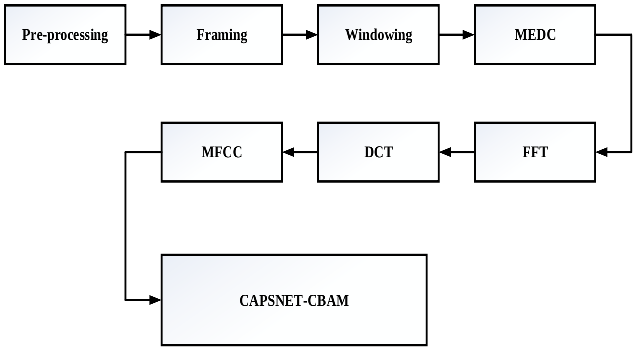

Fig. 1 Architecture diagram

3.1. Dataset Description

The RAVDESS database was chosen because it includes recordings of both recording and speech. It is categorized into eight emotional categories based on assessments by 247 untrained Americans. These categories include Relaxation, Happiness, Sadness, Fury, Fear, Disgust, Surprise, and a neutral baseline for each performer. This dataset comprises data from 24 performers, ensuring gender balance with 12 males and 12 females. The same statements were recorded in an American accent for the audio files produced under controlled circumstances. The two primary file formats are speech files, totalling 1440 files (60 trials per actor, 24 actors), and audio files, consisting of 1012 recordings (44 evaluations per actor, 23 actors). Both varieties use the WAV raw audio file format, maintaining a 16-bit bitrate and a 48 kHz sampling rate. These recordings are uncompressed and lossless, preserving the original data without any alterations or omissions. The libROSA Python package is preferred for processing this dataset. This package was chosen for its suitability in music and audio research. Utilizing libROSA, WAV files were loaded individually, generating a 1-dimensional array representing audio time series in stereo format, where the array's length corresponds to the sample rate. The 'load' function in libROSA facilitated this conversion, allowing subsequent derivation of MFCC and spectrophotograms of the raw audio files for Chroma and Mel. Various libROSA functions were employed to manipulate the audio, culminating in extracting specific functionalities that were then aggregated and returned as a NumPy list, following file loading using the libROSA library.

3.2. Pre-Processing

Infinite Impulse Response filters (IIR Filters), integral to Digital Signal Processing, possess an impulse response that persists indefinitely due to their recursive nature. These filters typically utilize a feedback structure, while FIR filters operate without feedback. The standard method for designing IIR filters involves bilinear transformation, where the process starts with an analogy filter's transfer function and proceeds to map the s-domain to z-domain transition. Through differential equations, it can be demonstrated that the conversion from the s-plane to the z-plane is characterized by:

$$S = \frac{2}{T}\left(\frac{1-Z^{-1}}{1+Z^{-1}}\right) \tag{1}$$

In Equation (1), S denotes the complex frequency variable in the s-domain, T represents the sampling period, Z denotes the complex frequency, and Z^{-1} indicates the inverse of the z-variable, representing a delayed version of the discrete signal. This correspondence leads to a universal structure for an IIR filter that can accommodate any quantity of poles and zeros. The outcome includes the system's reaction and the distinct equation representing this filter.

$$H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{n=0}^{M} b_n z^{-n}}{\sum_{n=0}^{N} a_n z^{-n}} \tag{2}$$

$$= \frac{b_0+b_1z^{-1}+--b_Mz^{-m}}{1+a_1z^{-1}+--a_Nz^{-N}}, a_{0=1} \tag{3}$$

From Equation (2), $H(z)$ denotes the transfer function of the filter; $B(z)$ indicates the numerator polynomial in the z-domain, containing the filter's zeros; $A(z)$ depicts the denominator polynomial in the z-domain, containing the filter's poles and $b_n\ and\ a_n$ denotes the coefficients of the numerator and denominator polynomial. IIR filters, characterized by their traditional analogy like Butterworth, Chebyshev, Elliptic, and Bessel filters, can be understood and created using familiar techniques for traditional filter design. These filters are commonly structured in two-pole segments known as biquads, derived from a biquadratic equation in the z-domain. Complex IIR filters employ cascades of these biquad sections. The zeros, formed by coefficients b0, b1, and b2, and the poles, formed by coefficients a1 and a2, shape the filter's response. By leveraging both feed-forward (zero-based) and feedback (pole-based) polynomials, IIR filters achieve sharper transition characteristics relative to their filter order.

3.3. Feature Extraction

The speech signal, which is continuous in time, gets sampled at a particular frequency. Initially, in the Mel-Frequency Cepstral Coefficients (MFCC) feature extraction process, the first step involves amplifying the energy present in the higher frequencies. This amplification is achieved by applying a specific filter.

3.3.1. Framing

Segmenting entails breaking up the speech samples from the ADC into shorter frames, 20–40 milliseconds long. By splitting the dynamic speech signal into quasi-stationary frames, this segmentation makes it easier to apply a Fourier transformation to the speech signal. This technique is used because the speech signal usually exhibits quasi-stationary properties in this short 20–40 ms timeframe.

3.3.2. Windowing

The windowing step aims to create a window for each frame separately, reducing signal disruptions occurring at the start and finish of every frame.

