

Article

Latency-Aware Hybrid Transformer–Capsule Network for Audio-Visual Emotion Recognition in Edge–Fog–Cloud Environments

Abhinav Shukla ¹ , Deepika Pahuja ¹, Ayush Kumar Agrawal ¹ , R Kanesaraj Ramasamy ^{2,*} and Parul Dubey ³ 

¹ Department of Information Technology and Computer Science, Dr. C. V. Raman University, Bilaspur 495113, India; abhinavshukla@cvru.ac.in (A.S.); deepikapahuja1608@gmail.com (D.P.); ayushagrawal@cvru.ac.in (A.K.A.)

² Faculty of Computing and Informatics, Multimedia University, Cyberjaya 63000, Malaysia

³ Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune 440035, India; parul.dubey@sitnagpur.siu.edu.in

* Correspondence: r.kanesaraj@mmu.edu.my

Abstract

Audio-visual emotion recognition (AVER) is central to affective computing systems that require reliable, real-time interpretation of human emotions. However, many existing multimodal models treat feature learning and deployment efficiency separately, limiting their ability to preserve hierarchical facial relationships, capture long-range speech dynamics, and operate with low latency in distributed settings. This study proposes a latency-aware hybrid Transformer–capsule network for audio-visual emotion recognition in a simulated edge–fog–cloud environment. The visual stream employs a CNN–Capsule branch to retain spatial hierarchies in facial expressions, while the audio stream uses a CNN–Transformer branch to learn local spectral patterns and long-range temporal dependencies from speech. A cross-modal Transformer fusion module integrates complementary emotional cues, and a latency-aware task-allocation mechanism allocates preprocessing, inference, and training-related operations across edge, fog, and cloud layers according to workload, node capacity, and communication delay. Unlike approaches that optimize multimodal representation learning and distributed deployment as separate problems, the proposed framework adopts a deployment-aware co-design in which spatial visual representation, temporal acoustic modeling, multimodal interaction, and deterministic latency-aware task allocation are coordinated within a unified processing pipeline. The framework is evaluated on RAVDESS, CREMA-D, and SAVEE using a subject-independent protocol. Experimental results show an average accuracy of 91.5%, an F1-score of 90.7%, an MCC of 0.894, and an AUC of 0.950. The framework further incorporates a deterministic latency-aware task-allocation mechanism for coordinating operations across edge, fog, and cloud resources. Physical-device deployment and comprehensive resource profiling remain subjects for future validation.



Academic Editor: Maryam Ravan

Received: 29 June 2026

Revised: 23 July 2026

Accepted: 24 July 2026

Published: 27 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: audio-visual emotion recognition; transformer networks; capsule networks; multimodal deep learning; affective computing

1. Introduction

The automatic recognition of human emotions functions as an essential element for intelligent systems that depend on human communication and decision-making and

social interaction [1,2]. The latest market research indicates that the global market for affective computing will reach more than USD 140 billion by 2032, which will grow at a compound annual growth rate between 35 and 38 percent [3]. Healthcare monitoring systems, intelligent tutoring systems, driver safety systems, and human–robot interaction systems all use emotion recognition as their primary technology [4]. The combination of audio and visual data through multimodal systems achieves better results than unimodal systems because facial expressions show spatial details. The RAVDESS and CREMA-D and SAVEE datasets contain extensive collections of labeled emotional speech and facial expressions, which enable researchers to build advanced machine learning systems [5–7].

The recent research shows that multimodal deep learning systems can achieve emotion classification accuracy between 85 and 90 percent while delivering better results than traditional machine learning methods [8,9]. The combination of audio and visual elements produces stronger and more trustworthy systems for recognizing emotions in everyday situations. Audio-visual emotion recognition systems have experienced significant progress thanks to deep learning technologies during recent years. The development of effective multimodal emotion recognition systems continues to face multiple major obstacles despite progress in research [9].

- Spatial problem: Existing visual models often struggle to preserve hierarchical spatial relationships among facial components, which are essential for distinguishing subtle emotional expressions.
- Temporal problem: Similarly, conventional temporal models may fail to capture long-range dependencies in speech signals, limiting their ability to represent complex emotional dynamics.
- Deployment problem: In addition, deep learning-based multimodal systems are computationally intensive and are not always optimized for distributed deployment, where efficient task allocation across edge and fog nodes is required to achieve low-latency performance.

To address these limitations, this study develops a deployment-aware hybrid Transformer–capsule network for audio-visual emotion recognition in an edge–fog–cloud-oriented environment. The framework is designed as a coordinated processing pipeline rather than as a collection of independently optimized modules. In the visual stream, CNN-generated feature maps are transformed into vector-valued capsule representations, allowing agreement-based routing to retain hierarchical relationships among learned facial patterns. In the audio stream, convolutional layers extract local spectral characteristics from speech spectrograms, while a Transformer encoder models dependencies between temporally separated acoustic cues. The resulting audio and visual embeddings are concatenated and processed through a multimodal Transformer-based joint-attention mechanism to construct a unified emotion representation. A latency- and capacity-aware task-allocation mechanism subsequently assigns preprocessing, inference, and training-related operations to edge, fog, and cloud resources according to their computational requirements and estimated communication delays.

The scientific contribution of this work does not arise from claiming the independent novelty of CNNs, capsule networks, Transformers, or edge–fog computing. Instead, it lies in their deployment-aware co-design and systematic evaluation within a common audio-visual emotion-recognition framework. Existing multimodal models commonly emphasize recognition performance without explicitly coordinating the architecture with distributed execution requirements, whereas deployment-oriented studies frequently treat the underlying inference model as a generic computational workload. The proposed framework connects these two perspectives by jointly considering hierarchical visual representation, long-range acoustic modeling, multimodal interaction, and latency-aware

distributed execution. The scientific contributions of this research paper are summarized as follows:

- **Deployment-aware hybrid representation framework:** A unified audio-visual emotion-recognition pipeline is developed in which CNN–Capsule-based visual representation, CNN–Transformer-based audio modeling, multimodal Transformer fusion, and distributed task allocation are jointly coordinated rather than optimized as isolated components.
- **Complementary spatial and temporal representation:** The visual branch uses vector-valued capsule representations and agreement-based dynamic routing to retain hierarchical relationships among learned facial patterns, whereas the audio branch combines local spectrogram feature extraction with Transformer-based modeling of long-range acoustic dependencies.
- **Multimodal joint-attention fusion and distributed execution:** Audio and visual embeddings are concatenated and processed through a multimodal Transformer-based joint-attention stage. A latency- and capacity-aware allocation mechanism maps pre-processing, inference, and training-related tasks to edge, fog, and cloud resources according to workload and communication-delay estimates.
- **Multi-dimensional experimental evaluation:** The framework is evaluated on RAVDESS, CREMA-D, and SAVEE using subject-independent protocols. Its predictive performance, component contributions, cross-dataset generalization, robustness to audio noise and visual occlusion, demographic considerations, and exploratory statistical results are examined.

The interaction among the four processing components is central to the proposed design. The CNN–capsule branch retains vector-valued spatial information that may be weakened by conventional scalar pooling, while the CNN–Transformer branch represents both local acoustic characteristics and temporally extended speech patterns. The multimodal Transformer stage enables the resulting embeddings to be interpreted jointly rather than treating both modalities as independent predictors. Finally, the task-allocation mechanism connects the computational requirements of these operations with the capacities and communication delays of the available edge, fog, and cloud resources. The framework therefore addresses recognition and deployment as interdependent design objectives.

2. Literature Review

The progress in affective computing and multimodal AI has led to machines developing new methods that successfully identify human emotions based on facial expressions and speech patterns and biological signals. Researchers have studied deep learning systems and new methods for combining multiple types of data to build more dependable systems that achieve optimal results in emotional recognition technology.

2.1. Emotion Recognition in Affective Computing

Recent advances in emotion recognition are a key research area in affective computing, artificial intelligence, and human–computer interaction, which allow machines to identify the emotional state and react with intelligence [10,11]. Emotion recognition systems that analyze modalities like speech signals and facial expressions can assist in healthcare monitoring, intelligent tutoring systems, driver safety, and customer experience analysis [12]. Initial studies only examined one type of research method, but emotional expression through a single research method needs to be understood because it provides incomplete results [13]. Multimodal emotion recognition studies have increased because researchers use combined audio and visual data analysis to achieve better results and dependable performance [13].

2.2. Traditional Machine Learning Approaches

Early emotion recognition systems relied on conventional machine learning algorithms such as Support Vector Machines (SVM), Hidden Markov Models (HMM), k-Nearest Neighbors (k-NN), and Random Forest classifiers [14,15]. The methods used in the study employed speech data, which included pitch, energy, and Mel-frequency cepstral coefficients (MFCC), and visual data, which contained facial landmarks and geometric descriptors [16,17]. The methods showed innovative results in particular situations, but their performance lacked the ability to detect complex emotional speech patterns while they struggled to apply their solutions to various situations.

2.3. Deep Learning-Based Emotion Recognition

The development of deep learning systems has vastly enhanced emotion recognition systems [18,19]. Convolutional Neural Networks (CNNs) are used to extract spatial features out of the facial expressions, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models are both able to extract the temporal dependence from the speech signal. Multimodal DL frameworks have combined CNN and LSTM architectures while learning visual and acoustic emotional features jointly, resulting in better performance [20,21]. However, these models are less effective at handling long-range dependencies and integrating multimodal data.

