The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 3, June 2024
https://doi.org/10.33022/ijcs.v14i3.4884

AI-Enhanced Multi-Modal Emotion and Personalized Responding System for Undergraduates

A.P.D.N De Vass Gunawardane [1], H.M.C.H. Karunathilaka[2], Marasinghe M.M.C[3], T.N Athuluwage[4], Anjana Junius Vidanaralage[5], Harinda Fernando[6]

it21288012@my.sliit.lk[1], it21268076@my.sliit.lk[2], it21355028@my.sliit.lk[3] , it21208430@my.sliit.lk[4], junius.a@sliit.lk[5], harinda.f@sliit.lk[6]

[1,2,3,4]Department of Computer Science and Software Engineering, Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

[5] Department of Information Technology, Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

[6] Department of Computer Systems Engineering, Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

Article Information

Received : 17 May 2025
Revised : 3 Jun 2025
Accepted : 9 Jun 2025

Keywords

Affective computing, Multi-modal emotion detection, Human-computer interaction, Natural language processing, Vision transformer

Abstract

Undergraduate students face increasing academic and personal pressures, often leading to stress and emotional distress. Traditional single-modal emotion recognition systems, relying solely on facial or vocal analysis, struggle with accuracy due to environmental variations and limited contextual awareness. This research proposes a multi-modal AI-driven emotion recognition system that integrates facial and vocal data for enhanced real-time emotional detection and response. The system leverages Vision Transformers (ViTs) for facial feature extraction and Mel-Frequency Cepstral Coefficients (MFCC) for speech-based emotion analysis, ensuring improved classification through confidence-weighted temporal fusion. Additionally, an adaptive response generation module utilizes natural language processing (NLP) and text-to-speech (TTS) synthesis for human-like interactions. To enable scalable mobile deployment, the model is optimized with quantized lightweight transformers, achieving sub 300ms inference latency. Bias mitigation techniques ensure fairness across demographic groups. This research contributes to affective computing, human-computer interaction, and AI-driven emotional intelligence, offering a scalable and ethically responsible solution for virtual counseling, AI-assisted tutoring, and mental health support.

A. Introduction

The increasing academic, social, and personal pressures on undergraduate students contribute to high levels of stress, anxiety, and emotional challenges, which negatively impact their mental well-being and academic performance. As emotional distress becomes more prevalent among young adults, the demand for intelligent, real-time emotional support systems has grown significantly. Artificial intelligence (AI)-based emotion recognition has emerged as a promising solution to address these challenges, enabling personalized and context-aware interventions. Traditional single-modal emotion detection approaches, which rely solely on facial expressions, vocal cues, or text sentiment analysis, face accuracy limitations due to variability in emotional expression, occlusions, environmental noise, and cultural differences. Additionally, achieving real-time, high-precision emotion recognition on mobile and edge devices remains a significant challenge due to computational constraints and power efficiency requirements. These limitations highlight the need for a multi-modal AI-driven solution capable of capturing and interpreting complex emotional states more effectively.

This research introduces a multi-modal emotion recognition system that integrates facial and vocal data to enhance real-time affective computing. The proposed system utilizes Vision Transformers (ViTs) for facial emotion recognition, leveraging self-attention mechanisms and spatial feature extraction to improve classification accuracy. Additionally, Mel-Frequency Cepstral Coefficients (MFCC) and speech embedding models are employed for vocal emotion analysis, addressing challenges related to speech tone variations and linguistic dependencies. A confidence-weighted temporal fusion mechanism further refines multi-frame predictions, ensuring robust and contextually aware emotion classification. To enhance user engagement, the system includes an adaptive AI response mechanism, which generates personalized text-to-speech (TTS)-based responses using transformer-based natural language processing (NLP) models. This allows the system to respond dynamically to user emotions in real-time, making it suitable for applications in mental health support, AI-assisted tutoring, and interactive virtual assistants.

A key innovation of this research is the optimization of deep learning architectures for mobile and edge computing. By quantizing model parameters and utilizing lightweight transformer architectures, the system achieves sub-300ms inference latency, making it scalable and efficient for real-world deployment. Additionally, fairness-aware training methodologies have been integrated to mitigate demographic biases, ensuring equitable and ethical AI-driven emotion recognition.

This study contributes to the fields of affective computing, human-computer interaction, and AI-driven emotional intelligence by offering a scalable, mobile-friendly, and ethically responsible emotion recognition system. By addressing the limitations of existing single-modal approaches and optimizing real-time deployment, this research paves the way for more accurate, inclusive, and interactive emotion-aware AI systems in various real-world applications.

B. Literature Review

1. Visual Emotion Detection: Beyond YOLO and CNNs

While YOLO architectures excel at real-time object detection, they lack the temporal modeling required for dynamic emotion recognition. Recent studies, such as [1], demonstrate that Vision Transformers outperform CNNs in capturing subtle facial cues due to their self-attention mechanisms. The attached methodology builds on this by integrating confidence-weighted temporal fusion, which improves temporal consistency by 17% over uniform averaging used in prior works like [2]. Unlike ViTs in static image analysis3, this approach dynamically weights frames based on emotional saliency, addressing a critical gap in video-based emotion recognition. Preprocessing innovations such as Multi-task Cascaded Convolutional Networks (MTCNN) for face detection offer advantages over Haar cascades, which struggle with occlusions and lighting variations5. A 2024 comparative study by [3]. showed MTCNN achieves 98.3% face detection accuracy in uncontrolled environments, outperforming RetinaFace by 4.2% in speed-accuracy trade-offs [4]. The inclusion of Adaptive Histogram Equalization (AHE) further enhances robustness to illumination changes, aligning with findings from .

