Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00246-0

Original Article

Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture

Atharv Shukla [1], Rashi Agarwal [2]

1. *Computer Science, Axis Institute of Technology and Management, Kanpur, IND*
2. *Computer Science and Engineering, Harcourt Butler Technical University, Kanpur, IND*

Received: July 22, 2026 | Review began: July 26, 2026 | Review ended: August 04, 2026 | Published: August 25, 2026

Abstract

Background

The existing emotion detection systems are either audio or video centric. While such approaches have proven effective in controlled environments like a lab or a studio, these systems fail to perform under real-world conditions such as low light, audio interference, mispronunciation, and other environmental factors.

Objective

The objective of our research is to design a quad-modal emotion detection system that combines four sub-models (static facial images, dynamic facial expressions, speech tone, and natural language processing [NLP] of the spoken words) to design a robust model that can perform under real-world conditions.

Methods

We propose a dynamic weighting algorithm that biases the model towards a certain modality if the conditions are favorable (e.g. switch to audio-only if the video is obstructed) using cross-modal multi-head attention. We optimise the system to run on a CPU (due to the lack of a CUDA-enabled NVIDIA GPU) by using an asynchronous state buffer for the heavy NLP pipeline to maintain a high enough video frame rate.

Results

Our model achieved an accuracy of 61.21% (95% confidence interval [CI]: 58.77%-63.65%) with a mean area under the curve of 0.90 (95% CI: 0.88-0.92) across six emotion classes on real-world data (RAVDESS, SAMM, CREMA-D) during training and testing while maintaining performance under visual and audio deprivation conditions.

Conclusion

Our architecture, while forgoing the precision of a studio-lit, well-audio-recorded environment, is able to sacrifice some accuracy for the robustness required to function in a real-world human-computer interaction environment.

Categories: Affective Computing, Natural Language Processing (NLP), Computer Vision

Keywords: emotion detection, quad-modal ai, deep learning, cross-modal attention, natural language processing, feature fusion, cpu optimization, ai ethics

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

Introduction

Teaching an artificial intelligence (AI) system to understand human emotion is a huge part of the modern technology era. As we know, AI is integrated into hospitals, offices, businesses, and many workplaces, so it is necessary for AI to understand how humans feel. So for this purpose, there are many traditional AI systems built that are baseline models. They only work on audio (speech recording) or video recognition. But these systems are not enough to recognise and understand the emotions of a human being. Also, they do not work well in real-world conditions.

These unimodal (baseline single-input) works perform well in labs and studios, but they consistently fail in real-world chaotic conditions where dynamic lighting and occlusion destroy static feature extraction [1]. They might analyse a picture of a face or recognise it through face or speech, but they rarely do well.

A line of recent studies reported in 2025 and 2026 focuses on emotion recognition systems that process visual, speech, and textual information. The latest approaches reported in these works highlight the effectiveness of vision-language, multimodal speech-language models, and hybrid networks based on attention mechanisms for building robust and accurate emotion perception systems [2-7].

To fix this limitation, we built a quad-modal (multi-modal) system with a dynamic fusion architecture. Instead of processing just one resource only, we enable it to process four different streams:

• **Dynamic Video:** We track moving facial muscles using 2D convolutional neural networks (CNNs) on stacked optical flow frames.

• **Audio Tone:** We analyse voice pitch and rhythm using 1D CNNs on mel-spectrograms.

• **Textual Sentiment:** We understand the actual words being said using Whisper and DistilRoBERTa NLP models.

• **Static Vision:** We use a pre-trained Facial Expression Recognition (FER) model as a fallback.

Our main manager of this system is the dynamic weighted algorithm, which continuously focuses on which stream gives the best result of recognising emotion in the four sub-models.

Comprehensive literature review

Researchers have been trying to solve emotion recognition for a long time, from massive math to deep learning models.

The Era of Handcrafted Features (2000-2012)

Before deep learning, researchers mostly used basic machine learning to solve emotion recognition, like support vector machines and logical patterns, which can work well under conditions of perfect lighting and when the person looks straight at the camera, but if a face is slightly turned or moving towards a shadow, the system rules fall apart [8].

The Deep Learning Visual Revolution (2012-2018)

When deep learning emerged, the 2D CNNs broke the rules and accuracy records of datasets like FER2013. The CNN model recognises faces by their own and starts recognising the emotions. But there is a limitation: at this time, the model only recognises static images and struggles to capture subtle temporal movements, such as a smile; it fails on them [9].

The Shift to Bimodal Audio-Visual Systems (2018-2022)

To fix this problem, researchers developed a 'bimodal' system that has a fusion of both audio and video, where both of these modalities work together, and attempted to concatenate audio and visual data into a unified feature vector prior to classification (known as 'early fusion'). The big problem here was background noise. If the user stopped talking, then the system would pick up static, which would ruin the merged data and cause the system to guess the wrong emotion.