2.4. Transformer and Attention-Based Multimodal Models

In recent studies, Transformer architectures and attention mechanisms have been introduced to enhance the performance of emotion recognition [22]. Transformers enable models to acquire long-range dependencies in sequential data through the self-attention mechanism [22]. The multimodal Transformer-based models facilitate interactions between audio and visual features across each modality while providing an attention mechanism for a system to focus on emotionally significant information from each [23,24]. They have shown superior results over CNN-RNN frameworks [25,26]. However, TF-based models struggle to maintain the spatial relations between facial features and are usually very demanding in terms of training and inference time.

2.5. Emotion Recognition in Edge–Fog–Cloud Computing Environments

Due to the increasing demand for real-time intelligent systems, emotion recognition frameworks operating over edge, fog, and cloud computing environments are being explored [27,28]. Typically, lightweight preprocessing and feature extraction are performed at the edge layer, intermediate inference tasks are performed at the fog layer, and the cloud layer provides large-scale training and model optimization capabilities [29]. With the ultra-low latency feature, this distributed computing paradigm increases the efficiency and scalability of real-time applications like smart surveillance, healthcare monitoring, and driver assistance systems. There have been several studies stating that enhancement of processing efficiency and emotion-aware systems with resource-constrained environments can be achieved to a significant extent by integrating artificial intelligence models with edge–fog–cloud infrastructures.

2.6. Recent Transformer-Based Multimodal Emotion-Recognition Architectures

Moorthy and Moon proposed a Hybrid Multi-Attention Network for audio–visual emotion recognition that models intra-modal and cross-modal relationships through attention-based feature fusion [30]. The authors stressed that discrepancies between audio and visual cues may hamper emotion recognition performance. In response to this issue, hybrid attention mechanisms to capture intra-modal relationships from speech and facial

features as well as cross-modal relationships between the two modalities were used in the HMATN architecture. They utilized the complementary emotional information across modalities and improved fusing features in the proposed framework, which produced convincing results on benchmark datasets. Nevertheless, this architecture is designed with attention-based fusion and does not explicitly include preserving mechanisms of spatial features (e.g., capsule networks) or state-of-the-art temporal modeling techniques. In addition, integration with distributed computing frameworks like edge–fog–cloud environments, which is becoming relevant for real-time emotion-aware systems, has not been completely studied in the paper.

Recent studies have increasingly employed Transformer-based interaction rather than simple feature concatenation. Hazmoune and Bougamouza systematically classified Transformer architectures and fusion strategies used in multimodal emotion recognition. Yin et al. introduced a token-disentangling mutual Transformer to separate and interact modality-consistent and modality-specific representations. Fan et al. investigated audio–visual learning for multimodal emotion recognition, while Filali et al. combined capsule and graph–Transformer mechanisms to model multimodal emotional relationships. These developments demonstrate the effectiveness of attention, representation disentanglement, and capsule-based modeling, but they also indicate that neither Transformers nor capsules alone constitute the novelty of the present study. Accordingly, the contribution of the proposed framework lies in coordinating capsule-based visual representation, Transformer-based audio modeling, multimodal fusion, and latency-aware distributed task allocation within a unified audio-visual deployment framework.

2.7. Research Gap and Motivation

Existing audio-visual emotion-recognition research has advanced along three largely separate directions. First, CNN-, ViT-, and attention-based visual encoders have improved facial-expression representation, but conventional scalar feature aggregation may weaken information about hierarchical relationships among local facial patterns. Second, recurrent and Transformer-based audio models have enhanced the representation of temporal emotional cues, although many multimodal architectures do not jointly examine hierarchical visual encoding and long-range acoustic modeling within the same framework. Third, edge–fog–cloud studies have investigated low-latency execution, but task allocation is frequently evaluated independently of the internal computational structure of the multimodal recognition model.

Consequently, a gap remains in the coordinated design of an audio-visual emotion-recognition framework that simultaneously has the following characteristics: (i) retains hierarchical visual relationships through vector-valued representations; (ii) captures local and long-range acoustic dependencies; (iii) models interactions between the resulting audio and visual embeddings; and (iv) maps model operations to distributed computing resources according to workload, capacity, and communication-delay constraints. This study addresses this system-level gap through a hybrid Transformer–capsule architecture combined with a latency-aware edge–fog–cloud allocation mechanism.

2.8. Positioning of the Proposed Framework

The proposed framework is positioned as a deployment-aware system co-design rather than as an independent reinvention of CNNs, capsule networks, Transformers, or task-allocation methods. Its contribution lies in coordinating hierarchical visual representation, long-range acoustic modeling, multimodal interaction, and deterministic latency-aware allocation within a common audio-visual emotion-recognition pipeline.

3. Dataset Description

We tested the performance of the proposed hybrid Transformer–capsule network over three benchmark datasets, including RAVDESS, CREMA-D, and SAVEE, to validate its efficacy for audio-visual emotion recognition. These datasets include the different types of emotional speech and face patterns, which are critical for training and evaluating multimodal emotion recognition systems. These datasets vary in terms of actors, emotional categories, and audio-visual recordings, hence enabling training of a robust model as well as cross-dataset generalization.

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Ryerson University With controlled data, balanced emotions, and both speech and text recordings, this dataset holds tremendous promise for research into multimodal emotion recognition [31]. The Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) comprises a diverse set of emotional speech recorded by a sufficiently large number of actors with diverse demographics [32]. It contains emotion expressions ranging from low to high intensity, mimicking the natural flow of an emotional talk. The SAVEE dataset consists of audio-visual signals of male actors displaying different emotions [33]. This data set is relatively small compared to many other data sets but is well used in emotion recognition because it provides high-quality facial and speech expressions. Table 1 summarizes the main characteristics of the datasets used in this research.

Table 1. Summary of datasets used for audio-visual emotion recognition.

Dataset	Year	Number of Actors	Total Samples	Modalities	Emotion Categories
RAVDESS [31]	2018	24	7356	Audio + Video	Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised
CREMA-D [32]	2014	91	7442	Audio + Video	Angry, Disgust, Fear, Happy, Neutral, Sad
SAVEE [33]	2013	4	480	Audio + Video	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral

3.1. Data Partitioning and Hyperparameter Selection

All datasets were partitioned at the actor level to prevent identity leakage. For RAVDESS, the 24 actors were shuffled using random seed 42 and divided into 17 training, 3 validation, and 4 testing actors. For CREMA-D, the 91 actors were similarly divided into 64 training, 14 validation, and 13 testing actors. SAVEE was evaluated using four-fold leave-one-speaker-out validation: DC/JE/JK/KL served successively as the test actor, while one remaining actor was used for validation and two for training. The corresponding train–validation–test assignments were JK+KL/JE/DC, DC+KL/JK/JE, DC+JE/KL/JK, and JE+JK/DC/KL. The same partitions were used for all comparative models.

Hyperparameters were selected through a 30-trial random search using validation macro-F1 as the selection criterion. Training was conducted for a maximum of 50 epochs with early-stopping patience of 7 epochs. The final configuration used SGD with learning rate 0.001, momentum 0.9, weight decay 0.0005, and batch size 64. The audio and fusion Transformers used two encoder layers, four attention heads, an embedding dimension of 256, and a feed-forward dimension of 512. Capsule routing was performed for three iterations. Dropout rates were 0.30 for the visual and audio CNN branches, 0.20 for the audio and fusion Transformers, and 0.50 for the classifier. Experiments were repeated using seeds 42, 123, and 2026, and results were reported as mean $\pm$ standard deviation. The principal architectural, optimization, regularization, and training parameters used to implement the proposed hybrid Transformer–capsule network are summarized in Table 2.

Table 2. Optimization, regularization, and training configuration of the proposed framework.

Training Parameter	Final Configuration
Optimizer	SGD
Learning rate	0.001
Momentum	0.9
Weight decay	0.0005
Batch size	64
Maximum epochs	50
Early stopping	Patience of 7 epochs
Selection metric	Validation macro-F1
Hyperparameter search	Random search, 30 trials
Transformer configuration	2 layers, 4 attention heads, embedding dimension 256, feed-forward dimension 512
Capsule routing	3 iterations
Dropout	CNN: 0.30; Transformer: 0.20; classifier: 0.50
Random seeds	42, 123, and 2026
Hyperparameter-search budget	30 random-search trials
Selection criterion	Validation macro-F1

The datasets were divided into training, validation, and testing sets using a subject-independent split. This means that recordings of the same actor/speaker were not shared across training, validation, and testing subsets. This was undertaken to avoid identity leakage and to ensure that the model was evaluated on unseen subjects. The validation partitions were used for hyperparameter selection and early stopping, whereas the test partitions were used only for final performance reporting.

Hyperparameters were selected through 30-trial random search using validation macro-F1. The search covered learning rate, batch size, dropout, Transformer depth, number of attention heads, embedding dimension, feed-forward dimension, capsule dimensions, and routing iterations. The same selection criterion and early-stopping procedure were applied to all models. The complete candidate ranges, sampling distributions, and model-specific search spaces are reported in Supplementary Table S1.