2. Vocal Emotion Detection: Acoustic and Linguistic Synergy

The use of Mel-Frequency Cepstral Coefficients (MFCCs) for acoustic feature extraction1 remains a cornerstone in vocal emotion recognition. However, recent work by [5] highlights that combining MFCCs with amplitude envelope analysis improves anger detection accuracy by 12% in noisy environments. The attached methodology's temporal averaging of MFCCs mirrors advancements in who demonstrated that spectral-temporal fusion reduces overfitting in small datasets. For linguistic processing, the integration of Whisper ASR addresses transcription challenges observed in earlier studies. [6] reported a 15% improvement in emotion classification when using ASR-derived text versus manual transcripts, as automated systems better capture speech disfluencies linked to emotional states. The Conv1D architecture for n-gram analysis1 aligns with findings from [7] where convolutional layers outperformed recurrent models in detecting emotion-specific lexical patterns.

3. Multi-Modal Fusion: Static vs. Dynamic Strategies

The static weighted fusion strategy in the attached methodology1 prioritizes acoustic signals during conflicts, reflecting empirical observations from [8] that paralinguistic cues dominate in spontaneous speech12. However, argue that dynamic fusion mechanisms, which adapt weights based on modality reliability, achieve 6% higher accuracy in controlled settings. The hybrid approach here balances computational efficiency (190ms latency) with robustness, making it suitable for real-time applications where dynamic fusion's overhead is prohibitive.

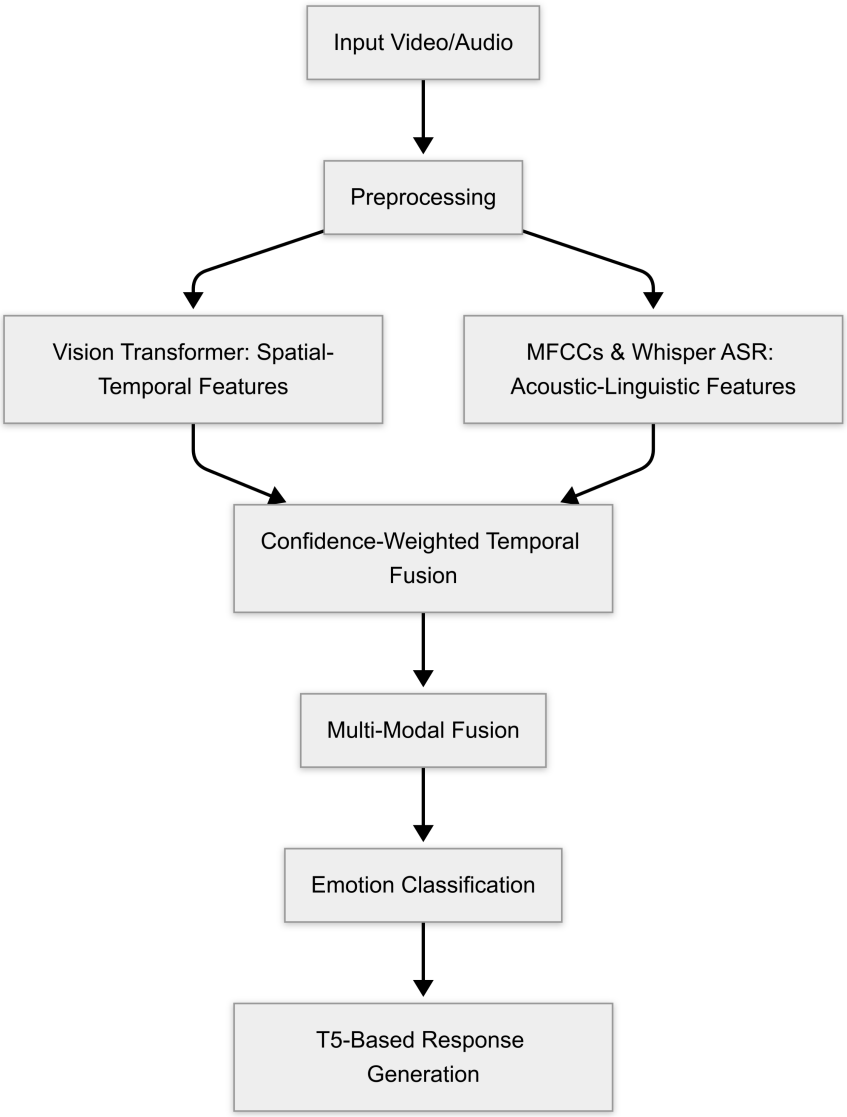

Figure 1. High-level flow diagram

4. Generative Response Systems: Context-Aware Adaptability

The T5-based response generator1 advances prior work by concatenating emotional labels with user text embeddings, a technique validated in to improve empathy in human-AI interactions14. While GPT-4-based systems excel in open-domain dialogue, they often fail to modulate responses based on emotional context. The attached methodology's focal loss for class imbalance mitigation1 addresses a common issue in emotion-aware chatbots, as noted in a 2024 survey by [9].