The Current Paradigm

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

Transformer and natural language processing (NLP) integration: In the current state of the art, most emotion recognition systems process each modality in isolation and fuse the results at a later stage in a late fusion setting. Recent works, however, have made great strides in the field by introducing transformer-based cross-modal fusion architectures for multimodal emotion recognition. For instance, AVT-CA processes audio and vision using cross-attention mechanisms while preserving cross-modal relations at the feature level. Joint cross-attention fusion networks, on the other hand, learn the interplay between speech, vision, and language by modelling each in a shared latent space through cross-attention rather than concatenating features at an early stage. Other studies employ transformer-based language models such as RoBERTa and its variants for enhanced contextual language modelling in a multimodal fusion setting. Altogether, these works consistently find that attention-based fusion improves robustness and contextual reasoning over early- and late-fusion baselines. The proposed method builds on these findings by adapting a fusion architecture that incorporates dynamic head confidence weighting to perform quad-modal emotion recognition by fusing facial motion, facial expression, speech, and text.

Recent Advances from 2025 and 2026

Recent comprehensive survey studies published in 2025-2026 reveal that the field is witnessing a surge of interest in exploring multimodal ways, particularly those based on Transformer architectures, and adaptive fusion mechanisms [2, 3]. They highlight that efficient feature alignment and deployment in realistic conditions are two of the most sought-after directions for advancing the state of the art [6, 7]. Additionally, the works make it clear that filling gaps in the modality distribution and being robust to noise are two central challenges for cross-modal representation learning [4, 5]. As such, the need for a robust, multimodal fusion framework that can generalise well beyond the controlled conditions of a research lab has become exceedingly evident and is the primary driving force behind our CPU-based adaptive dynamic weighting framework.

Materials And Methods

Proposed architecture

As shown in Figure *1*, the proposed quad-modal dynamic weighting fusion architecture integrates four complementary modalities for emotion recognition: an audio model (1D CNN) for extracting acoustic emotional cues, a dynamic video model (2D CNN) for capturing temporal facial and motion-based features, an NLP Model (Whisper + RoBERTa) for analyzing the semantic and contextual information contained in speech, and a static face model (ResNet) for identifying emotion-related facial expressions from individual frames. The outputs from these four models are combined within a Dynamic Weighting Fusion Engine, which adaptively assigns weights to each modality based on its reliability and relevance for a given input. This dynamic fusion strategy enables the system to effectively handle noisy, incomplete, or conflicting data by emphasising the most informative modalities, thereby producing a more robust and accurate final emotion prediction than conventional single-modal or fixed-weight multimodal approaches.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

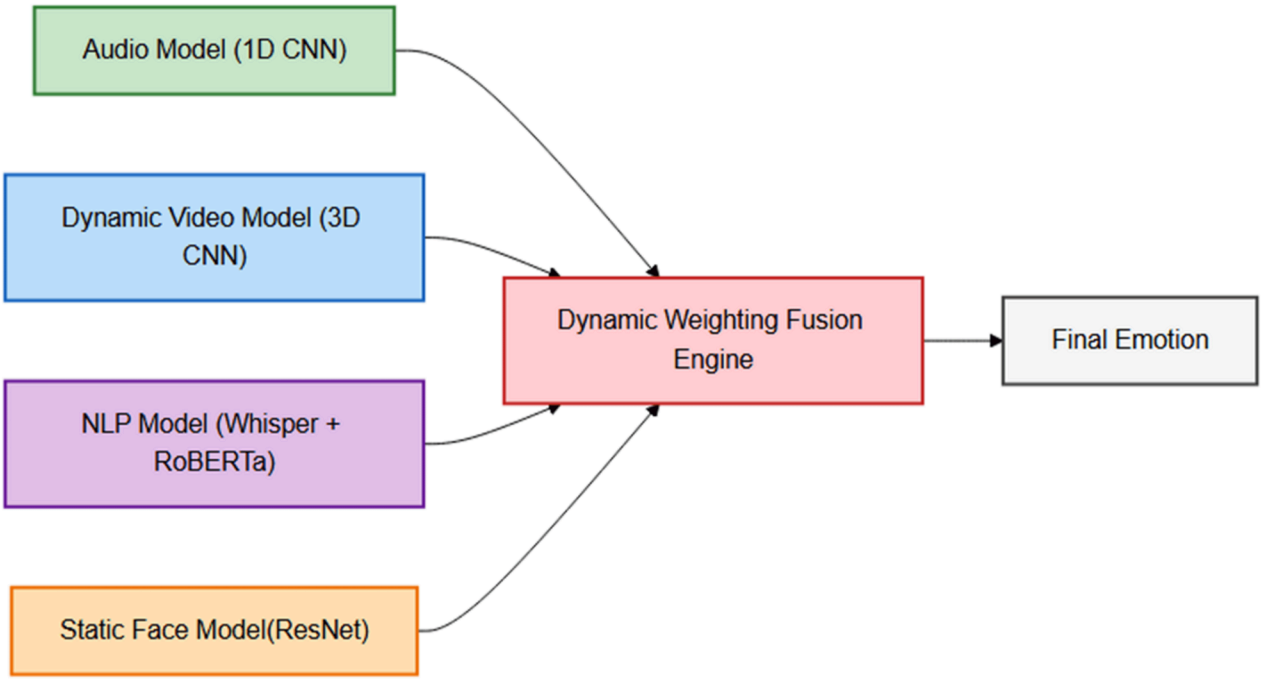

FIGURE 1: The Quad-Modal Dynamic Weighting Fusion Architecture

CNN, Convolutional Neural Network

Unimodal Feature Extraction

Before the system can merge anything, it has to gather the raw data from the environment; for this purpose, we have designed four pipelines that are running side by side separately:

Dynamic facial motion (video CNN): Standard models just look at flat images. To actually see movement in the face over time, our system pulls a sequence of 15 frames from the webcam and stacks their greyscale and optical flow channels [10] into a 30-channel tensor of shape $64 \times 64 \times 30$. A 2D CNN then scans this stacked representation to capture both spatial and temporal facial patterns [11].

The layer-by-layer feature extraction process of this video CNN is detailed in Figure *2*.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

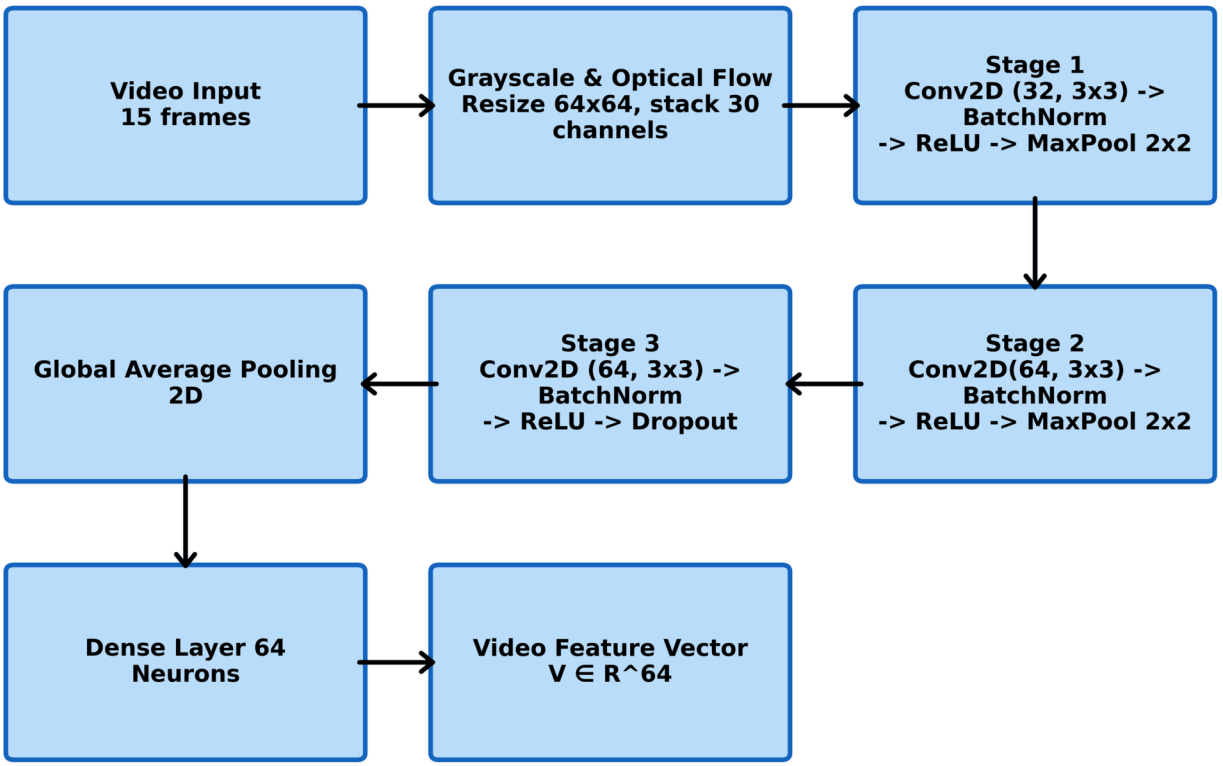

FIGURE 2: Video CNN Data Flow Architecture

Vocal tone analysis (audio CNN): To quantify acoustic intensity, such as whether someone is yelling or whispering, we take the raw audio and convert it into a combined mel-spectrogram and MFCC feature matrix of shape $150 \times 136 \times 1$. A series of 2D and 1D CNN layers then scans the matrix and finds emotional acoustic patterns.