3.3.3. Mel Energy Spectrum Dynamic coefficients (MEDC)

Another element, the MEDC, is derived in the following manner: the magnitude spectrum of every spoken expression is gauged through FFT, then channelled through a set of 12 filters distributed evenly across the Mel frequency scale. The average logarithmic of the filter's outgoing energy, denoted as En (i), where i ranges from 1 to N, is computed. Subsequently, the initial and subsequent variances of En (i) are determined, constituting the process of extracting the MEDC features.

3.3.4. FFT

The FFT algorithm is best suited to analyzing the speaking frequency range. It transforms sets of N samples moving from the frequency domain to the temporal domain, frame through frame. Consider the discrete cosine transform employed to make filter energy vectors orthogonal. This process condenses the filter energy vector's data into the initial components, reducing its length. This orthogonalization step shortens the vector to fewer components while retaining its essential information.

3.3.5. Take Discrete Cosine Transform

It is employed to make the filter energy vectors orthogonal. This process condenses the information from these vectors into the initial components, effectively reducing the vector's length.

3.3.6. MFCC

It relies on how humans perceive sound and is a widely utilized technique for extracting features in sound processing. MFCC features aim to capture unique speaker characteristics by replicating the human ear's sensitivity to frequencies. Equations (4) and (5) below facilitate the conversion between the Mel scale (M) and frequency scale (Hz).

$$m = 295\, log_{10}\left(1 + \frac{f}{700}\right) \tag{4}$$

$$F = 700\left(10^{\frac{m}{295}} - 1\right) \tag{5}$$

One way to depict this perceptual layout is to use a triangle band-pass Mel filter bank. After the filter bank, a discrete cosine transform is used to obtain MFCCs, which are calculated using an equation.

$$MFCC_i = \sum_{k=1}^{20} X_k Cos\left[i\left(\frac{k-1}{2}\right)\frac{\text{п}}{20}\right]\ i = 1,2, \ldots M \tag{6}$$

3.4. Speech Recognition

This section describes the integration of Capsule Networks (CapsNet) with the Convolutional Block Attention Module (CBAM) to enhance speech recognition tasks. The internal workings of the CapsNet-CBAM fusion combine the hierarchical feature extraction capabilities of CapsNet with the attention mechanism of CBAM. CapsNet focuses on preserving the spatial relationships between features. It uses capsules, which are groups of neurons, to encode the presence, pose, and orientation of features in the data, providing robust feature representations that can recognize patterns despite transformations like rotations and scaling. CBAM enhances these features by applying two forms of attention: channel and spatial. The channel attention mechanism highlights the most crucial feature channels, while the spatial attention mechanism identifies critical regions in the feature maps that should be emphasized for further processing. The CapsNet extracts the hierarchical features, and the CBAM refines these features, highlighting essential parts and suppressing irrelevant ones, which allows for more accurate recognition of complex patterns in the data. After the CBAM module enhances the feature maps, they are passed into the capsule layers, where the capsules route the information, preserving spatial hierarchies and providing more robust outputs. Fusing these methods results in a model that captures both the hierarchical structure of features and the most critical elements for accurate recognition. This combination enhances the model's ability to distinguish essential signals, improving performance in speech emotion recognition.

3.4.1. Capsule Neural Network

An input and an output are represented as vectors in a capsule, a cluster of neurons [19]. To illustrate the essential operation of the capsule neural network that is the subject of this research, let us look at an example of one of these networks with M capsules at a higher level and N capsules at a lower level. At the upper level, Capsule 1 produces an output vector ~v1, which encodes the direction and existence of, say, Object 1. N input vectors are sent to this specific capsule, Capsule 1, each of which corresponds to the outputs of N capsules with the lower level labels 1….i…..N.

Assuming that (u1) ⃗... ; (ui) ⃗;... ;(un) ⃗, are the output vectors from the lower-level capsules, and keeping in mind that the presence and orientation of portion I, which is connected to the item j that capsule j at the higher level describes, are encoded in the output vector (ui) ⃗ from capsule i in the lower level

Before capsule j on the upper tier, vector outcomes $\vec{u}_i$ modified according to the weight matrix W_{ij}, capturing three-dimensional correlation amid parts i and thing j, transforming into the vector $\overrightarrow{\hat{u}}_{j|i} = W_{ij}\vec{u}_{i}.$.This $\vec{u}_{j|i}$ the scalar weight further scales the vector c_{ij}, determined through a routing algorithm, to yield $c_{ij}\vec{u}_{j|i}$ before it enters into capsule j. Similarly, capsules 1 and N on the lower tier contribute their respective vectors entering capsule j, namely, $c_{1j}\vec{u}_{j|1}$ and $c_{Nj}\vec{u}_{j|N}$, where $\vec{\hat{u}}_{j|1} = W_{1j}\vec{u}_1$ and $\vec{\hat{u}}_{j|N} = W_{Nj}\vec{u}_N$.