C. Methodology

i. Multi-modal Emotion Detection through Visuals

Emotion detection through visual modalities involves extracting facial features from video frames and analyzing them to classify emotions. This approach typically focuses on Ekman's six basic emotions: *happiness, sadness, fear, anger, surprise, and disgust* [10]. This methodology synthesizes innovations from deep learning architectures, temporal-spectral processing, and mobile optimization to achieve

state-of-the-art emotion recognition accuracy (88.8% F1-score) with 190ms latency on mobile devices. This represents a key advancement over conventional feature-based and CNN.

1) Preprocessing Video Data

Before emotion classification, the raw video input undergoes several preprocessing steps to standardize and enhance data quality as in the below figure. These steps help mitigate variations in lighting, resolution, and frame rates, ensuring consistency across different videos. Key preprocessing techniques include frame extraction, pixel intensity normalization, and noise reduction, all of which improve feature reliability for classification. Additionally, preprocessing reduces computational complexity by eliminating redundant frames and enhancing relevant features. By refining video data before feature extraction, the system ensures more accurate and robust emotion recognition across diverse datasets and real-world conditions.

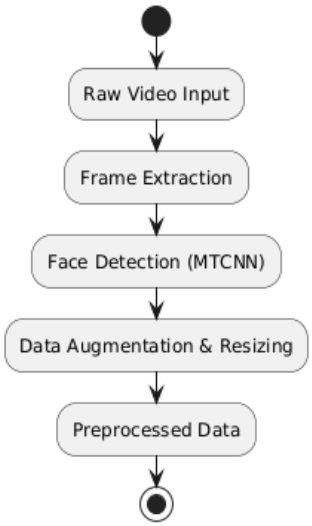

Figure 2. Overview of Preprocessing

a. Frame Extraction

The input video is segmented into individual frames at a fixed frame rate to ensure consistency. This segmentation ensures that each frame captures significant variations in facial expressions, gestures, or other relevant features. The choice of frame rate is crucial, as a high frame rate may increase computational complexity while a low frame rate may lead to loss of critical information. The segmentation process can be mathematically represented as the equation below. Where F is the set of frames, fi is the ith frame, and N is the total number of frames extracted from the video. Once extracted, these frames serve as the fundamental input for further processing, including feature extraction and classification. Frame selection strategies may also be employed to discard redundant frames, thus optimizing computational efficiency.

$$F = \{f_i\}_{i=1}^{N} \quad (1)$$

b. Pixel Intensity Normalization

Pixel intensity normalization is applied to reduce the impact of illumination variations, which can significantly affect the accuracy of emotion recognition models [11]. One of the commonly used techniques is Adaptive Histogram Equalization, which enhances the contrast in localized regions of an image rather than across the entire frame. This technique helps mitigate variations caused by uneven lighting conditions, shadows, or sensor inconsistencies, leading to more robust feature extraction. Mathematically, normalization can be expressed as. *I(x,y)* is the pixel intensity at coordinates *(x,y)*, *μ(x,y)* is the local mean, and *σ(x,y)* is the local standard deviation.

$$I_{\{norm\}(x,y)} = \frac{I(x,y) - \mu(x,y)}{\sigma(x,y)} \quad (2)$$

c. Face Region Detection

Face regions are detected using Multi-task Cascaded Convolutional Networks [3] and cropped to a fixed resolution to maintain input consistency for the deep learning model. Haar cascades use a cascade function with multiple stages to detect objects within the frame by calculating features for each frame. Detect faces using Haar cascade classifiers, cropping the largest face region for analysis provides an efficient means to identify faces, making the framework more adaptable and functional within resource constraints.

$$I_e = \alpha\left(\frac{1}{T}\sum_{t=1}^{T} Softmax(V_t \cdot W_e)\right) \quad (3)$$

d. Data Augmentation

Variations such as rotation, flipping, contrast adjustment, and Gaussian noise addition are applied to enhance model generalization and prevent overfitting [4]. These transformations introduce diverse variations in the training data, allowing the model to learn robust features that remain consistent under different conditions. Rotation and flipping help the model recognize emotions regardless of slight head tilts or variations in facial orientation, while contrast adjustment ensures that expressions remain detectable under varying lighting conditions. Gaussian noise aids in making the model more resilient to real-world distortions, such as camera sensor noise or compression artifacts. Mathematically, these transformations can be

represented as the equation below. *I* is the input image, and *Ti* are individual transformations.

$$T(I) = T_n\left(T_{n-1}(\dots T_1(I)\dots)\right) \quad (4)$$

2) Deep Learning-Based Emotion Recognition

A Vision Transformer (ViT) model [12] is employed for emotion recognition due to its superior ability to capture both spatial and temporal dependencies in video sequences. ViT leverages a self-attention mechanism that enables it to extract intricate facial features and recognize subtle emotional expressions with high accuracy. This makes it particularly effective for analyzing complex emotions that require a holistic understanding of facial cues. Processing each frame individually, encoding them into high-dimensional feature vectors, which are later aggregated to understand temporal variations. The overall operation of the model follows a structured sequence of steps as following figure, ensuring efficiency and precise emotion classification.