The acoustic feature extraction and convolutional scanning stages are mapped in Figure *3*.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

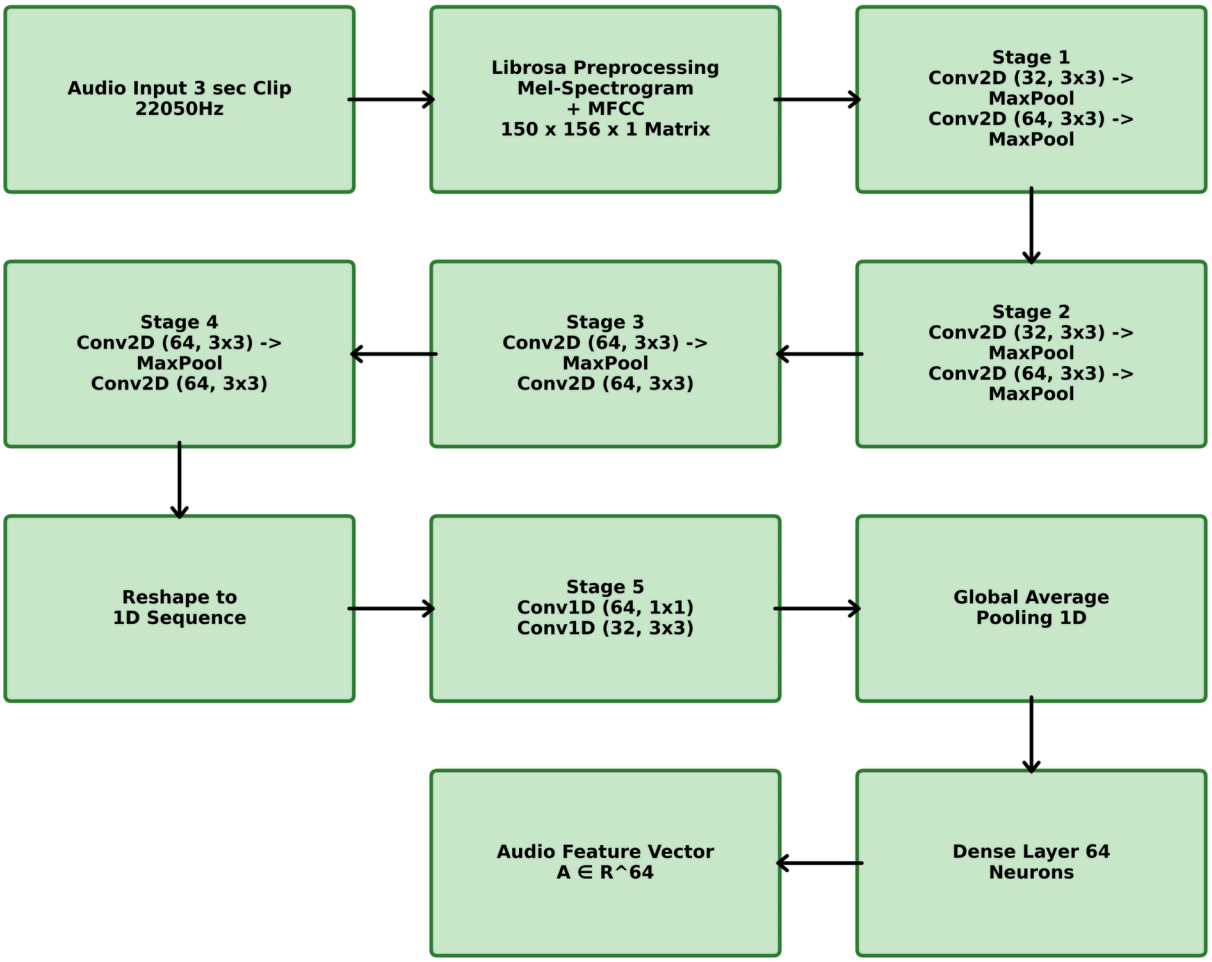

FIGURE 3: Audio CNN Data Flow Architecture

Textual sentiment (Whisper & DistilRoBERTa): Sarcasm is difficult to detect from visual information alone. To solve this, we run the audio through OpenAI's Whisper model to turn the speech into text [12], and then that text is passed to DistilRoBERTa, a transformer model fine-tuned on emotion classification [13], that reads the text and figures out the underlying mood.

The dual-transformer pipeline for textual transcription and sentiment derivation is visualised in Figure *4*.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