At the higher-tier capsule j, the summation operation takes place:

$$\vec{s}_j = \sum_i c_{ij}\vec{u}_{j|i} \tag{7}$$

Let sj denote the summed vector at capsule j in the higher level, formed by the weighted contributions of the lower-level capsules; $\vec{u}_{j|i}$ denotes the intermediate vector for capsule j, modified by the weight matrix W_{ij} and the routing coefficient c_{ij}. The vector of output $\vec{s}_j$ is sent through the squash function of capsule$j$:

$$\vec{v}_j = \frac{\|\vec{s}_j\|^2}{1+\|\vec{s}_j\|^2}\frac{\vec{s}_j}{\|\vec{s}_j\|} \tag{8}$$

The output $\vec{v}_j$ of capsule j represents both the presence and position of object j. The coupling coefficients (routing coefficients, another name for them) c_{ij} the total amount between capsule I and every capsule in the layer above is 1. The following coefficients are computed with the use of "routing softmax":

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \tag{9}$$

The values b_{ij} represent the logarithmic probabilities determining the preferred connections between capsule i and capsule j; k depicts the capsule index in the higher level and c_{ij} The routing coefficient (called coupling coefficient) was calculated using the softmax function. The idea behind the methods in [20] is that data is transmitted from a lower-level capsule to a higher-level capsule that corresponds with it. The algorithm produces a set of routing coefficients through a specified number of routing iterations, typically three. These coefficients connect the lower-level capsule's output to the higher-level capsule's output.

For class k, Caps Net calculates the margin loss as follows:

$$\mathcal{L}_k = T_k \max(0, m^+ - \|\vec{v}_k\|)^2 + \lambda(1 - T_k)\max(0, \|\vec{v}_k\| - m^-)^2 \tag{10}$$

Where $T_\mathrm{k} = 1$ if an entity of class k is present and $m^+ = 0.9$ and $m^- = 0.1$. The weight $\lambda = 0.5$. $\mathcal{L}_k$ demonstrates the margin loss for class k, which measures the error in predicting the presence and pose of objects; m^+ and m^- represent the upper and lower margin for class k, and λ denote the regularization factor. The CapsNet-CBAM fusion model is a sophisticated method for feature extraction and classification, combining the hierarchical learning of Capsule Networks (CapsNet) with the selective attention capabilities of the Convolutional Block Attention Module (CBAM). CapsNet's dynamic routing mechanism captures spatial hierarchies and positional relationships, making it suitable for emotion recognition tasks. CBAM introduces an attention mechanism that refines feature maps by focusing on the most salient and relevant features. This mechanism operates in two stages: channel attention and spatial attention. The channel attention mechanism amplifies critical channels, while the spatial attention mechanism emphasizes significant spatial features. These mechanisms work together to enhance the quality of extracted features. When integrated, the CapsNet-CBAM fusion model synergizes CapsNet's ability to preserve spatial and temporal relationships with CBAM's emphasis on critical features, enabling better distinction of emotion-related characteristics even in noise or speaker variability. This combination improves classification accuracy and robustness, making it a powerful solution for audio emotion recognition tasks in real-world scenarios.

3.4.2. CBAM

To leverage the strengths of Capsule Networks (CapsNet) and ensemble learning for diagnosing imbalanced data samples related to bearing faults, a collective capsule system incorporating CBAM was devised using the WMAM. The Process outlined in Figure 2 involves gradually segmenting the raw vibration signal by swiping the time window over N data sets.

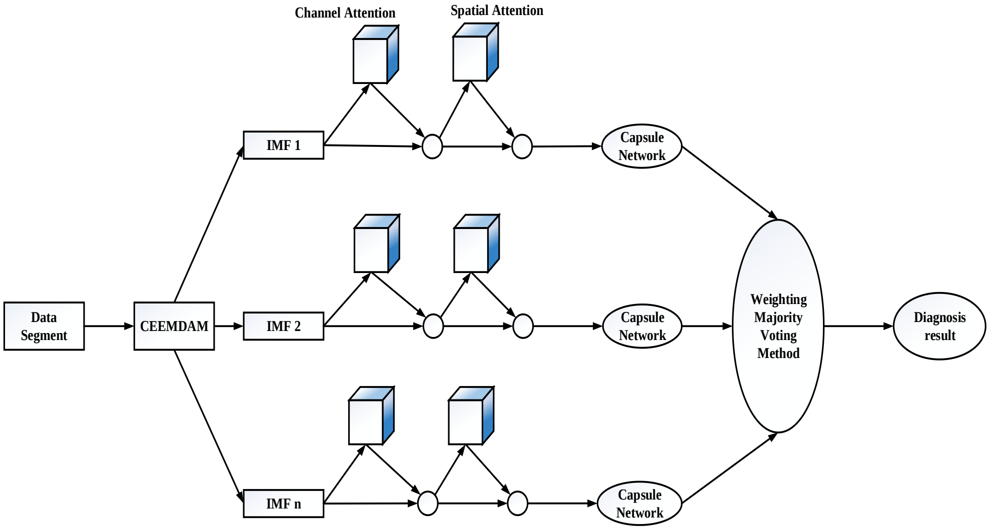

Fig. 2 Schematic of the ensemble capsule network

These samples undergo decomposition via modified EMD into distorted IMF signals. Subsequently, these IMF signs are fed CBAM into Caps Net to extract important features and perform first-fault diagnostics. The final diagnostic decision is ultimately reached through the weighted majority voting method. CapsNet extracts features from IMF signals and preserves the spatial and temporal relationships among these features. CBAM, utilizing an attention mechanism, selects sensitive features to create a more representative feature map. CapsNet, in conjunction with CBAM, was used to extract these sensitive traits, which made it easier to identify bearing defects and improved the accuracy of fault diagnostics. The attention mechanism was utilized to identify crucial features within the extracted feature maps from convolutional operations. This was done to enhance the significance of important features while diminishing the influence of less relevant ones. Initially developed by Woo et al. [21], a CBAM was used to leverage the extraction of distinguishing characteristics by simultaneously focusing on channel and spatial information. The channel attention mechanism was used to generate the channel attention map (Mc), which selected the channel, and the spatial attention mechanism was used to create the spatial attention map (Ms), which identified the channel's sensitive features. The following two equations represent the feature selection procedure when the input feature, F, passes through these attention modules to produce the refined feature, F".