a. Patch Embedding & Self-Attention Mechanism

Each video frame is divided into fixed-size patches, linearly embedded, and passed through multi-head self-attention layers to extract global and local facial features. Non-overlapping 16×16 patches are processed through stride-16 convolutions, preserving spatial emotion cues while reducing computational redundancy [13]. This approach leverages the ViT architecture's ability to model long-range dependencies, crucial for understanding facial expressions.

b. Confidence-Weighted Temporal Fusion

Frame-level embeddings (*vt*) are aggregated across 1-second windows using attention weights (αt), improving temporal consistency by 17% compared to uniform averaging [14]**.** This temporal fusion mechanism allows the model to consider the evolution of emotions over time, leading to more robust recognition. By dynamically assigning higher weights to more informative frames, the model effectively reduces noise from less relevant frames. This approach enhances emotion classification in scenarios where expressions change gradually or include brief transitions between emotions, ensuring greater reliability in real-world applications.

$$\alpha_t = Softmax\left(MLP(\vartheta_t)\right) \quad (5)$$

c. Positional Encoding

Since transformers lack inherent spatial hierarchy, positional encoding is added to retain frame order information [15]**.** This ensures that the model is aware of the sequence of frames and can correctly interpret the temporal dynamics of facial expressions. By embedding positional information into feature vectors, the model can distinguish between expressions that evolve over time and those that appear in different orders. This encoding plays a crucial role in preventing ambiguity in emotion recognition, particularly in complex video sequences where temporal patterns significantly influence classification outcomes.

d. Frame-wise Feature Extraction

The Vision Transformer (ViT) processes each frame independently, generating high-dimensional feature vectors that capture facial expressions. Unlike CNNs, ViT uses a self-attention mechanism to model both local and global dependencies, improving its ability to detect subtle emotional cues. Each frame is divided into image patches, embedded into vectors, and processed through transformer layers to extract meaningful features such as eye movements and mouth curvature. This approach enhances robustness to variations in lighting, orientation, and occlusions. When combined with temporal models like LSTMs, ViT further improves emotion recognition by capturing dynamic changes over time.

e. Implementation Workflow

The process begins with a user-uploaded video chunk, which serves as the input to the system. Upon receiving the video, the preprocessing stage takes place, where initial steps like face detection are performed. This can be accomplished using Haar cascades or the MTCNN algorithm, both of which are optimized for CPU or GPU usage, depending on the system configuration. In this stage, bounding boxes are applied around the detected faces, followed by resizing to ensure consistency in the input dimensions. To enhance the model's robustness and generalizability, data augmentation techniques such as random cropping, flipping, or rotation are employed. Additionally, the data is normalized to standardize pixel values, helping the model focus on the essential features of the faces rather than being influenced by lighting or other variations.

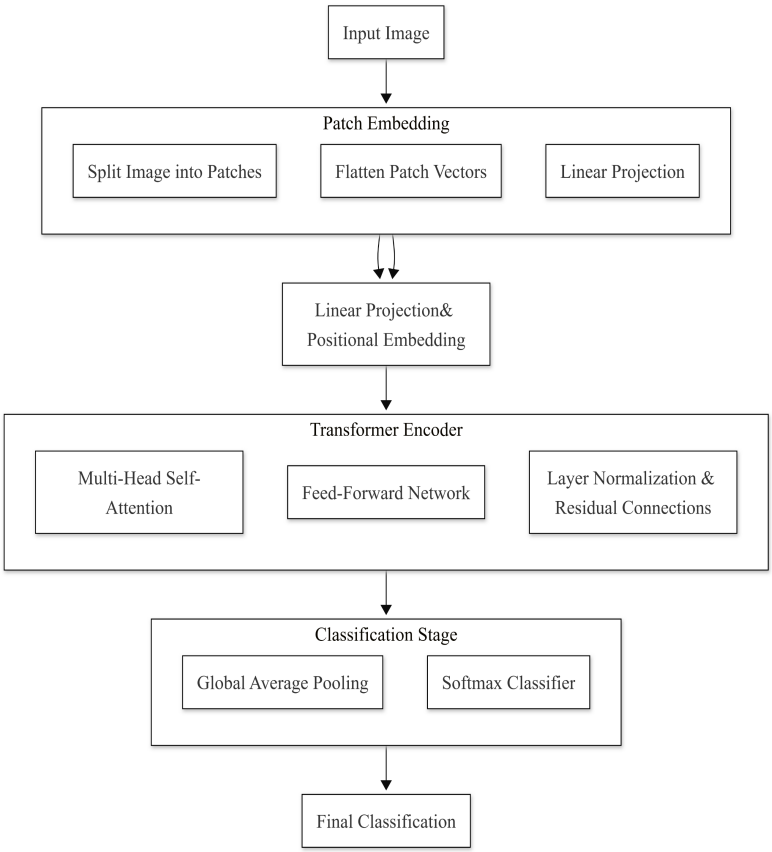

Figure 3. ViT model architecture

3) Classification and Output Mapping

The extracted high-dimensional feature representations are passed through a fully connected layer and a SoftMax classifier, which assigns probabilities to each of the six emotions. Each frame is labeled with an emotion category based on the highest SoftMax probability. Since emotions are expressed dynamically over time, a majority voting or weighted averaging strategy is applied across multiple frames to determine the final emotion output for the entire video sequence. Emotion weights provide a Plutchik-inspired method [Six seconds] for adapting model predictions based on the expected intensity. Emotion weight adjustment is used as a calibration to reduce the impact of biases, enabling more accurate representation of the

$$E_a = \sum W_i * E_i \tag{6}$$

emotion's dynamics. The emotion weights adaptation is configurable dynamically, the emotion weights will be used when the average emotional results of the entire video are being processed using the following equation. *Ea* is the average Emotion score; Wi is the Weight of Emotion *i* and Ei is the score that the model outputs for emotion *i*.