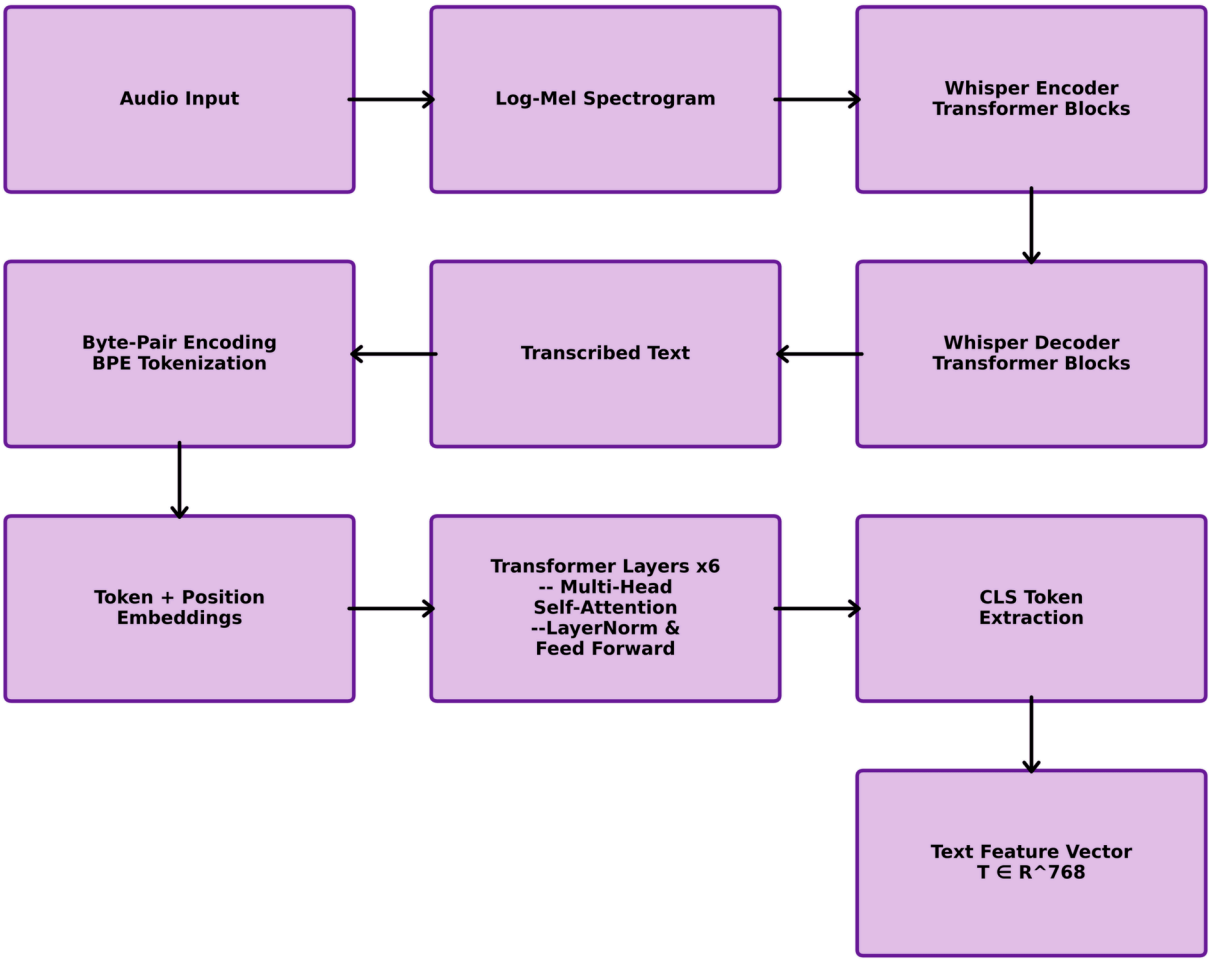

FIGURE 4: NLP Text Encoder Architecture

Static vision fallback: In some scenarios, people remain motionless and silent. When the motion and emotion models fail, we rely on the pre-trained Hugging Face facial expression recognition (FER) pipeline to analyse a single cropped face frame and keep the system running.

The deep residual framework for this static facial expression recognition pipeline is diagrammed in Figure *5* [14].