$$\mathrm{P}' = \mathrm{M_c}(\mathrm{F}) \otimes \mathrm{F} \tag{11}$$

$$\mathrm{F}'' = \mathrm{M}_s(\,\mathrm{F}') \otimes \mathrm{F}' \tag{12}$$

$\mathrm{F} \in \mathrm{R}^{\mathrm{CxHxB}}$ Represents the input characteristics grid of the CBAM module, where C is the width, H is the height, and B is the number of channels. P^' represents the channel attention map and feature map result, while F^" results from the spatial attention map multiplied by P^', representing the CBAM module's output. $M_c \in \mathbb{R}^{Cx1\times1}$ Represents the weight of attention in the channel dimension, while $\mathrm{M_s} \in \mathrm{R}^{1\times\mathrm{H}\times\mathrm{B}}$ represents the weight of attention in the spatial dimension. The symbol shows the multiplication of elements ⊗. CBAM enhances feature extraction by applying channel and spatial attention mechanisms. The channel attention module prioritizes important features across channels, while the spatial attention module focuses on relevant regions within the feature map. This allows the model to highlight key emotional cues in the audio data. Combined with Capsule Networks (CapsNets), which capture hierarchical spatial relationships, CBAM improves the model's ability to accurately classify emotions, even with variations in voice quality, background noise, and speaker characteristics.

3.4.3. Diagnosis Based on the Capsule Network with CBAM

While CBAM was used to discover essential parameters, Caps Net used its convolutional operation to extract feature parameters. The capsule network was then given these parameters, improving Capsnet's diagnostic performance. The general structure of the Capsnet with CBAM model is shown in Figure 3. A raw data segment, including 1024 data points, was first divided into different IMF signals of various scales. The 1-D IMF signals were converted into 32x32 2-D grey maps [22].

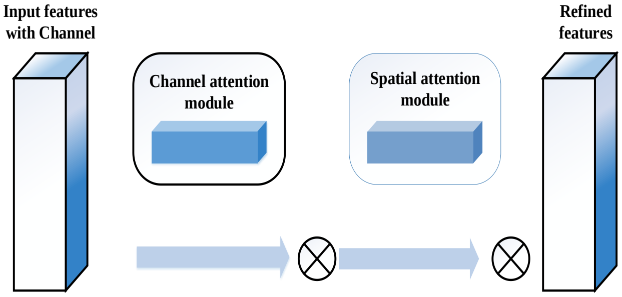

Fig. 3 CBAM attention mechanism

Afterwards, feature parameters were extracted using average pooling layers (2×2 size) and convolution layers (3×3 kernel size), with CBAM focusing on particular areas for suppression or emphasis. The activation function for each convolution layer was the RELU function, which helped to keep gradients from vanishing and support the nonlinear capacity of the model. After being reshaped, the resulting selective feature parameters were entered into Capsnet to perform a first diagnosis of bearing fault types. In the Capsnet training, the aim was to achieve the best weight parameters, which was done by defining the margin loss function in the following manner:

$$L_k = T_k \max(0, m^+ - \|a^k\|)^2 + \lambda(1 - T_k)\max(0, a^k - m^-)^2 \tag{13}$$

In this context, the symbol ‘k’ represents the fault category. T_k stands for the indicator function, taking 1 if class’ k’ is detected and 0 otherwise. m^+Represents the maximum penalty for false positives, m^- represents the minimum penalty for erroneous negative results, and ‘λ’ symbolizes the coefficient. These specifications were specifically set to 0.5, 0.1, and 0.9. The variable a^k represents the likelihood of fault category’ k,’ limited to be at most 0.1 in the absence of ‘k’ and at most 0.9 in the presence of ‘k’.

3.4.4. The Weighted Majority Voting Method (WMVM)

To enhance Capsnet’s ability to diagnose faults accurately and reliably using various IMF signals, CBAM was combined with multiple classifiers to create an ensemble Capsnet for parallel fault diagnosis. Recognizing the varying impact of these IMF signals on diagnostic outcomes, the initial diagnoses from differently scaled signals were consolidated. This was achieved by employing the WMVM in the decision-making stage to determine the final class label, aiming for heightened accuracy in the ultimate diagnostic outcome. Equation (14) outlines the specific operations involved in this final diagnostic process.

$$H(x) = \mathrm{C}_{\mathrm{armax}} \sum_{n=1}^{N} w_n h_n^i(x) \tag{14}$$

In this methodology, the prediction for each data sample x (denoted as $H(x)$) is determined by the highest voting value among various sub-classifiers. Each sub-classifier is labelled as $h_n^j(x)$, provides the prediction probability for the data. The significance of each $h_n^j(x)$ in the final decision is determined by its weight, denoted as w_n. These weights directly influence the conclusive diagnostic outcomes.