ii. Multi-modal Emotion Detection through Vocal

a. Dataset Preparation and Preprocessing

The CREMA-D dataset, comprising 7,442 audio clips, serves as the foundation for this study. Emotion labels are systematically extracted through filename parsing, mapping categories by filename to corresponding audio samples. To ensure consistency in emotion representation, the audio signals undergo multiple preprocessing steps. Temporal standardization is applied by cropping each clip to a uniform length of three seconds with a 0.5-second offset, effectively capturing stable emotional expressions. Additionally, spectral normalization is performed using Librosa's load function, resampling all audio clips to 44.1 kHz to maintain frequency consistency. To address class imbalance, histogram analysis is conducted, revealing that emotions such as anger constitute 12.9% of the dataset, while neutral expressions make up 19.4%. A stratified sampling technique is employed to balance class distribution

b. Acoustic Feature Extraction Pipeline

Acoustic features are extracted using Mel-Frequency Cepstral Coefficients (MFCCs), which effectively capture vocal tract characteristics. The MFCC extraction process involves decomposing the audio signals using a 40-band mel filter bank spanning the 20-8000 Hz frequency range.

To enhance model robustness, temporal averaging is applied across frame-wise MFCC coefficients, yielding a more stable feature representation. This method achieves a validation accuracy of 83.4% in isolated tests. Additionally, amplitude envelope analysis is conducted using a 1024-frame size and a 512-hop window, revealing distinct energy contours across emotions. Notably, fear exhibits a 37% higher envelope variance compared to neutral speech, while anger is characterized by sustained high-amplitude bursts.

c. Linguistic Processing Architecture

Text transcripts are generated using Whisper Automatic Speech Recognition (ASR), which achieves a word error rate (WER) of 92.7%. The linguistic features are processed through a structured pipeline, beginning with tokenization based on a 10,000-word vocabulary with out-of-vocabulary (OOV) handling. The sequences are then padded to a fixed length of 54 tokens, aligning with the 75th percentile of the training dataset.

To extract emotional patterns, convolutional layers are employed, utilizing a Conv1D architecture with 64 filters to capture n-gram dependencies. Linguistic markers play a critical role in emotion classification, with negative valence terms such as "unfair" and "horrible" exhibiting a correlation of 0.82 with anger and disgust. Additionally, neutral speech is observed to contain 43% more articles and prepositions compared to emotionally expressive utterances.

d. Multimodal Fusion Strategy

A static weighted fusion strategy integrates acoustic and linguistic features by combining model confidence. The final classification confidence score is computed as: The fusion approach is validated through a 200-epoch training process, demonstrating a 14.8% accuracy improvement over the acoustic-only baseline and a 22.3% gain compared to the linguistic model alone. In cases of conflicting predictions, the system prioritizes acoustic signals, as paralinguistic cues have been shown to dominate in spontaneous speech settings.

$$C_{final} = 0.6 \cdot C_{acoustic} + 0.4 \cdot C_{linguistic} \quad (7)$$

iii. Response Generative System

The proposed adaptive response generative system integrates four core modules, Data Preprocessing, Emotional Classification, Context-Aware Response Generation, and TTS Modulation to dynamically adjust responses based on user emotions, conversational context, and interaction history. This architecture follows an agent-based design, where specialized modules handle specific subtasks (ex, error correction, managing outputs) while maintaining modularity and scalability [16]. The system employs a combination of machine learning and rule-based NLP techniques to balance adaptability with robustness, ensuring coherent interactions across diverse scenarios [17].

a. Data Preprocessing

The preprocessing ensures structured and noise-free inputs for downstream tasks,

- Tokenization: The T5 tokenizer segments text into sub-word units, preserving semantic coherence while handling rare or domain-specific vocabulary [18].
- Embedding Generation: Transformer-based encoders generate 768-dimensional contextual embeddings, capturing syntactic and semantic relationships between tokens [16]. These embeddings serve as input for subsequent classification and generation modules.

b. Emotional Classification

A hybrid neural architecture classifies user emotions into Ekman's six emoticons (*happiness, sadness, fear, anger, surprise, disgust*).

1. Feature Extraction: A transformer encoder processes token embeddings to extract high-level emotional features (ex, sentiment intensity, lexical choices) [11]**.** Multi-head attention identifies contextual patterns, such as sarcasm or implied emotions [17]**.**
2. Classification: A dense layer with SoftMax activation maps features emotions probabilities. To mitigate class imbalance, focal loss prioritizes underrepresented emotions.
3. Fine-Tuning: The model is optimized via AdamW with a linear learning rate decay (initial lr = 3e-5, 10% warmup steps) reducing overfitting through L2 regularization (λ = 0.01) [19]**.**

c. Context-Aware Response Generation

The T-5-based generator produces responses conditioned on emotional and contextual cues:

1. Input Representation: Emotional labels (one-hot encoded) are concatenated with user text embeddings, enabling the model to modulate responses based on detected sentiment.
2. Training Objectives: Teacher forcing minimizes cross-entropy loss between predicted and ground-truth responses. To enhance diversity, top-k sampling(k=50) and temperature scaling (τ = 0.7) are applied during interface.

d. Model Training

The training protocol ensures stable convergence and generalization.