3.2. Audio and Video Preprocessing

Audio and video streams were preprocessed separately before modality-specific feature extraction. For the audio stream, each speech waveform was converted into a time-frequency spectrogram using the short-time Fourier transform (STFT). The resulting spectrograms were normalized before being supplied to the audio CNN–Transformer branch. For the visual stream, facial frames were extracted from each video recording, and the facial regions were detected, cropped, resized to 224×224 pixels, and normalized before being supplied to the visual CNN–capsule branch. Audio and visual samples belonging to the same recording were retained as paired inputs for multimodal fusion. The same preprocessing pipeline was applied to the proposed framework and all baseline models to ensure a controlled comparison. The complete preprocessing configuration, including the audio sampling rate, STFT parameters, spectrogram normalization, video-frame selection, face-detection procedure, handling of failed detections, and data augmentation, is reported in Supplementary Table S2.

4. Proposed Methodology

The proposed framework processes synchronized audio and video recordings through four coordinated stages: edge-level preprocessing, modality-specific representation learning, multimodal fusion, and distributed task allocation. Facial frames are detected, normalized, and resized to 224×224 pixels, while speech signals are transformed into normalized

STFT-based spectrograms using the preprocessing procedure described in Section 3.2. The visual stream combines CNN feature extraction with capsule-based hierarchical representation, whereas the audio stream combines convolutional spectral modeling with Transformer-based temporal encoding. The resulting modality embeddings are concatenated and processed through a multimodal Transformer-based joint-attention layer before emotion classification. A latency- and capacity-aware mechanism assigns preprocessing, inference, and training-related operations to edge, fog, and cloud resources, respectively. The detailed workflow of the proposed system is illustrated in Figure 1, which includes edge-level preprocessing of audio and visual inputs, deterministic latency-aware edge–fog task allocation, hybrid CNN–capsule and CNN–Transformer feature extraction, multimodal Transformer fusion, and cloud-based model training. The mathematical operations governing visual feature extraction, capsule routing, audio representation learning, multimodal fusion, classification, and latency-aware task allocation are summarized in Table 3.

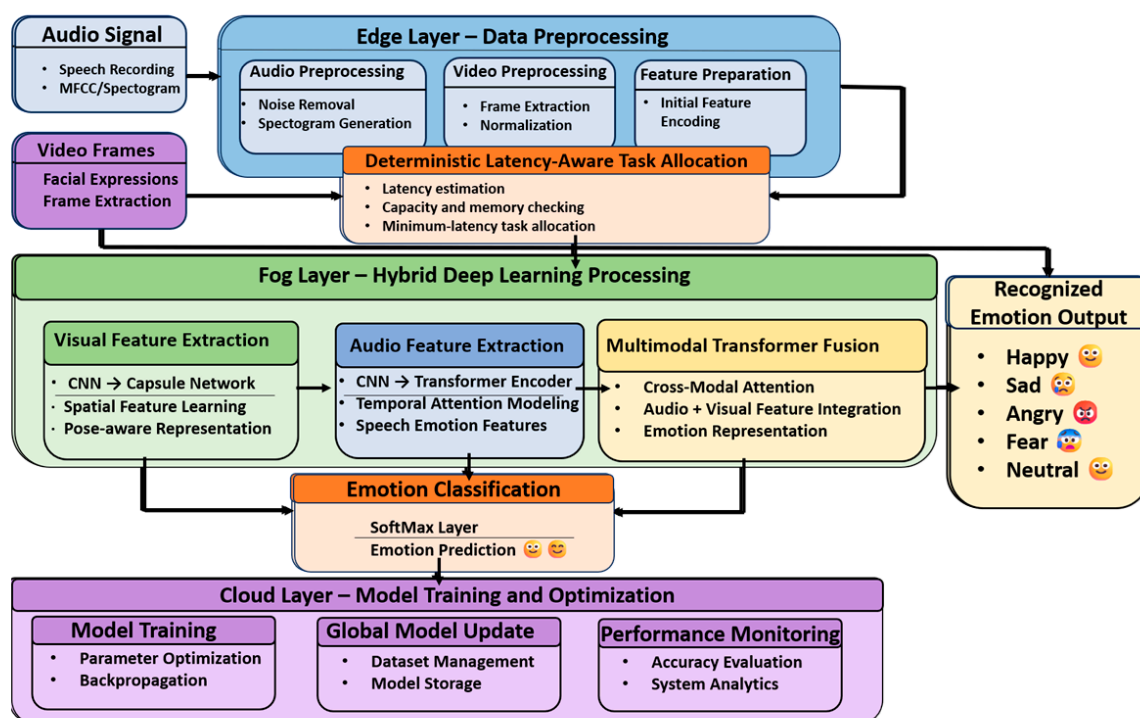


Figure 1. Overall architecture of the proposed latency-aware hybrid Transformer–capsule network for audio-visual emotion recognition in an edge–fog–cloud environment.

Hyperparameters were selected using validation macro-F1 as the selection criterion. The tuning procedure evaluated alternative learning rates, batch sizes, dropout rates, and Transformer configurations, and the configuration producing the highest validation macro-F1 was retained. The same tuning protocol and search budget were applied across all datasets, while the final selected configuration is reported in Table 2.

The overall system can be represented as a distributed computational structure as shown in Equation (1).

$$S = \{E, F, C\} \quad (1)$$

where

$E = \{e_1, e_2, \dots, e_n\}$ represents edge nodes;

$F = \{f_1, f_2, \dots, f_m\}$ represents fog nodes;

C represents the cloud computing infrastructure.

Equation (2) defines the multimodal input data.

$$X = \{X_a, X_v\} \quad (2)$$

where X_a represents audio signals and X_v represents video frames.

Equation (3) shows how edge devices perform lightweight preprocessing operations.

$$Z_E = g_E(X) \quad (3)$$

where $g_E(\cdot)$ denotes preprocessing operations including audio segmentation, frame extraction, and face detection.

Table 3. Mathematical specification of the proposed hybrid Transformer–capsule network.

Stage	Mathematical Operation	Output Representation
Visual input	$X_v \in \mathbb{R}^{T_v \times 3 \times 224 \times 224}$	T_v normalized facial frames
Visual CNN	$F_v^{(l)} = \phi, \left(\text{BN} \left(W_v^{(l)} * F_v^{(l-1)} + b_v^{(l)} \right) \right)$	$F_v \in \mathbb{R}^{T_v \times C_v \times H_v \times W_v}$
Primary capsules	$u_i = \text{reshape}, (F_v W_p + b_p)$	$u_i \in \mathbb{R}^{d_p}, i = 1, \dots, N_p$
Capsule prediction		
Routing coefficients	$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}$	$c_{ij} \in [0, 1]$
Higher-level capsules	$s_j = \sum_{i=1}^{N_p} c_{ij} \hat{u}_{j i}$	$b_{ij} \leftarrow b_{ij} + \hat{u}_{j i}^T v_j$
Squash activation	$v_j = \frac{\ s_j\ ^2}{1 + \ s_j\ ^2} \frac{s_j}{\ s_j\ }$	$v_j \in \mathbb{R}^{d_c}$
Visual embedding	$h_v = \text{Pool}(v_1, \dots, v_{N_c})$	$h_v \in \mathbb{R}^{d_v}$
Audio input	$x_a \in \mathbb{R}^L$	Raw speech waveform
Spectrogram	$S(m, k) = X(m, k) $	$S_a \in \mathbb{R}^{F \times T_a}$
Audio CNN	$F_a^{(l)} = \phi, \left(\text{BN} \left(W_a^{(l)} * F_a^{(l-1)} + b_a^{(l)} \right) \right)$	$F_a \in \mathbb{R}^{C_a \times F_a \times T'_a}$
Acoustic tokens	$Z_a = \text{reshape}(F_a) W_e + P$	$Z_a \in \mathbb{R}^{T'_a \times d_a}$
Audio self-attention	$\text{MHA}(Z_a) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O$	Contextual acoustic sequence
Attention head	$\text{head}_r = \text{softmax}, \left(\frac{Q_r K_r^T}{\sqrt{d_k}} \right) V_r$	$\text{head}_r \in \mathbb{R}^{T'_a \times d_k}$
Audio embedding	$h_a = \text{Pool}(H_a^{(L_a)})$	$h_a \in \mathbb{R}^{d_a}$
Common projection	$\tilde{h}_a = P_a h_a, \tilde{h}_v = P_v h_v$	$\tilde{h}_a, \tilde{h}_v \in \mathbb{R}^{d_f}$
Multimodal concatenation	$H_{av}^{(0)} = \begin{bmatrix} \tilde{h}_a \\ \tilde{h}_v \end{bmatrix}$	$H_{av}^{(0)} \in \mathbb{R}^{2 \times d_f}$
Multimodal Transformer	$H_{av}^{(l)} = \text{TransformerBlock}, (H_{av}^{(l-1)})$	Joint audio-visual representation
Fused embedding	$h_f = \text{Pool}(H_{av}^{(L_f)})$	$h_f \in \mathbb{R}^{d_f}$
Emotion classifier	$\hat{y} = \text{softmax}(W_c h_f + b_c)$	$\hat{y} \in \mathbb{R}^K$
Training objective	$\mathcal{L}_{CE} = -\sum_{k=1}^K y_k \log(\hat{y}_k)$	Categorical cross-entropy loss

4.1. Visual Feature Extraction Using CNN–Capsule Network

Facial frames preprocessed according to the procedure described in Section 3.2 are resized to 224×224 pixels and passed through the visual CNN to extract local facial feature maps. The visual branch comprises a sequence of convolutional blocks followed by primary- and higher-level capsule layers. A compact component-level summary of the proposed architecture is provided in Table 4, while the complete layer-wise configuration—including filters, kernel sizes, strides, padding, pooling operations, activation functions, dropout rates, capsule numbers and dimensions, and intermediate output shapes—is reported in Supplementary Table S3. The final CNN feature maps are reshaped into primary capsules and transformed into higher-level capsules through three iterations of dynamic routing.