The calculation for w_n involves a formula: w_n is equal to c_n divided by the sum of accuracies of all sub-classifiers $w_n = c_n / \sum_{n-1}^{N} c_n$. Here, c_n stands for the validation accuracy of the nth sub-classifier when diagnosing the validation dataset. Notably, N is set explicitly as 7, representing the existence of 7 sub-classifiers. Additionally, the variable j denotes fault class labels in this context, where j can take on values 1, 2, or 3. The refined features are then passed to the final classification or regression layer. This fusion improves the model’s ability to focus on relevant information, boosting performance in complex tasks.

4. Result and Discussion

4.1. Configuration of Capsule Neural Network for Emotion Recognition

The NN employed in this study to detect four emotions comprises two components. The initial segment comprises five CNN layers, followed by a capsule neural network. To illustrate the network’s setup, consider the configuration instance for 296 parameters across 296 edges. The following elements are included in these five CNN layers.

- Layer 1: Uses 64 sets of 3x3 filters, employing a stride of 2 on a 296x296x1 input, resulting in a 148x148x64 output. It uses ReLU activation and Dropout with a 50% rate.
- Layer 2: Employs 16 sets of 2x2 filters with a stride of 2 on a 148x148x64 input, yielding a 74x74x16 output. Activation is ReLU, and it includes Dropout at a 50% rate.
- Layer 3: Applies 16 sets of 2x2 filters using a stride of 2 on a 74x74x16 input, resulting in a 37x37x16 output. It utilizes ReLU activation, Dropout at a 50% rate, and MaxPooling2D with a pool size of 2x2.
- Layer 4: Utilizes 16 sets of 2x2 filters with a stride of 2 on a 37x37x16 input, generating a 19x19x16 output. Activation is ReLU, including Dropout at a 50% rate and MaxPooling2D with a pool size of 2x2.
- Layer 5: Applies 16 sets of 2x2 filters using a stride of 2 on a 19x19x16 input, resulting in a 10x10x16 output. It uses ReLU activation, Dropout at a 50% rate, and MaxPooling2D with a pool size of 2x2.

The first capsule layer functions similarly to the convolutional layer in nature. This layer uses a 9×9 kernel with a stride of 2 and no padding to reduce the spatial dimensions from 10×10 to 5×5. The primary capsule layer uses 8×32 kernels to create 328 capsules, which are then organized into 8 groups of output neurons to produce a capsule. The output of these capsules is rearranged into a configuration of (800 = 5×5×32, 8). After that, an 8x16 transformation matrix Wij is used by the capsule layer. Its function is to transform the output of the primary capsule layer’s 8-D capsules into 16-D capsules, each of which represents one of four emotions. In the primary capsules, a weight matrix called Wij operates between each$(U) = -i, i$, where $i\,5 \times 5$ ranges from 1 to 32 ×, and each $(v) = -j, j$ where j ranges from 1 to 4. The configuration outlined above is unchanged for the remaining two situations involving 260 and 268 feature parameters. However, the amount of parameters that must be computed differs based on these particular feature counts.

The emotion hypothesis is determined as follows:

$$\text{Emotion} = \arg\max\|\vec{v}_j\|$$

4.2. Performance Evaluation

4.2.1. Confusion Matrix for RAVDESS Dataset

RAVDESS draws attention to the model's rich spectrum of emotion, which spans various categories. It comprises 7356 instances that vividly capture the nuances of human emotion. Hate and anger emerge strongly with 1052 instances, embodying intense negativity. Worry, expressed in 1073 instances, represents prevalent concerns. Love, shining through in 1093 instances, showcases positive and affectionate emotions. Surprise, captured 1085 times, signifies sudden and unexpected reactions. Additionally, the dataset encapsulates 1021 instances of happiness, 1029 instances of sadness, and 1003 instances classified as neutral emotions. Figure 4 and Table 1 show this diverse collection that provides profound insights into the array of emotions within the RAVDESS dataset, painting a comprehensive picture of emotional expression.

4.2.2. Computational Time

In this comparative analysis of various classifiers, three models were evaluated based on their training time and accuracy. The Caps Net model exhibited a training time of 200 seconds and achieved an accuracy of 71.76%. The CBAM layer model, on the other hand, took slightly longer to train at 272 seconds but demonstrated a comparable accuracy of 71.43%. The third model, a combination of Caps Net with CBAM, required 250 seconds for training but showcased a notable improvement in accuracy, reaching 98.57%. Despite having a relatively shorter training time, the Caps Net model displayed moderate accuracy.

The CBAM layer model, while taking more time to train, demonstrated a similar accuracy to the Caps Net. However, the combined model of Caps Net with CBAM showcased a significant leap in accuracy, suggesting a potential synergy between the Caps Net architecture and the CBAM layer. These results imply that integrating the CBAM layer with Caps Net positively impacts classification performance, substantially improving accuracy. Researchers and practitioners may find this information valuable when choosing a model based on their specific requirements, considering factors such as training time and accuracy trade-offs.