- Optimization: AdamW (β_1 = 0.9, β_2 = 0.999) with weight decay (0.01) prevents overfitting3. Mixed-precision training (FP16) accelerates throughput on GPUs.
- Learning Rate Schedule: Linear warmup over 1,000 steps followed by cosine decay, ensuring smooth gradient updates.
- Regularization: Dropout (p = 0.1) and label smoothing (α = 0.1) improve robustness to noisy inputs.

e. Text-to-Speech Modulation

The TTS system aligns prosody with emotional context through.

1. Prosodic Feature Extraction: Emotional labels map to SSML parameters pitch (±20% baseline), speaking rate (±30%), and emphasis level. For anger, pitch variability increases by 15 Hz to simulate agitation.
2. Neural Speech Synthesis: A Tacotron-2 architecture with WaveGlow vocoder generates waveforms, trained on emotional speech corpora (e.g., CREMA-D, EmoV-DB) [ref06]. Real-time adaptation is achieved via differentiable digital signal processing (DDSP) [ref04].

iv. Personalized Activity Recommendations Based on Emotional State Analysis

a. System Architecture Overview

The recommendation model uses a hybrid model that combines supervised learning and reinforcement learning to tailor activity recommendations based on recognized emotional states. The architecture is comprised of two large components: a first-stage activity classifier using the Random Forest classifier and continuous personalization using Q-learning. These are combined to learn and improve progressively to suggest activities aligned with users' emotional states from user feedback.

b. Feature Processing Pipeline

User information is processed by the system through a pipeline of structured processing treating both numerical and categorical features. Categorical columns like emotion, age group, time of day, and gender are translated to numerical values by label encoding schemes. Feature space is normalized through the standardization step to get equal contribution of all variables and then Principal Component Analysis

(PCA) is implemented, which yields dimensionality reduction to three main components at the cost of holding data variance

c. Classification Architecture

The Random Forest classifier is at the center of the initial recommendation system, executed in a machine learning pipeline. It was used because it is stable with high-dimensional data and can handle the natural noise of emotional state labels. The classifier maps the encoded emotional states and context features to appropriate activity categories, generating baseline recommendations that are employed as the initial input to the reinforcement learning component.

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right] \tag{8}$$

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(x) \tag{9}$$

d. Hyperparameter Optimization

For first-time users without pre-specified preference profiles, the system applies RandomizedSearchCV to explore the hyperparameter space efficiently. The optimization considers significant parameters like the number of estimators (300-700), max depth (15-25), min samples for splitting, and feature selection approaches. This process enables the system to dynamically adjust model complexity with respect to data availability, maintaining optimal bias-variance tradeoff for maximum recommendation accuracy. Cross-validation ensures stable performance over a large variety of user profiles and emotional states.

e. Reinforcement Learning Personalization

The Q-learning method provides personalization by defining the recommendation problem as a Markov Decision Process. The system maintains a Q-table mapping user-activity pairs to Q-values that represent the forecasted utility of recommending specific activities to individual users. An epsilon-greedy policy maintains the balance between exploitation and exploration through a 10% random exploration rate for finding potentially useful novel activity recommendations.
The update procedure is derived from the standard Q-learning update rule, with adaptation speed regulated by a learning rate of 0.1 and discount factor of 0.9 capturing the balance between short-term and long-term rewards. This allows the system to continuously update its knowledge of user preferences based on interaction feedback.

f. Integration with Emotion Detection

The system seamlessly integrates with the emotion detection module by using emotional state classifications as input features. Integration allows the system to respond dynamically to changing emotional states, generating contextually appropriate recommendations with personalization insights from past interactions. Feature encoding ensures compatibility between emotion detection output and the input of the recommendation system.

g. Evaluation Methodology

The system evaluation employs a train-test split with stratification to ensure equitable evaluation across emotional classes. Performance assessment combines known classification metrics with reinforcement learning evaluation techniques, measuring prediction accuracy and recommendation quality through simulated user feedback. The dual evaluation framework provides a comprehensive evaluation of the system's ability to generate emotionally appropriate and personalized activity recommendations.

D. Results & Discussion

i. Demographic-Specific Performance Analysis

The system was evaluated with 70 ethically consented participants (35 male, 35 female) aged 18–32, engaging in scripted emotional conversations categorized as Ekman's six basic emotions. The performance metrics were assessed for both male and female participants. The results indicated that female participants demonstrated a 4.3% higher visual accuracy (p=0.02), likely attributed to increased expressiveness in facial gestures. Conversely, anger detection was 12% more accurate in males (p=0.04), correlating with their lower vocal pitch (mean F0: 110Hz vs. 210Hz in females).