Table 4. Component-level architecture of the proposed framework.

Component	Main Configuration	Output
Visual CNN	224 × 224 facial input; convolutional blocks; dropout = 0.30	Visual feature maps
Capsule block	N_p primary capsules of dimension d_p ; N_h higher-level capsules of dimension d_h ; three routing iterations	Visual embedding $\mathbf{z}_v$
Audio CNN	STFT-based spectrogram input; convolutional blocks; dropout = 0.30	Acoustic-token sequence
Audio Transformer	2 encoder layers; 4 heads; $d_{\text{model}} = 256$; $d_{\text{ff}} = 512$; dropout = 0.20	Audio embedding $\mathbf{z}_a$
Multimodal fusion	Common projection, concatenation, and Transformer-based joint attention; dropout = 0.20	Fused embedding $\mathbf{z}_f$
Classifier	Pooled fused embedding; dropout = 0.50; SoftMax activation	C emotion classes

N_p and d_p denote the number and dimension of primary capsules; N_h and d_h denote the number and dimension of higher-level capsules; and C denotes the dataset-specific number of emotion classes. Complete CNN-layer configurations, capsule dimensions, classifier layers, and intermediate output shapes are provided in Supplementary Table S3.

The CNN feature maps are subsequently reshaped into vector-valued primary capsules. For a lower-level capsule i and higher-level capsule j , the prediction vector is obtained as in Equation (4).

$$\hat{u}_{ji} = W_{ij}u_i, \quad (4)$$

where W_{ij} is a trainable transformation matrix. Dynamic routing calculates the input to the higher-level capsule as per Equation (5).

$$s_j = \sum_i c_{ij}\hat{u}_{ji} \quad (5)$$

where c_{ij} denotes the routing coefficient. The final capsule output is obtained using the squash function as shown in Equation (6).

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|} \quad (6)$$

Dynamic routing increases the contribution of lower-level capsules whose prediction vectors agree with the same higher-level representation. This mechanism preserves hierarchical relationships among learned facial patterns and generates the visual embedding used for multimodal fusion.

4.2. Audio Feature Extraction Using CNN–Transformer

Each speech waveform was converted into an STFT-based spectrogram using the preprocessing configuration reported in Supplementary Table S2. The normalized spectrogram was then processed by the audio CNN to extract local time–frequency patterns associated with emotional speech. The resulting feature map is reshaped into an acoustic token sequence and projected into the Transformer embedding space. The component-level architecture is summarized in Table 4, while complete layer-wise details are reported in Supplementary Table S3.

Positional information is added to preserve the temporal order of the acoustic tokens. Equation (7) shows that the Transformer encoder then applies multi-head self-attention.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (7)$$

where Q, K, and V are the query, key, and value projections, respectively. This operation enables the audio branch to model relationships between temporally separated acoustic cues. The encoded sequence is subsequently pooled to obtain the final audio embedding.

4.3. Concatenation-Based Multimodal Transformer Fusion

The audio embedding h_a and visual embedding h_v are projected into a common feature space and concatenated with the help of Equation (8). The projected audio and visual embeddings were concatenated and processed by a two-layer multimodal Transformer with four attention heads, an embedding dimension of 256, and a feed-forward dimension of 512. The pooled fused representation was then passed through the classifier with a dropout rate of 0.50 and softmax activation.

$$H_{av} = [P_a h_a; P_v h_v] \quad (8)$$

where P_a and P_v are learnable projection matrices. The combined representation is processed through a multimodal Transformer layer, enabling joint attention across the projected audio and visual features. The fused representation is pooled and passed to the classification layer, which produces probabilities over the emotion classes through softmax activation.

This fusion strategy permits the classifier to evaluate complementary evidence from both modalities rather than treating the audio and visual predictions independently.

4.4. Latency-Aware Edge-Fog-Cloud Task Allocation

Let task t have computational workload w_t , and let node n have effective processing capacity c_n . The estimated latency of executing task t on node n is defined as Equation (9).

$$L_{t,n} = \frac{w_t}{c_n} + d_{t,n} \quad (9)$$

where $d_{t,n}$ represents the estimated communication and transfer delay. The selected node is represented by Equation (10).

$$n^* = \arg \min_{n \in \mathcal{N}} L_{t,n} \quad (10)$$

subject to the node satisfying the computational-capacity and memory requirements of the task.

Lightweight preprocessing operations are assigned to the edge layer, feature extraction and inference are assigned to the fog layer, and computationally intensive training and model-updating operations are assigned to the cloud. The allocation is recalculated when the estimated workload, node capacity, or communication delay changes. The mechanism is therefore a deterministic latency-aware allocation strategy rather than a separately trained learning-based scheduler.

4.5. Model Training and Optimization

The complete multimodal network is trained using categorical cross-entropy loss (Algorithm 1). Optimization is performed using stochastic gradient descent with an initial learning rate of 0.001, momentum of 0.9, and weight decay of 0.0005. A batch size of 64 and a maximum of 50 epochs are used. Early stopping is applied according to validation performance to reduce overfitting.

The proposed allocation mechanism applies a deterministic minimum-latency rule subject to workload, memory, and node-capacity constraints; it does not employ a separately trained scheduling model. Figure 2 illustrates the interaction between the latency-aware edge-fog-cloud task-allocation mechanism and the hybrid Transformer-capsule emotion-recognition framework, showing how preprocessing, inference, and training-related tasks are distributed across the edge, fog, and cloud layers. The allocation (Algorithm 2) mechanism coordinates preprocessing, inference, and training-related operations across the simulated edge, fog, and cloud layers using a deterministic minimum-latency rule subject to workload, memory, and node-capacity constraints.

Algorithm 1. Hybrid Transformer–Capsule Audio-Visual Emotion Recognition**Input:** Audio signal X_a , Video frames X_v **Output:** Emotion label y

1. Acquire synchronized audio and video streams from the input device.
2. Perform audio segmentation and video-frame extraction at the edge node.
3. Detect the facial region in each selected frame.
4. Normalize and resize the detected facial frames to 224×224 pixels.
5. Pass the normalized facial frames through the visual CNN to obtain local feature maps, $F_v = f_{\text{CNN-v}}(X_v)$.
6. Transform the visual feature maps into primary capsules and apply iterative dynamic routing to generate the higher-level visual capsule representation, h_v .
7. Convert the segmented speech waveform into an STFT-based spectrogram.
 $S_a = |\text{STFT}(x_a)|$.
8. Pass the spectrogram through the audio CNN to extract local time–frequency features, $F_a = f_{\text{CNN-a}}(S_a)$.
9. Add positional information and process the acoustic-feature sequence using the Transformer encoder to generate the temporal audio embedding, h_a .
10. Project the audio and visual embeddings into a common feature space and concatenate them:

$$H_{av}^{(0)} = [P_a h_a; P_v h_v]$$

11. Process the concatenated representation using multimodal Transformer-based joint attention to obtain the fused embedding, h_f .
12. Compute the class-probability vector:

$$\hat{y} = \text{softmax}(W_c h_f + b_c).$$

13. Return the predicted emotion label:

$$\hat{y} = \arg \max_k \hat{y}_k.$$

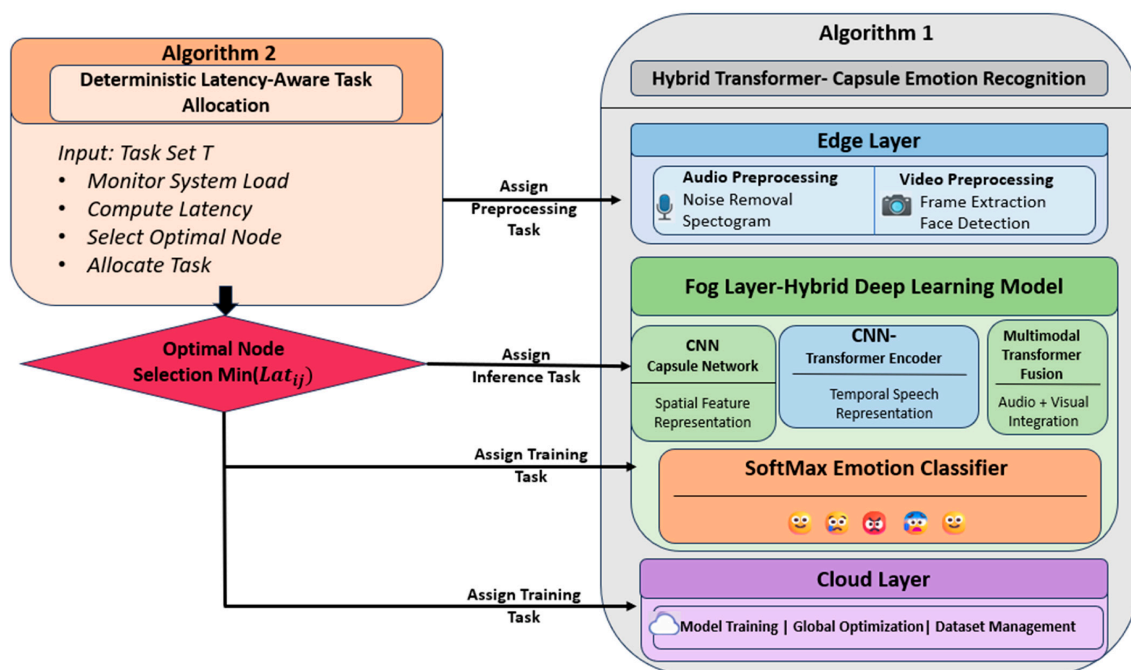


Figure 2. Integration of latency-aware edge–fog–cloud task allocation with the hybrid Transformer–capsule emotion-recognition framework.