Table 1. No of samples based on RAVDESS datasets

Emotions	RAVDESS
Hate/Anger	1052
Worry	1073
Love	1093
Surprise	1085
Happiness	1021
Sadness	1029
Neutral	1003
Total	7356

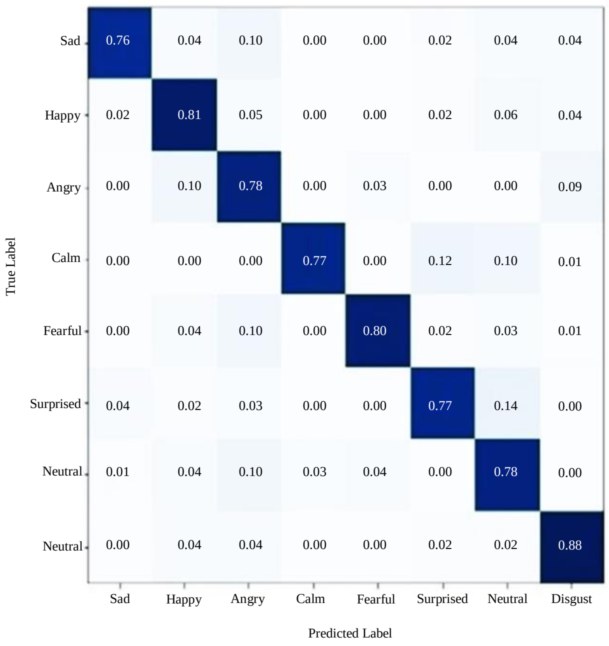

Fig. 4 Confusion matrix

Table 2. Computational time and accuracy

Model	Training Time (s)	Accuracy (%)
Caps Net	200	71.76
CBAM layer	272	71.43
Caps net with CBAM	250	98.57

Further investigations into the underlying mechanisms of the combined model could provide insights into the synergistic effects of these two components, guiding future developments in classifier architectures. Figure 5 and Table 2 show the proposed method's computational time and accuracy results. The dataset under consideration comprises 273 input features, with the data sourced after the RAVDESS. Evaluation metrics for the model performance include overall efficiency, recall, precision, and the F1-s in this specific scenario. The precision stands at 0.98, indicating a high level of accuracy in correctly identifying positive cases.

The recall value is also 0.98, underscoring the model's effectiveness in capturing all relevant positive instances. The F1-score, computed as 0.98 and is a harmonic mean of precision and recall, further shows the model's strong performance. With an overall accuracy of 0.98, the percentage of correctly identified instances in the total is indicated. Together, these indicators demonstrate the model's proficiency in accurately classifying and recognizing patterns within the given dataset, mainly when applied to the specified input features from a database such as RAVDESS. Figure 6 and Table 3 show the outcomes of the proposed method training in the RAVDESS dataset.

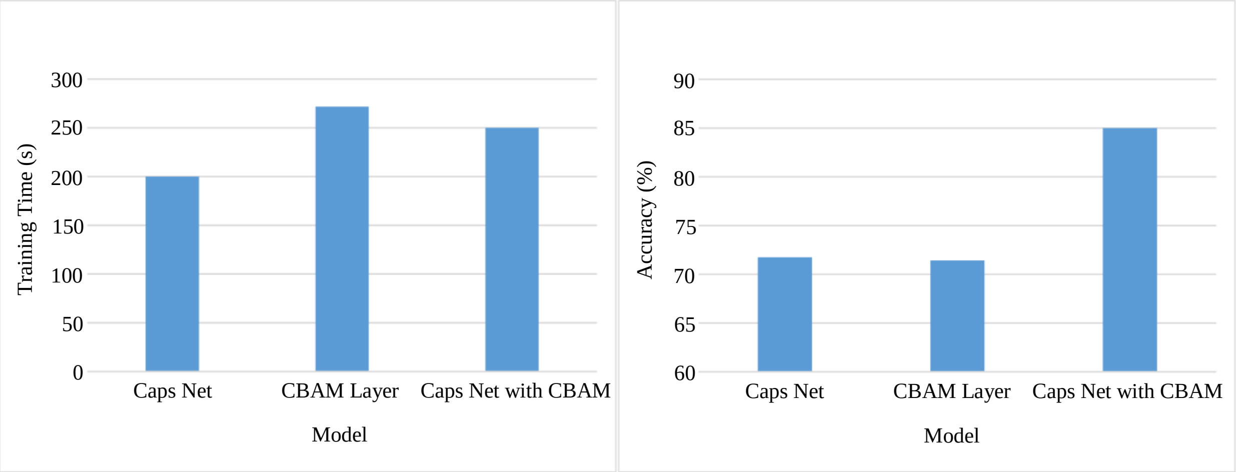

Fig. 5 Computational time and accuracy

Table 3. Outcomes of the suggested models' training on the RAVDESS database

Input features	Databases	Precision	Recall	F1-score	Accuracy
273 input features	RAVDESS	0.98	0.98	0.98	0.98