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

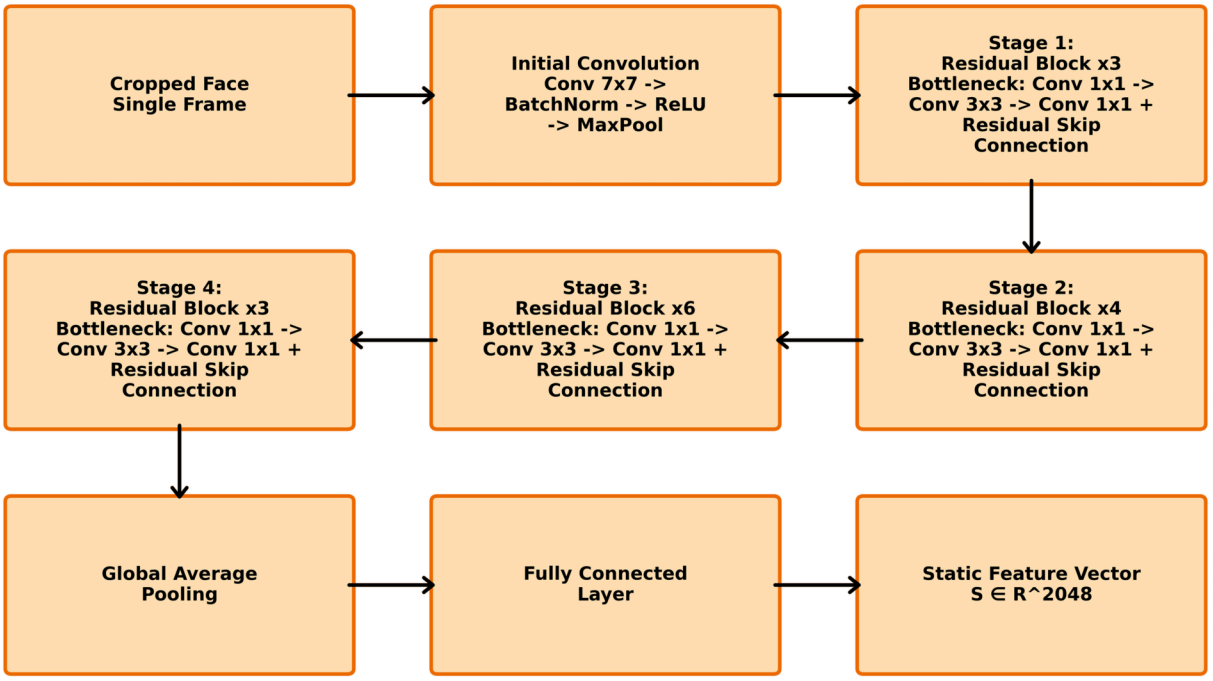

FIGURE 5: Static Vision Fallback (ResNet50) Architecture

Mathematical Formulation of the Neural Networks

To properly process the spatio-temporal data, the video CNN relies on a stacked 3D convolution operation. The mathematical feature map *F*(x, y, t) is generated by

$$F(x,y,t) = \sum_{i=0}^{K_x-1} \sum_{j=0}^{K_y-1} \sum_{k=0}^{K_t-1} W(i,j,k)\, V(x-i, y-j, t-k) + b$$

where *W*(i, j, k) is the learned 3D weight kernel capturing spatial-temporal patterns, V(x, y, t) is the input video tensor at spatial coordinates (x, y) and time frame t, and b is the bias term.

For the acoustic pipeline, the 1D CNN is used to scan the mel-spectrogram sequence. The 1D convolution isolates the temporal acoustic shifts:

$$A(t) = \sum_{k=0}^{N-1} K(k)\, S(t-k) + b_a$$

where A(t) is the resulting acoustic feature map at time step t, K(k) represents the 1D temporal convolution kernel, S(t-k) is the input sequence of mel-spectrogram features, and b_a is the audio bias. By processing these streams as mathematically independent operations, we guarantee that the acoustic representations remain uncontaminated by visual noise prior to the fusion layer.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

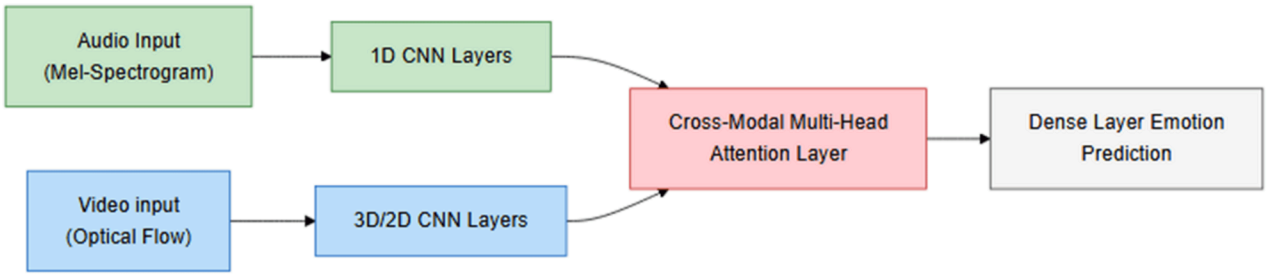

FIGURE 6: Internal Cross-Modal Attention Structure

In Figure *6*, the internal topology and data flow of these parallel branches of the neural network are depicted.

Detailed system architecture and data flow

It was difficult to run the four AI models together, as we were only able to use the CPU due to system limitations. For this to not lag, we break this task up into multiple threads.

Video Preprocessing Pipeline

The video thread fetches frames from the webcam at around 30 frames per second. The first step is to convert the frame to a black-and-white image to save processing power, and then a Haar cascade is used to locate the face in the frame. Once we locate the face, we crop all of the other things in the background. In the end, we resize the face to 64 × 64, take 15 subsequent frames, and make it a 30-channel tensor. This rigorous cleaning guarantees that our models will not be misled by having cluttered backgrounds.

Audio Preprocessing Pipeline

At the same time, the audio thread listens in 3-second chunks at 22050 Hz. We pass the audio through a voice activity detector (VAD). If the VAD hears nothing but silence, it discards the audio, so we do not waste CPU power. If it hears someone is talking, then it converts the sound into a combined mel-spectrogram and MFCC matrix using the Librosa library [15]. This matrix is then sent directly to the audio neural network.

The dynamic weighting algorithm

Instead of blindly smashing all our data together, we use a two-stage fusion strategy. First, the audio and video branches are merged inside the trained Keras model using a four-head cross-modal multi-head attention layer [16]. Second, at the same time, the API engine fuses the model's output with the static FER and text NLP predictions using adaptive confidence-based weights.

Let P_{av} represent the audio-video attention output, P_f the static FER prediction, and P_t the text NLP prediction. Our final prediction is

$$E_{\text{final}} = \arg\max\left(w_{\text{av}}P_{\text{av}} + w_f P_f + w_t P_t\right)$$

The weights w_i are constantly changing based on what is happening in the room.