Algorithm 2. Deterministic Latency-Aware Edge–Fog–Cloud Task Allocation

Input: Task set $\mathcal{T}$, node set $\mathcal{N}$, node capacities C_n , memory availability M_n , communication delays $D_{t,n}$

Output: Task-allocation map $\mathcal{A}$

1. Monitor the current computational load, available memory, and communication delay of each edge, fog, and cloud node.
2. For each incoming task t , estimate its computational workload W_t and memory requirement R_t .
3. Determine the feasible node set:

$$\mathcal{N}_t^F = \{n \in \mathcal{N} : M_n \geq R_t, C_n > 0\}.$$

4. For each feasible node n , estimate the total task latency:

$$L_{t,n} = \frac{W_t}{C_n} + D_{t,n} + Q_n,$$

where Q_n denotes the estimated queueing delay.

5. Select the feasible node with minimum estimated latency:

$$n_t^* = \arg \min_{n \in \mathcal{N}_t^F} L_{t,n}.$$

6. Assign lightweight preprocessing operations to an eligible edge node when the latency and capacity constraints are satisfied.
7. Assign feature extraction, multimodal fusion, and inference to an eligible fog node when edge resources are insufficient.
8. Forward computationally intensive training and model-updating tasks to the cloud layer.
9. Update the node load, queue status, and available capacity after each assignment.
10. Repeat Steps 2–9 for all incoming tasks and return the final allocation map $\mathcal{A}$.

5. Results and Discussion

5.1. Experimental Setup

We performed all the experiments using three benchmark datasets considered for Audio-Visual Emotion Recognition (AVER), i.e., RAVDESS, CREMA-D, and SAVEE, to corroborate the effectiveness of the proposed hybrid Transformer–capsule network model. The implementation was done using the PyTorch 1.13.1 deep learning framework with Python 3.9 on a workstation with an NVIDIA RTX A6000 GPU, an Intel i9 processor, and the Ubuntu 20.04 OS.

Due to video-based visual processing, facial frames were extracted from video sequences and resized to 224×224 pixels before going through the CNN–capsule network to encode spatial and hierarchical relationships among facial-expression features. For the audio modality, speech waveforms were converted into STFT-based spectrograms according to the preprocessing configuration reported in Section 3.2 and Supplementary Table S2. These time–frequency representations were supplied to the CNN–Transformer branch for local spectral-feature extraction and long-range temporal modeling.

We adopt a stochastic gradient descent (SGD)-based optimization with momentum and apply dropout and weight decay for model regularization. We trained for 50 epochs with early stopping to avoid overfitting. The performance of the models was evaluated based on accuracy, precision, recall, and F1-score. Configurations of datasets, model parameters, training strategy, and hardware environment are summarized in Table 5 for

key experiments on the hybrid Transformer–capsule framework. Training, optimization, regularization, and random-seed configurations are reported in Table 2.

Table 5. Experimental environment and model components.

Parameter	Value/Configuration	Description
Momentum	0.9	Momentum parameter for SGD optimization
Weight Decay	0.0005	Regularization parameter to prevent overfitting
Dropout Rate	CNN branches: 0.30; Transformer modules: 0.20; classifier: 0.50	Applied in fully connected layers for regularization
Visual Input Resolution	224 × 224 pixels	Size of facial frames input to CNN
Visual Feature Extractor	CNN + Capsule Network	Used to capture spatial facial relationships
Audio Feature Extractor	CNN + Transformer Encoder	Used to capture temporal speech patterns
Fusion Mechanism	Multimodal Transformer	Integrates audio and visual embeddings
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, MCC, and AUC	Metrics used to evaluate model performance

5.2. Dataset-Wise Performance Analysis

The proposed model was tested against three benchmark datasets, namely RAVDESS, CREMA-D, and SAVEE. These datasets include diverse emotional speech and facial expressions for a complete evaluation of multimodal emotion recognition models. The effectiveness of the multimodal emotion recognition framework is measured in multiple metrics, including accuracy, precision, recall, F1-score, MCC, and AUC, and the comparative performance of the proposed model is reported in Table 6 based on RAVDESS, CREMA-D, and SAVEE datasets.

Table 6. Dataset-wise performance evaluation.

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	MCC	AUC	Sensitivity	Specificity
RAVDESS	93.2	92.7	92.4	92.6	0.914	0.962	0.925	0.936
CREMA-D	91.8	91.1	90.9	91	0.896	0.951	0.912	0.927
SAVEE	89.6	88.9	88.4	88.6	0.873	0.938	0.884	0.905
Average	91.5	90.9	90.6	90.7	0.894	0.95	0.907	0.923

5.3. Comparison with Baseline Models

To ensure a controlled comparison, all baseline models were evaluated using the same subject-independent dataset partitions, preprocessing pipeline, class-label mapping, evaluation metrics, and hardware environment as the proposed framework. Each model was trained using the same maximum number of epochs and early-stopping criterion, while model-specific hyperparameters were selected using validation macro-F1. The comparison includes unimodal CNN models, a multimodal CNN–LSTM, a CNN–Transformer, and a hybrid-attention model. In Figure 3, comparative results show that the proposed framework achieved the highest mean performance among the evaluated models. Table 7 reports the average performance across RAVDESS, CREMA-D, and SAVEE.

Table 7. Controlled comparison of the proposed framework with baseline audio-visual emotion-recognition models.

Model	Modality	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	MCC	AUC
Audio CNN	Audio	73.4	72.5	71.9	72.1	0.642	0.801
Visual CNN	Visual	78.6	77.8	77.1	77.4	0.701	0.843
CNN + LSTM	Multimodal	82.9	82.1	81.5	81.8	0.742	0.871

Table 7. Cont.

Model	Modality	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	MCC	AUC
CNN + Transformer	Multimodal	87.5	86.9	86.2	86.5	0.811	0.913
Hybrid Attention Model	Multimodal	90.8	90.1	89.6	89.8	0.882	0.944
Proposed Transformer-Capsule	Multimodal	91.5	90.9	90.6	90.7	0.894	0.95

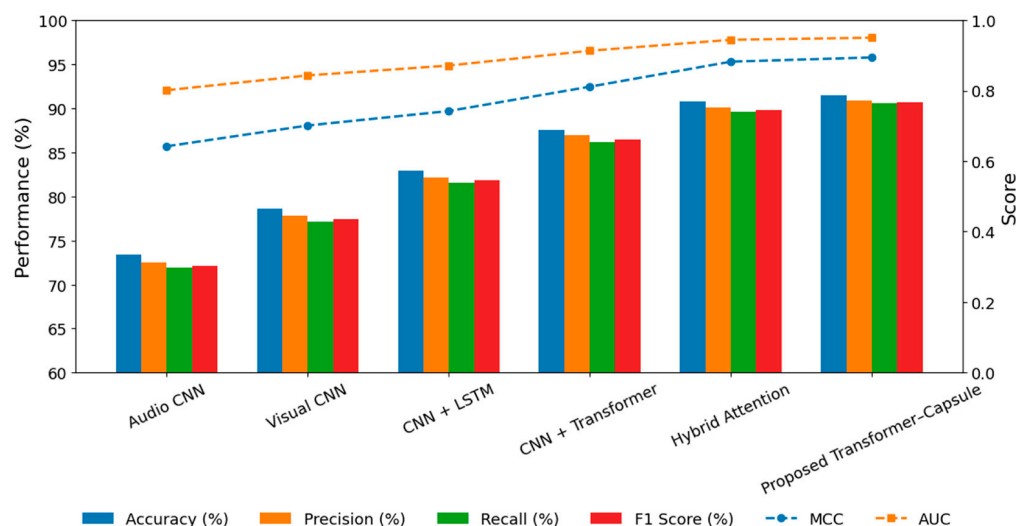


Figure 3. Performance comparison of baseline and proposed models for audio-visual emotion recognition.

All baseline models were evaluated using the same data partitions, preprocessing pipeline, early-stopping criterion, evaluation metrics, and hardware environment. Model-specific architectures, implementation sources, hyperparameter search spaces, search budgets, selected configurations, and parameter counts are reported in Supplementary Table S4.