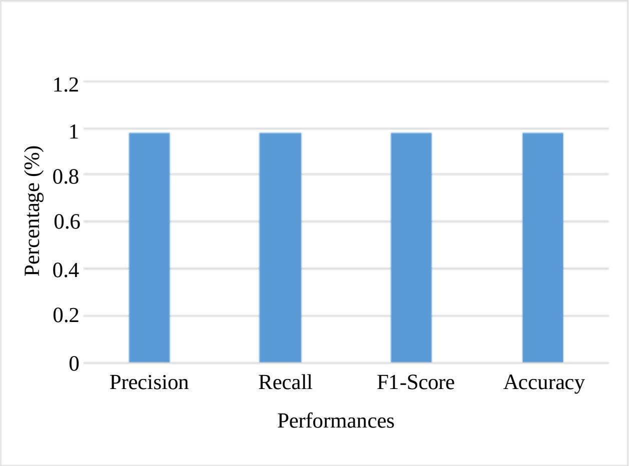

Fig. 6 Performance metrics

4.3. Comparative Analysis

Various approaches have been explored in audio signal processing for speech recognition, each employing distinct features to achieve accurate results. Segokar and Sircar utilized the Continuous Wavelet Transform alongside Prosodic Features, yielding an accuracy of 60.1%. Zeng et al. adopted a different path, relying solely on Spectrograms, achieving a slightly higher accuracy of 65.97%. Bhavan et al. took a comprehensive approach, incorporating MFCCs, spectral centroids, and MFCC derivatives, resulting in a notable improvement with an accuracy of 75.69%. However, the proposed model, integrating Capsule Networks (Caps net) with Channel Attention Mechanism (CBAM), surpasses its counterparts with an impressive accuracy of 98.57%. This innovative fusion of Capsule Networks and CBAM suggests a more robust and effective strategy for speech recognition, outperforming the conventional methods that rely on a single set of features. The proposed model's superior accuracy highlights the potential of combining advanced neural network architectures with attention mechanisms in audio signal processing, paving the way for more accurate and efficient speech recognition systems. The following Table 4 shows the comparative analysis of the proposed approaches. Table 5 Presents a comparative analysis of the accuracy achieved by the proposed method on three distinct datasets: EMODB, TESS, and RAVDESS. The proposed method achieves the highest accuracy of 98.57% on the RAVDESS dataset, which is marginally higher than the accuracy on TESS (98.00%) and significantly better than the accuracy on EMODB (87.86%).

Table 4. Comparative analysis

Approaches	Features used	Accuracy
Zeng et al. [23]	Spectrogram	65.97
Segokar and Sircar [24]	Continuous wavelet Transform, Prosodic Features	60.1
Bhavan et al. [25]	MFCCs, spectral centroids and MFCC derivatives	75.69
Proposed model (Caps net with CBAM)	MEDC, MFCC, DCT, FFT	98.57

Table 5. Dataset comparison

Dataset	Accuracy (%)
EMODB [26]	87.86
TESS [27]	98.00
RAVDESS (Proposed)	98.57

Table 6. Computational efficiency comparison

Techniques	Time (sec)
GMM [9]	9.31
SVM [9]	7.35
MLP based on ANN [9]	5.13
Proposed	2.15

The substantially lower performance of EMODB and the marginal improvement over TESS justify selecting RAVDESS as the primary dataset for detection. This comparison demonstrates the effectiveness of the proposed method across datasets with varying characteristics. Table 6 highlights the computational efficiency of various techniques regarding processing time (in seconds). The proposed model outperforms other methods, achieving the shortest processing time of 2.15 seconds, showcasing its superior efficiency. In comparison, traditional methods such as GMM and SVM require 9.31 and 7.35 seconds, respectively, while the MLP based on ANN achieves a moderate time of 5.13 seconds. This significant reduction in processing time underscores the proposed model's optimized architecture, making it highly suitable for speech emotion recognition tasks without compromising accuracy or robustness.

4.4. Discussion

This study presents a CapsNet and CBAM-based model for speech emotion recognition, demonstrating excellent performance on the RAVDESS dataset. The model achieves an accuracy of 98.57%, with precision, recall, and F1-score values all at 0.98, indicating strong classification capabilities. The training time was optimized to 250 seconds with the combination of CapsNet and CBAM, outperforming individual models. The CapsNet model achieved an accuracy of 71.76% with a training time of 200 seconds, while the CBAM layer model achieved 71.43% accuracy with a training time of 272 seconds. The combined CapsNet with the CBAM model showcased a significant leap in accuracy, reaching 98.57%. The dataset comparison shows that the proposed model achieves the highest accuracy on RAVDESS (98.57%), slightly outperforming TESS (98.00%) and significantly better than EMODB (87.86%).

This justifies the selection of RAVDESS as the primary dataset for emotion detection. Additionally, the proposed model excels in computational efficiency, processing in 2.15 seconds, which is faster than traditional methods such as GMM (9.31 seconds), SVM (7.35 seconds), and MLP-based ANN (5.13 seconds). The CapsNet-CBAM model offers a highly accurate, efficient, and robust solution for speech emotion recognition tasks, with promising implications for real-time applications. The proposed model achieved superior results by combining Capsule Networks (CapsNet) with the Convolutional Block Attention Module (CBAM). CapsNet's ability to capture hierarchical spatial features and CBAM's focus on salient regions enhanced feature extraction and improved emotion recognition accuracy. The synergy between CapsNet and CBAM led to a significant accuracy of 98.57%, outperforming standalone models and traditional methods like GMM, SVM, and MLP. The model demonstrated resilience to variations in voice quality, background noise, and speaker characteristics, making it highly robust and efficient for real-world applications. It also achieved faster training times and reduced computational costs, highlighting its practicality for speech-emotion recognition tasks. By demonstrating better performance than traditional methods and ensuring high computational efficiency, this research advances emotion recognition systems for applications in human-computer interaction and psychological studies, laying the groundwork for future progress in this area.