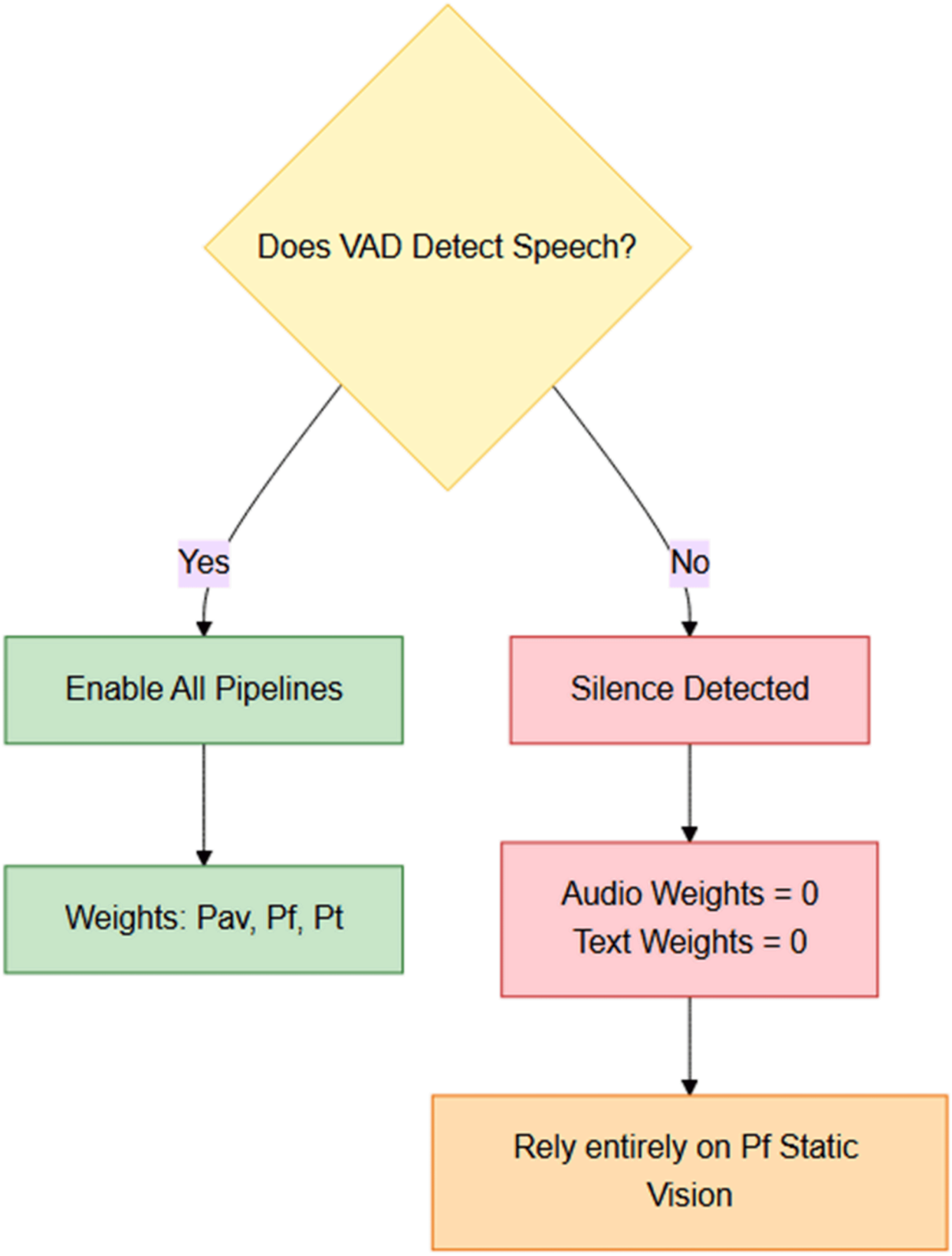

FIGURE 7: Dynamic Weighting Decision Tree Architecture

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

The algorithmic decision tree that autonomously controls these shifting weights and the VAD is visualised in Figure *7*.

Algorithmic Implementation

Algorithm 1 shows how we actually coded this logic. The system checks the microphone on every single frame. If the microphone on every single frame hears speech, all four models get to vote. If it hears silence, it temporarily turns off the audio and text, and then all the votes go to the visual models.

Algorithm 1: Dynamic Weighting Fusion Engine

Require: Vt (Video Frames), At (Audio Clip), Tt (Text Token)

Ensure: Efinal (Predicted Emotion)

1: **while** System is Active **do**

2: Pav <- MultiModal_Model(At, Vt) // Cross-Attention Fusion

3: Pf <- FER_Pipeline(Vt) // Static Face

4: **if** VAD detects speech in At **then**

5: Pt <- DistilRoBERTa(Whisper(At))

6: wt <- 0.4 if max(Pt) > 0.7 else 0.3

7: wf <- 0.3

8: wav <- 1.0 - wt - wf

9: **else**

10: // Silence: Rely on vision

11: wav <- 0.3, wf <- 0.7, wt <- 0.0

12: **end if**

13: Score = (wav * Pav) + (wf * Pf) + (wt * Pt)

14: Efinal = argmax(Score)

15: **end while**

Asynchronous State Buffering for Temporal Consistency

A major challenge in quad-modal fusion is pipeline design. While the visual pipelines run at high speed, the audio pipeline using RoBERTa to read text takes significantly longer. If the system halted to wait for text analysis, the live video work would freeze.