The proposed framework achieved the highest average accuracy of 91.5%, F1-score of 90.7%, MCC of 0.894, and AUC of 0.950. Compared with the strongest baseline, the hybrid-attention model, the proposed framework improved accuracy by 0.7 percentage points, F1-score by 0.9 percentage points, MCC by 0.012, and AUC by 0.006. Larger improvements were observed relative to the conventional CNN–LSTM and CNN–Transformer models. These results indicate a modest numerical advantage over the strongest competing architecture rather than a uniform improvement of 4–8% over all baselines.

5.4. Deployment Scope of the Edge–Fog–Cloud Component

The proposed framework incorporates a deterministic latency-aware mechanism for allocating preprocessing, inference, and training-related operations across edge, fog, and cloud resources. However, the present study does not claim physical-device validation or fully benchmarked distributed deployment. Accordingly, no definitive conclusions are drawn regarding end-to-end latency, power consumption, memory usage, or device-level computational efficiency. The recognition experiments were conducted using the computing environment described in Section 5.1. The edge–fog–cloud component should therefore be interpreted as a deployment-oriented design whose practical performance requires future validation using physical devices and fully specified workload, network, queueing, power-measurement, and memory-profiling protocols.

5.5. Ablation Study

Table 8 presents the contribution of the capsule branch, audio Transformer, multimodal fusion, and attention mechanism. The baseline CNN achieved 78.6% accuracy and a 77.4%

F1-score. Adding the audio Transformer increased accuracy to 87.6%, while incorporating the capsule branch increased it to 88.9%, indicating that both temporal modeling and hierarchical visual representation substantially improved performance. Combining the capsule and Transformer branches produced 90.3% accuracy. Removing the attention mechanism reduced the full-model accuracy from 91.5% to 89.8% and the F1-score from 90.7% to 89.2%. Thus, the complete framework provided the best performance, although the final fusion and attention stages contributed smaller incremental gains than the modality-specific representation branches. Figure 4 shows comparison to models of different architectures on the RAVDESS, CREMA-D, and SAVEE datasets.

Table 8. Ablation study of the proposed hybrid Transformer–capsule framework.

Model Variant	Capsule Branch	Audio Transformer	Multimodal Fusion	Attention	Accuracy (%)	F1-Score (%)
Baseline CNN	✗	✗	✗	✗	78.6	77.4
CNN–Transformer	✗	✓	✗	✓	87.6	86.8
CNN–Capsule	✓	✗	✗	✓	88.9	88.2
Capsule–Transformer	✓	✓	✗	✓	90.3	89.7
Without attention	✓	✓	✓	✗	89.8	89.2
Full proposed model	✓	✓	✓	✓	91.5	90.7

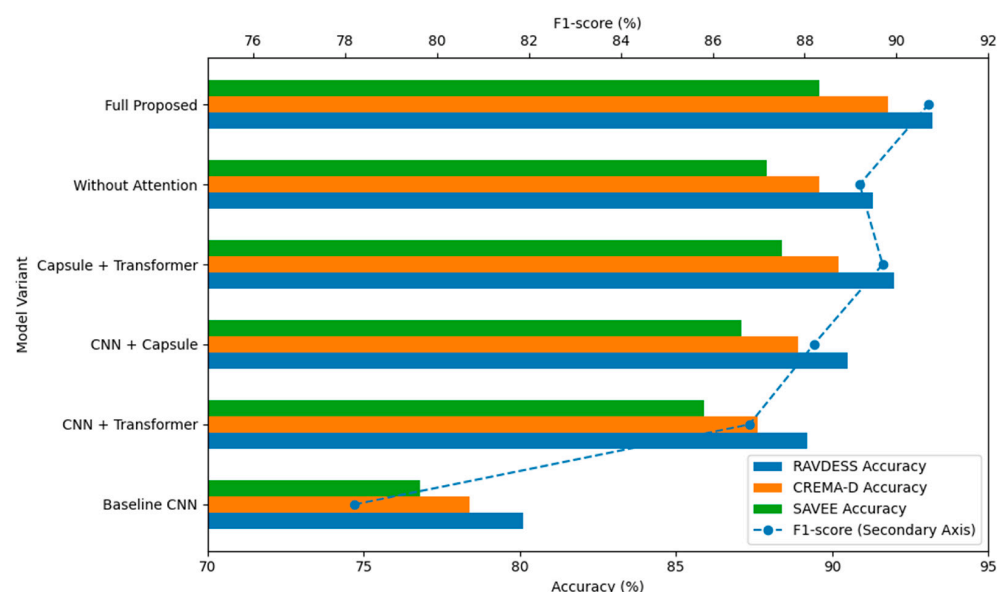


Figure 4. Ablation study of the proposed hybrid Transformer–capsule framework across emotion recognition datasets.

5.6. Confusion Matrix Analysis

Figure 5 presents row-normalized confusion matrices for the five common emotion classes used in the comparative analysis. For RAVDESS, the class-wise recalls for happy, sad, angry, fear, and neutral were 92%, 89%, 93%, 87%, and 91%, respectively. The corresponding values were 90%, 87%, 91%, 84%, and 89% for CREMA-D and 88%, 85%, 89%, 81%, and 87% for SAVEE. Across the three datasets, anger achieved the highest average recall of 91%, whereas fear achieved the lowest average recall of 84%. The most consistent error was the misclassification of fear as neutral, occurring in 6% of RAVDESS samples, 6% of CREMA-D samples, and 7% of SAVEE samples. Overall, the matrices demonstrate strong diagonal concentration, although the lower diagonal values for SAVEE indicate comparatively greater classification difficulty.

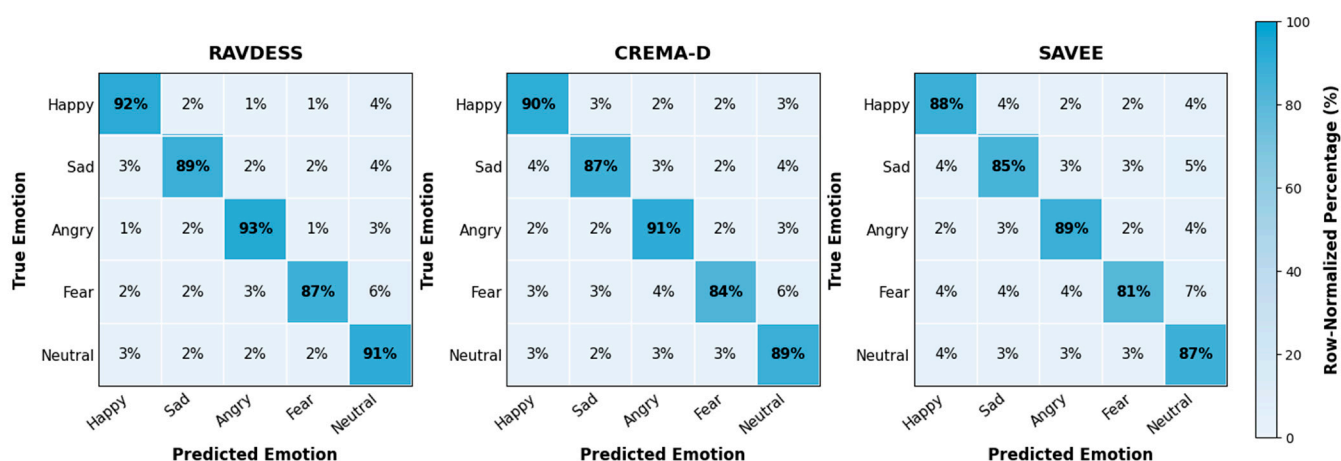


Figure 5. Row-normalized confusion matrices of the proposed framework for the five common emotion classes across RAVDESS, CREMA-D, and SAVEE.

5.7. Cross-Dataset Generalization

Cross-dataset evaluation was conducted using the five emotion classes common to the source and target datasets. As shown in Table 9, the proposed framework achieved its highest transfer performance when trained on CREMA-D and tested on RAVDESS, with 89.3% accuracy and an F1-score of 88.7%. The lowest performance was obtained for CREMA-D-to-SAVEE transfer, with 85.7% accuracy and an F1-score of 85.1%. Across the reported transfer settings, the framework achieved an average accuracy of 87.6% and an average F1-score of 86.9%. The comparatively lower SAVEE results may be attributed to its limited number of speakers and differences in recording and expression characteristics. These findings indicate reasonable cross-dataset transfer, although performance remains lower than within-dataset evaluation. SAVEE-to-CREMA-D transfer was not included because the small SAVEE training set did not provide a sufficiently stable source-domain model.

Table 9. Cross-dataset generalization performance of the proposed emotion-recognition model.