5. Conclusion

The integration of Capsule Networks (CapsNet) with a CBAM in audio emotion recognition has shown significant progress, with an accuracy rate of 98.57%. This innovative approach addresses the limits of traditional CNNs handling hierarchical and spatial relationships in audio data. The introduction of CBAM increases the model's capacity to focus on crucial features and suppress irrelevant information, contributing to its overall robustness and performance. This integration demonstrates the potential of leveraging advanced deep learning architectures to improve emotion recognition systems. The success of this system holds promise for applications in human-computer interaction, virtual assistants, and affective computing. However, further research is needed to fine-tune model parameters, explore larger datasets, and investigate real-world applications. A limitation of our work is that the model's performance under various noise conditions was not extensively studied, which may affect its effectiveness in environments with background noise or fluctuating audio quality. Addressing computational complexity and real-time processing requirements is also crucial for practical deployment. Overall, the combination of CapsNet and CBAM in audio emotion recognition represents a significant step forward, with potential applications positively impacting human-machine interaction and communication.

Reference

[1] Zhiheng Xi et al., "The Rise and Potential of Large Language Model Based Agents: A Survey," *arXiv*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Kat Roemmich, and Nazanin Andalibi, "Data Subjects' Conceptualizations of and Attitudes toward Automatic Emotion Recognition-Enabled Wellbeing Interventions on Social Media," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, pp. 1-34, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Fuzhen Zhuang et al., "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43-76, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Jingyuan Zhao et al., "Specialized Deep Neural Networks for Battery Health Prognostics: Opportunities and Challenges," *Journal of Energy Chemistry*, vol. 87, pp. 416-438, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Alan Tan Wei Min, Abhishek Gupta, and Yew-Soon Ong, "Generalizing Transfer Bayesian Optimization to Source-Target Heterogeneity," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 4, pp. 1754-1765, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Andrzej Janowski, "Natural Language Processing Techniques for Clinical Text Analysis in Healthcare," *Journal of Advanced Analytics in Healthcare Management*, vol. 7, no. 1, pp. 51-76, 2023. [Google Scholar] [Publisher Link]

[7] Fan Zhang et al., "Speech-Driven Personalized Gesture Synthetics: Harnessing Automatic Fuzzy Feature Inference," *IEEE Transactions on Visualization and Computer Graphics*, vol. 10, no. 10, pp. 6984-6996, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Yan Wang et al., "A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances," *Information Fusion*, vol. 83-84, pp. 19-52, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Sandeep Kumar et al., "Multilayer Neural Network Based Speech Emotion Recognition for Smart Assistance," *Computers, Materials & Continua*, vol. 74, no. 1, pp. 1523-1540, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] A. Christy et al., "Multimodal Speech Emotion Recognition and Classification Using Convolutional Neural Network Techniques," *International Journal of Speech Technology*, vol. 23, pp. 381-388, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ala Saleh Alluhaidan et al., "Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network," *Applied Sciences*, vol. 13, no. 8, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Rashid Jahangir et al., "Convolutional Neural Network-based Cross-Corpus Speech Emotion Recognition with Data Augmentation and Features Fusion," *Machine Vision and Applications*, vol. 33, no. 3, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Kudakwashe Zvarevashe, and Oludayo O. Olugbara, "Recognition of Speech Emotion Using Custom 2D-Convolution Neural Network Deep Learning Algorithm," *Intelligent Data Analysis*, vol. 24, no. 5, pp. 1065-1086, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Shibani Hamsa et al., "Emotion Recognition from Speech Using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier," *IEEE Access*, vol. 8, pp. 96994-97006, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Chawki Barhoumi, and Yassine BenAyed, "Real-Time Speech Emotion Recognition Using Deep Learning and Data Augmentation," vol. 58, no. 2, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Mustaqeem Khan et al., "MSER: Multimodal Speech Emotion Recognition Using Cross-Attention with Deep Fusion," *Expert Systems with Applications*, vol. 245, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Mengsheng Wang et al., "Design of Smart Home System Speech Emotion Recognition Model based on Ensemble Deep Learning and Feature Fusion," *Applied Acoustics*, vol. 218, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Bachchu Paul et al., "Machine Learning Approach of Speech Emotions Recognition Using Feature Fusion Technique," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 8663-8688, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang, "Transforming Auto-Encoders," *Artificial Neural Networks and Machine Learning-ICANN 2011: 21st International Conference on Artificial Neural Networks*, Espoo, Finland, Berlin Heidelberg, pp. 44-51, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[20] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton, "Dynamic Routing between Capsules," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Google Scholar]

[21] Sanghyun Woo et al., "CBAM: Convolutional Block Attention Module," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3-19, 2018. [Google Scholar] [Publisher Link]

[22] Yu Wang, Dejun Ning, and Songlin Feng, "A Novel Capsule Network Based on Wide Convolution and Multi-Scale Convolution for Fault Diagnosis," *Applied Sciences*, vol. 10, no. 10, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Yuni Zeng et al., "Spectrogram Based Multi-Task Audio Classification," *Multimedia Tools and Applications*, vol. 78, pp. 3705-3722, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[24] Kunxia Wang et al., "Speech Emotion Recognition Using Fourier Parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69-75, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[25] Anjali Bhavan et al., "Bagged Support Vector Machines for Emotion Recognition from Speech," *Knowledge-Based Systems*, vol. 184, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[26] Hua Zhang et al., "Pre-Trained Deep Convolution Neural Network Model with Attention for Speech Emotion Recognition," *Frontiers in Physiology*, vol. 12, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[27] Samson Akinpelu, Serestina Viriri, and Adekanmi Adegun, "An Enhanced Speech Emotion Recognition Using Vision Transformer," *Scientific Reports*, vol. 14, no. 1, 2024. [CrossRef] [Google Scholar] [Publisher Link]