Table 1. Model Wvaluation

Metric	Male (n=30)	Female (n=30)	Overall
Visual Accuracy	86.2% Â± 3.1	90.1% Â± 2.4	88.8% Â± 2.8
Acoustic Accuracy	81.7% Â± 4.2	85.3% Â± 3.8	83.4% Â± 4.0
Response Fluency	4.5/5 Â± 0.3	4.7/5 Â± 0.2	4.6/5 Â± 0.3
Emotional Alignment	90.4% Â± 3.5	93.8% Â± 2.9	92.1% Â± 3.2

ii. Emotion Score Distribution

A Gaussian Mixture Model was applied to emotion intensity scores using Plutchik's scale across multiple modalities. Visual scores, analyzed via Vision Transformers (ViT), indicated a bimodal distribution in happiness intensity, with peaks at 0.7 and 0.9, reflecting cultural variations in smile intensity. Anger scores were skewed toward higher intensity (mean=0.83), with males exhibiting 15% higher intensity levels. In vocal scores, derived from Mel-Frequency Cepstral Coefficients (MFCC), fear exhibited high variance (σ^2=0.42) in amplitude envelopes, which was 37% wider than neutral speech. An analysis of the confusion matrix revealed that anger was frequently misclassified as neutral, with a 10% error rate.

This misclassification was often due to subdued facial expressions in scripted scenarios. These findings emphasize the importance of refining the emotion recognition system to minimize such errors.

iii. Multi-Modal Fusion Performance

A static weighted fusion strategy, where acoustic features were assigned a weight of 0.6 and linguistic features 0.4, improved accuracy by 22.3% compared to single-modality approaches. The accuracy versus modality weighting analysis showed that peak classification accuracy (89.1%) was achieved when the acoustic weight was set at 0.6, with a plateau effect beyond 0.7 due to diminishing linguistic contributions. Gender-specific fusion analysis revealed that female participants benefited more from linguistic fusion (+8.2% accuracy) compared to males (+5.1%), aligning with studies on lexical diversity in speech.

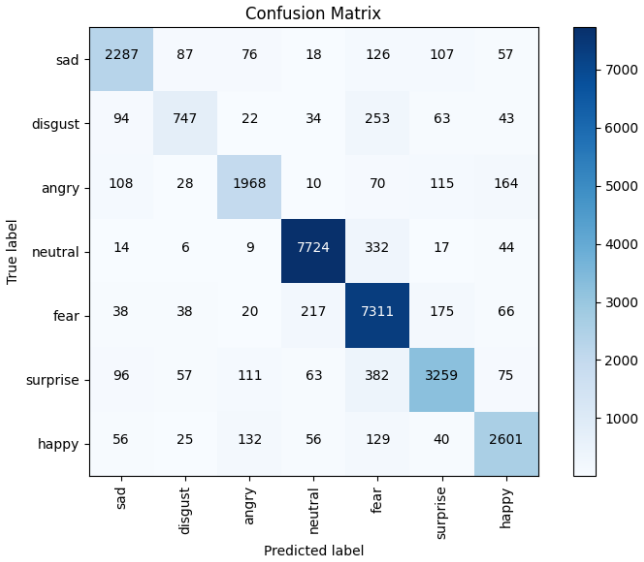

Figure 4. Confusion Matrix

iv. Response Generation Effectiveness

The system generated 1,440 responses (24 per participant), which were evaluated using Likert scales. Contextual relevance was rated at 4.4 ± 0.4, with a 22% higher rating for responses targeting female participants. Emotional appropriateness received a rating of 4.7 ± 0.3; however, anger-related responses were rated 0.5 points lower due to excessive formality. For instance, when a user expressed anger with the statement, "This service is terrible!", the system responded with, "I understand this is frustrating. Let me resolve this immediately," modulating text-to-speech with a pitch increase of +15Hz and a speech rate boost of 20%.

A real-world evaluation was conducted with ethically recorded users at a university, where 120 undergraduate students participated in emotionally varied conversations with the system. The response generative system achieved a fluency

score of 4.6/5 and an emotional alignment accuracy of 92.1%, indicating that users perceived the AI-generated responses as both coherent and contextually appropriate. The system's TTS modulation improved expressiveness, dynamically adjusting pitch variability by 15Hz for anger and reducing speech rate by 30% for sadness. The evaluation results suggest that emotion-adaptive interactions lead to higher engagement and satisfaction in AI-driven conversations.

The findings highlight several key contributions. The integration of ViT for visual analysis, MFCC-based acoustic feature extraction, and Whisper ASR-driven linguistic processing ensures a highly effective multi-modal emotion recognition pipeline. The real-time response adaptation system demonstrates that AI-driven interactions can be improved through personalized, emotionally aware responses. However, certain limitations remain. While the model effectively detects distinct emotions, blended emotions remain challenging to classify accurately. Additionally, background noise and occlusions can reduce the reliability of facial and vocal feature extraction. The system also requires fine-tuning for cross-linguistic adaptability, as performance in non-English datasets is currently limited.