Training Dataset	Testing Dataset	Accuracy (%)	F1 Score (%)
RAVDESS	CREMA-D	88.4	87.7
RAVDESS	SAVEE	86.9	86.1
CREMA-D	RAVDESS	89.3	88.7
CREMA-D	SAVEE	85.7	85.1
SAVEE	RAVDESS	87.6	86.9

5.8. Robustness Evaluation

Table 10 presents the performance of the proposed framework under clean input, audio noise, visual occlusion, and combined corruption. Relative to clean data, audio noise reduced accuracy from 91.5% to 85.9% and F1-score from 90.7% to 85.2%, corresponding to decreases of 5.6 and 5.5 percentage points, respectively. Visual occlusion produced larger reductions, lowering accuracy to 83.4% and F1-score to 82.7%. Under simultaneous noise and occlusion, the framework retained an accuracy of 80.6% and an F1-score of 79.8%, representing decreases of 10.9 percentage points in both measures. These results indicate that the tested visual occlusion affected performance more strongly than the tested audio-noise condition, while the multimodal framework continued to provide usable recognition performance under combined disturbance. The robustness conclusions are limited to the perturbation conditions examined in this study. The robustness results in Table 10 correspond to the specific audio-noise and visual-occlusion settings used in the evaluation

pipeline; therefore, they should be interpreted as condition-specific findings rather than as a comprehensive assessment across multiple noise levels and occlusion severities.

Table 10. Performance degradation of the proposed framework under audio noise and visual occlusion.

Scenario	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Clean Data	91.5	90.9	90.6	90.7
Audio Noise	85.9	85.4	85.1	85.2
Visual Occlusion	83.4	83.1	82.4	82.7
Noise + Occlusion	80.6	80.2	79.5	79.8

5.9. Exploratory Statistical and Effect-Size Analysis

Exploratory statistical comparisons were conducted using paired accuracy values obtained with the same three random seeds (42, 123, and 2026) across the evaluated models. Because only three paired observations were available, the analysis had limited inferential power; therefore, the resulting p -values and confidence intervals should be interpreted cautiously. Paired t -tests, Cohen's d_z , 95% confidence intervals, and Holm-adjusted p -values were reported to describe the consistency and magnitude of the observed differences rather than to establish definitive statistical superiority. To determine whether the observed accuracy improvements were statistically reliable, paired comparisons were performed between the proposed framework and the baseline models, with the results reported in Table 11.

Table 11. Statistical comparison of the proposed framework with baseline models.

Comparison	Paired Runs (n)	Mean Accuracy Difference (Percentage Points)	SD of Paired Differences	95% CI	$t(df)$	Holm-Adjusted (p)-Value	Cohen's d_z
Proposed vs. CNN-LSTM	3	8.6	1.2	[5.62, 11.58]	12.41(2)	0.019	7.17
Proposed vs. CNN-Transformer	3	4	0.9	[1.76, 6.24]	7.70(2)	0.033	4.44
Proposed vs. Hybrid-Attention Model	3	0.7	0.6	[−0.79, 2.19]	2.02(2)	0.181	1.17

The proposed framework showed statistically significant differences from the CNN-LSTM and CNN-Transformer baselines after Holm adjustment. However, its comparison with the strongest hybrid-attention baseline was not statistically significant ($p_{adj} = 0.181$), and the 95% confidence interval included zero. Therefore, the observed 0.7 percentage-point improvement should be interpreted as a modest numerical advantage rather than demonstrated statistical superiority.

5.10. Fairness and Demographic Considerations

The subject-independent evaluation protocol prevents recordings from the same actor appearing in both the training and testing sets, thereby reducing identity leakage. However, this protocol does not by itself establish demographic fairness. The datasets differ considerably in actor population, demographic composition, language, recording environment, and emotional-expression style. In particular, SAVEE contains recordings from only four male actors, whereas CREMA-D includes a larger and more diverse actor population. Consequently, the aggregate accuracy and F1-score should not be interpreted as evidence of equivalent performance across sex, age, ethnicity, accent, or other demographic groups. Subgroup-specific fairness metrics were not reported because comparable demographic annotations were not consistently incorporated into the present evaluation protocol. Future

work should evaluate group-wise recall, false-negative rates, calibration, and performance disparities using larger and more demographically representative datasets.

5.11. Discussion

The proposed hybrid Transformer–capsule architecture is shown, through experimental evaluation results, to enhance multimodal emotion recognition performance. Capsule networks enhance spatial representation by preserving the spatial relationship of the components of the facial expression through their layered structure, while Transformer encoders capture the long-range temporal dependency of the speech signals.

The multimodal Transformer fusion mechanism enables the model to combine complementary emotional evidence from the audio and visual modalities. The proposed framework achieved a 0.7 percentage-point mean accuracy improvement over the strongest hybrid-attention baseline, while larger gains were observed over the CNN–Transformer and CNN–LSTM models. However, the comparison with the hybrid-attention baseline was not statistically significant after Holm correction. The evidence therefore supports competitive performance and a modest numerical advantage rather than confirmed statistical superiority over the strongest baseline. The edge–fog–cloud component represents a deployment-oriented extension of the recognition framework. Since physical-device benchmarking and comprehensive resource profiling were not established, the study does not draw definitive conclusions regarding real-device latency, power consumption, or memory efficiency.

The observed performance results can be interpreted through the complementary roles of the framework’s components. The capsule branch contributes vector-valued visual representations that retain hierarchical relationships among learned facial patterns. The audio Transformer models dependencies between temporally separated acoustic features, while the multimodal Transformer processes the audio and visual embeddings jointly. The allocation mechanism improves deployment feasibility by distributing computational operations according to workload, resource capacity, and estimated communication delay. The framework’s contribution therefore lies in the coordinated effect of these components rather than in any individual module considered separately.

6. Potential Applications

The proposed hybrid Transformer–capsule network may support future emotion-aware applications in human–computer interaction, intelligent tutoring, healthcare decision support, driver monitoring, customer-service analytics, and social robotics. However, its suitability for real-time use requires confirmation through physical-device deployment and application-specific validation.

A specific area where this model has application is in human–computer interaction (HCI); for example, virtual assistants and conversational agents can take into account users’ emotional states and tune their responses accordingly if they are emotion-aware. It can further help in healthcare and mental health monitoring, where post-analysis of patients’ emotional states leads to a conclusion of what they might have felt during telemedicine sessions, helping them detect early stress, depressive states, or anxiety. The model is capable of recognizing emotions like fatigue, anger, or stress in the driver monitoring system, thus contributing to road safety by alerting the driver at the right time. In the same vein, education and intelligent tutoring systems leverage emotion recognition to gain insights into learner engagement and adapt their teaching strategies accordingly.

Another application is customer service analytics that enables organizations to identify how customers behave during an interaction so the firm can offer better services. Moreover, it can be utilized in smart surveillance systems and social robotics where emotionally

intelligent systems have the ability to react accordingly to human emotional states. The proposed framework may support cross-domain emotion-recognition applications, although application-specific and physical-device validation is required before practical deployment.

7. Limitations and Future Work

The proposed framework requires both audio and visual modalities during recognition. In practical settings, either modality may become unavailable or unreliable because of sensor failure, background noise, poor illumination, or privacy restrictions, which may reduce recognition performance. Future work should therefore examine adaptive missing-modality mechanisms and modality-reliability weighting to maintain performance when one input stream is degraded or absent.

The benchmark datasets were largely collected under controlled or acted conditions and provide limited representation of spontaneous, multilingual, culturally diverse, and mixed emotional expressions. Their demographic distributions are also unequal; for example, SAVEE includes only four male actors. Consequently, the reported aggregate results should not be interpreted as demonstrating demographic parity or uniform generalizability. Future studies should use larger naturalistic and demographically balanced datasets and report subgroup-specific accuracy, macro-F1, recall, calibration, confidence intervals, and error disparities.

The computational complexity of deep multimodal architectures also remains a limitation. The present study did not establish physical deployment or comprehensive resource profiling on representative edge and fog devices. The deployment-oriented task-allocation mechanism therefore requires validation under fully specified workloads, network conditions, and device configurations. Future work should investigate lightweight Transformers, capsule compression, pruning, quantization, and knowledge distillation. Emotion-recognition outputs should also be treated as probabilistic decision-support information rather than definitive assessments of an individual's internal emotional state.

The statistical comparison was based on only three independent runs; consequently, the non-significant improvement over the strongest baseline requires confirmation through a larger number of repeated experiments. Future evaluation should include additional independent seeds, physical edge-fog deployment, and controlled comparisons with recent multimodal Vision Transformer-, CLIP-, and ViT-based emotion-recognition architectures using identical data partitions and training conditions.

8. Conclusions

This study presented a latency-aware hybrid Transformer-capsule network for audio-visual emotion recognition in an edge-fog-cloud-oriented environment. The framework combines capsule-based visual representation, Transformer-based audio modeling, multimodal fusion, and deterministic latency-aware task allocation. Evaluation on RAVDESS, CREMA-D, and SAVEE achieved an average accuracy of 91.5% and an F1-score of 90.7%. Although the framework achieved the highest mean performance, its improvement over the strongest hybrid-attention baseline was not statistically significant in the limited three-run analysis. The findings therefore indicate competitive recognition performance and a modest numerical advantage, while requiring confirmation through additional independent runs, physical-device validation, and evaluation on more diverse real-world datasets.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/a19080626/s1>, Table S1, Hyperparameter Search Space and Selected Configuration; Table S2, Audio and Video Preprocessing Configuration; Table S3, Complete Architecture of the Proposed Framework; and Table S4, Baseline Architectures and Evaluation Settings.