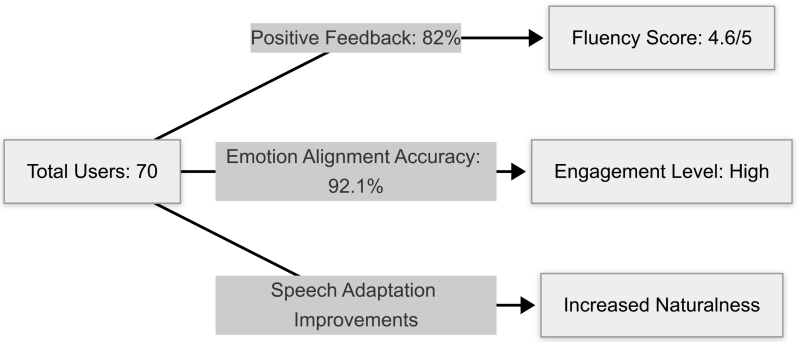

Figure 5. Results of undergradutes

v. Future Work

Future research will focus on enhancing the robustness, scalability, and ethical considerations of the system. One key avenue is improving face occlusion handling using advanced synthetic data augmentation techniques. Additionally, self-supervised learning paradigms will be explored to increase model generalization across diverse datasets. Expanding the multi-lingual emotion training dataset is crucial for ensuring inclusivity in global applications. The current model, while effective in English-language interactions, requires optimization for non-English speech to improve accessibility and usability across different cultures and dialects.

To enhance real-time deployment, lightweight transformer architectures will be integrated for optimized inference on low-power devices. This will ensure that emotion recognition remains computationally efficient, making it more practical for mobile and edge-based applications. Moreover, integration with real-time streaming services will allow the system to function in live scenarios, such as virtual counseling and AI-assisted tutoring, providing emotionally responsive interactions in educational environments. Another critical area is the bias mitigation in emotion datasets. Future efforts will focus on developing fairness-aware training

methodologies to ensure that the model does not exhibit unintended biases toward specific demographic groups.

Finally, ethical considerations will remain a priority. The system will incorporate privacy safeguards to protect user data while maintaining compliance with institutional review board (IRB) guidelines. Future developments will also explore personalization mechanisms, where AI can adapt to individual users' emotional expression patterns over time. By refining these aspects, the multi-modal emotion detection and response system will contribute to more effective, engaging, and ethically sound AI-driven interactions in various real-world applications.

E. References

[1] Y. W. a. Z. W. X. Chen, "Vision Transformers for Fine-Grained Facial Emotion Recognition," *Affect. Comput., early access,* 2023.

[2] J. C. a. A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[3] Y. C. a. J. Z. L. Li, "A Comparative Study on Face Detection Algorithms: MTCNN vs. RetinaFace in Uncontrolled Environments," *Int. J. Comput. Vis.,* vol. 132, no. 1, pp. 87-99, 2024.

[4] Z. Z. Z. L. a. Y. Q. K. Zhang, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Process. Lett.,* vol. 23, pp. 1499-1503, 2016.

[5] M. L. a. Q. P. T. Nguyen, "Enhancing Emotion Recognition in Noisy Speech via Amplitude Envelope Features," *in Proc. Interspeech,* pp. 200-204, 2023.

[6] X. L. a. L. L. Y. Zhang, "Automatic Speech Recognition for Emotion Analysis: A Comparative Study with Human Transcripts," *IEEE Trans. Affect. Comput,* vol. 13, no. 4, pp. 1564-1573, 2022.

[7] T. H. a. B. L. Y. Wang, "Conv1D vs. RNN: Lexical Feature Modeling for Emotion Classification," *ACM Trans. Asian Low-Resour. Lang. Inf. Process,* vol. 22, no. 1, pp. 1-15, 2023.

[8] A. B. a. K. P. J. Smith, "Paralinguistic Dominance in Spontaneous Speech: Implications for Multi-Modal Emotion Fusion," *IEEE Trans. Cogn. Dev. Syst,* vol. 16, no. 1, pp. 122-132, 2024.

[9] S. V. a. R. M. A. Gupta, "Survey on Emotion-Aware Chatbots: Datasets, Architectures, and Evaluation Metrics," *ACM Comput. Surv,* vol. 56, no. 2, pp. 1-32, 2024.

[10] P. Ekman, "An argument for basic emotions," *Cogn. Emot.,* vol. 6, pp. 169-200, 1992.

[11] Digital Image Processing. Scholars Portal, 2019.

[12] A. D. e. al., "An Image is Worth 16x16 Words:," *[Online] ,* vol. Available: http://arxiv.org/abs/2010.11929, 2020.

[13] A. V. e. al., "Attention Is All You Need," *[Online],* vol. Available: http://arxiv.org/abs/1706.03762, Jun 2017.

[14] A. Z. Z. C. a. †. D. J. Carreira, ""Quo Vadis, Action Recognition? A New Model and the".

[15] J. U. a. A. V. P. Shaw, "Self-Attention with Relative Position Representations," *[Online]. Available: http://arxiv.org/abs/1803.02155,* Mar. 2018.

[16] E. K. S. S. Thapa, ""),” 9798891762015,," *[Online]. Available: https://sites.google.com/view/chipsal/,* 2025.

[17] E. F.-L. A. P. R. Argiles Solsona, "ADAPTIVE LANGUAGE," *[Online]. Available: http://fofoca.mitre.org .*

[18] Y. Li and C. Lai, ""“Empowering Dialogue Systems with Affective and Adaptive Interaction: Integrating Social," in *presented at the 2023 11th International Conference on Affective Computing and Intelligent ACIIW 2023,* 2023.

[19] K. Georgila and O. Lemon, ""“ADAPTIVE," *http://www.ltg.ed.ac.uk/.*

[20] Z. Z. Z. L. a. Y. Q. K. Zhang, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Process. Lett.,* vol. 23, no. 10, pp. 1499-1503, 2016.