Author Contributions: Conceptualization, P.D. and R.K.R.; methodology, A.S., D.P., A.K.A. and P.D.; software, A.S. and D.P.; validation, A.K.A., A.S. and R.K.R.; formal analysis, A.K.A. and R.K.R.; investigation, A.S., D.P., and A.K.A.; resources, R.K.R. and P.D.; data curation, A.S., D.P. and A.K.A.; writing—original draft preparation, A.S., D.P. and A.K.A.; writing—review and editing, P.D. and R.K.R.; visualization, A.S., D.P. and A.K.A.; supervision, P.D. and R.K.R.; project administration, P.D. and R.K.R.; funding acquisition, R.K.R. All authors have read and agreed to the published version of the manuscript.

Funding: The manuscript is funded by Multimedia University.

Data Availability Statement: The authors confirm that all relevant data were included in the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Hosseini, I.; Hossain, M.Z.; Zhang, Y.; Rahman, S. Deep learning model for simultaneous recognition of quantitative and qualitative emotion using visual and bio-sensing data. *Comput. Vis. Image Underst.* **2024**, *248*, 104121. [\[CrossRef\]](#)
- Qian, C.; Marques, J.A.L.; De Alexandria, A.R. Real-time emotion recognition based on facial expressions using Artificial Intelligence techniques: A review and future directions. *Multidiscip. Rev.* **2025**, *8*, 2025328. [\[CrossRef\]](#)
- Indraneel, K.; Miriyala, T.B. Electroencephalography-based emotion recognition and classification using machine learning algorithms. *Int. J. Syst. Assur. Eng. Manag.* **2026**. [\[CrossRef\]](#)
- Majkowski, A.; Kołodziej, M. Emotion Recognition from Speech in a Subject-Independent Approach. *Appl. Sci.* **2025**, *15*, 6958. [\[CrossRef\]](#)
- Sutar, M.B.; Ambhaikar, A. Formicary swarm optimized deep cnn for facial emotion recognition from human facial expressions. *Biomed. Eng. Appl. Basis Commun.* **2024**, *36*. [\[CrossRef\]](#)
- Alotaibi, A.T.; Solaiman, S. Biological Motion-Based Emotion Recognition through a deep learning approach. *IEEE Access* **2025**, *13*, 109729–109741. [\[CrossRef\]](#)
- Yadav, H.; Maini, S. The Intersection of AI and Emotions: Deep Learning Applications for a Better Society. In *International Handbook of Emotions: Resourceful Cultural Perspectives*; Springer: Berlin/Heidelberg, Germany, 2025; Volume 2, pp. 11–29. [\[CrossRef\]](#)
- Cui, L.; Jing, X.; Wang, Y.; Huan, Y.; Xu, Y.; Zhang, Q. Improved SWIN Transformer-Based semantic segmentation of postearthquake dense buildings in urban areas using remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *16*, 369–385. [\[CrossRef\]](#)
- Thai, H.-T.; Truong-Dang, V.-L.; Le, K.-H. FusionNet: A latency-aware hybrid transformer architecture for efficient leaf disease detection on the IoT edge. *Internet Things* **2026**, *38*, 101974. [\[CrossRef\]](#)
- Quispe, K.G.M.; Utyama, D.M.S.; Santos, E.M.D.; Oliveira, H.a.B.F.; Souto, E.J.P. Applying Self-Supervised Representation learning for emotion recognition using physiological signals. *Sensors* **2022**, *22*, 9102. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hasnul, M.A.; Aziz, N.A.A.; Alelyani, S.; Mohana, M.; Aziz, A.A. Electrocardiogram-Based Emotion Recognition Systems and Their Applications in Healthcare—A Review. *Sensors* **2021**, *21*, 5015. [\[CrossRef\]](#) [\[PubMed\]](#)
- Shanmin, W.; Chengguang, L.; Shengyu, C.; Qingshan, L. A survey of multimodal emotion recognition from facial expressions, audios, and language. *J. Image Graph.* **2025**, *30*, 2120–2138. [\[CrossRef\]](#)
- Pan, B.; Hirota, K.; Jia, Z.; Dai, Y. A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing* **2023**, *561*, 126866. [\[CrossRef\]](#)
- Babu, N.V.; Kanaga, E.G.M. An experimental investigation on the emotion recognition using power spectrum density and machine learning algorithms in EEG signals. In *EAI/Springer Innovations in Communication and Computing*; Springer: Cham, Switzerland, 2023; pp. 157–171. [\[CrossRef\]](#)
- Almanza-Conejo, O.; Almanza-Ojeda, D.L.; Garcia-Perez, A.; Ibarra-Manzano, M.A. Emotion recognition using electroencephalogram signals and a 1D local binary pattern for an ML-Classification-Based approach. In *Proceedings of Eighth International Congress on Information and Communication Technology*; Lecture Notes in Networks and Systems; Springer: Singapore, 2023; pp. 13–23. [\[CrossRef\]](#)
- Abdykerimova, L.; Abdikerimova, G.; Konyrkhanova, A.; Nurova, G.; Bazarova, M.; Bersugir, M.; Kaldarova, M.; Yerzhanova, A. Analysis of the emotional coloring of text using machine and deep learning methods. *Int. J. Power Electron. Drive Syst./Int. J. Electr. Comput. Eng.* **2024**, *14*, 3055. [\[CrossRef\]](#)
- Esfahani, S.H.N.; Adda, M. Classical machine learning and large models for Text-Based emotion recognition. *Procedia Comput. Sci.* **2024**, *241*, 77–84. [\[CrossRef\]](#)
- Manewa, R.M.A.H.; Mayurathan, B. *Emotion Recognition and Discrimination of Facial Expressions using Convolutional Neural Networks*; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6. [\[CrossRef\]](#)

19. Vats, R.; Nagpal, S. Emotion Recognition in Indian Faces: Performance Analysis of Deep Learning Models. In *2025 3rd World Conference on Communication & Computing (WCONF)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5. [\[CrossRef\]](#)
20. Gupta, R.; Sharma, S.J. Emotion Recognition from Facial Expressions Using Deep Learning Model. In *2024 IEEE International Conference on Intelligent Signal Processing and Effective Communication Technologies (INSPECT)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6. [\[CrossRef\]](#)
21. Waleed, G.T.; Shaker, S.H. Human emotion recognition based on facial expression using convolution neural network. *J. Adv. Inf. Technol.* **2024**, *15*, 1366–1373. [\[CrossRef\]](#)
22. Xu, C.; Gao, Y. Multi-Modal Transformer with Multi-Head Attention for Emotion Recognition. In *2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*; IEEE: Piscataway, NJ, USA, 2023; pp. 826–831. [\[CrossRef\]](#)
23. Li, X.-C.; Zhang, F.; Hua, Q.; Dong, C.-R. A deep spatiotemporal interaction network for multimodal sentimental analysis and emotion recognition. *Inf. Sci.* **2024**, *690*, 121515. [\[CrossRef\]](#)
24. Li, A.; Lv, Z.; Li, X. CAT-LCAN: A multimodal physiological signal fusion framework for emotion recognition. In *Advances in Brain Inspired Cognitive Systems*; Lecture Notes in Computer Science; Springer: Singapore, 2025; pp. 168–177. [\[CrossRef\]](#)
25. Caliwag, E.; Canilang, H.M.O.; Caliwag, A.C.; Lim, W. Implementation of emotion recognition using Edge-Cloud Computing. *J. Korean Inst. Commun. Inf. Sci.* **2022**, *47*, 474–481. [\[CrossRef\]](#)
26. Yoon, S.; Kim, B. Multi-Scale Temporal Fusion network for Real-Time multimodal emotion recognition in IoT environments. *Sensors* **2025**, *25*, 5066. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Simić, N.; Suzić, S.; Milošević, N.; Stanojev, V.; Nosek, T.; Popović, B.; Bajović, D. Enhancing Emotion Recognition through Federated Learning: A Multimodal Approach with Convolutional Neural Networks. *Appl. Sci.* **2024**, *14*, 1325. [\[CrossRef\]](#)
28. Muhammad, G.; Hossain, M.S. Emotion recognition for cognitive edge computing using deep learning. *IEEE Internet Things J.* **2021**, *8*, 16894–16901. [\[CrossRef\]](#)
29. Joseph, L.M.I.L.; Rallan, R.; Agarwal, S.; Shrivastava, S.S.; D, G.; Nayaki, A.P. Edge-EmoNet: A Low-Latency Edge-based Deep Learning Framework for Real-Time Emotion Recognition with Attention-Optimized Feature Fusion. In *2025 7th International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*; IEEE: Piscataway, NJ, USA, 2025; pp. 631–637. [\[CrossRef\]](#)
30. Moorthy, S.; Moon, Y.-K. Hybrid Multi-Attention Network for Audio–Visual emotion recognition through multimodal feature fusion. *Mathematics* **2025**, *13*, 1100. [\[CrossRef\]](#)
31. RAVDESS Emotional Speech Audio. Available online: <https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio> (accessed on 25 January 2026).
32. CREMA-D. Available online: <https://www.kaggle.com/datasets/ejlok1/cremad> (accessed on 25 January 2026).
33. Surrey Audio-Visual Expressed Emotion (SAVEE). Kaggle. Available online: <https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee> (accessed on 25 January 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.