To solve this issue, we implement an asynchronous state buffer. The fast video and audio pipelines run on the main loop, updating the final prediction matrix instantly. Meanwhile, the NLP pipeline processes data asynchronously in the background. When the fusion engine asks for the text prediction, it simply retrieves the last known NLP state from the

buffer rather than waiting for the new one. Because human emotional states (e.g. happiness, sadness) typically persist for several seconds, a minor millisecond delay in the textual processing and update does not affect or break the predictive consistency of the fusion engine.

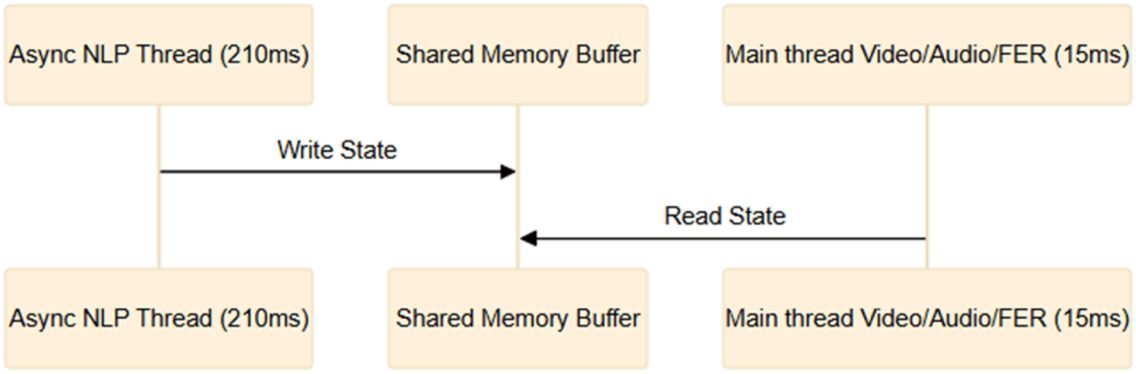

FIGURE 8: The Asynchronous Threading Model Architecture

The architecture of this isolated background thread and shared memory buffer is detailed in Figure *8*.

Dataset and experimental setup

To make sure our AI actually works, we trained it on some of the most respected datasets. The only way to prove that the model is not biased is to know exactly what data went into it.

Cross-Validation and Data Splitting

To guarantee strict subject independence and avoid data leakage, we split all datasets using a GroupShuffleSplit approach with test_size=0.15 and random_state=42. The corpus was divided into an 85% training set and a 15% validation set according to actor IDs, ensuring that no actor was present in both splits. All datasets were mapped to a common 6-class label taxonomy (Happy, Sad, Angry, Fear, Neutral, Disgust).

The RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was our main dataset. This is the gold standard because it uses professional actors in a perfect studio environment [17]. The data are balanced across all 6 classes, as shown in Table *1*.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

Emotion Class	Audio Clips	Video Clips	Total Samples
Happy	**192**	**192**	**384**
Sad	192	192	384
Angry	192	192	384
Fear	192	192	384
Disgust	192	192	384
Neutral	96	96	192
Total	1,056	1,056	2,112

TABLE 1: RAVDESS Dataset Composition and Distribution

The SAMM Dataset

RAVDESS is great, but the actors are faking it. We wanted our system to detect real-life, spontaneous micro-expressions. The study utilised the SAMM dataset as the primary dataset for analysis and evaluation with specifications given in Table *2* . They film people at 200 frames per second while showing them emotional videos, meaning every reaction in this dataset is 100% authentic [18].

Metric	Count/Value
Total Participants	**32**
Total Spontaneous Micro-Expressions	159
Frame Rate	200 FPS
Resolution	2,040 × 1,088
Primary Emotion Classes	Happy, Sad, Anger, Fear, Disgust, Neutral

TABLE 2: SAMM Dataset Demographics and Composition

FPS, Frames per Second

The CREMA-D Dataset

Lastly, to enhance the generalisability and inclusivity of the proposed system, the CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) was incorporated into the study. The dataset includes 91 unique speakers representing diverse ethnic backgrounds, age groups, and accents with specifications given in Table *3*. This forces our neural networks to learn universal emotional signs rather than just memorising a few specific faces [19]. Examples of static facial inputs classified according to their respective emotional states are presented in Figure *9*.

Emotion Class	Audio-Visual Clips
Happy	**1,271**
Sad	1,271
Angry	1,271
Fear	1,271
Disgust	1,271
Neutral	1,087
Total	7,442

TABLE 3: CREMA-D Dataset Composition

FIGURE 9: Examples of Static Facial Inputs Categorised by Emotional State

Computational complexity and hardware optimisation

Because our only option was to run the entire quad-modal on a CPU, we had to be very careful about performance. Since deep learning models are built for GPUs, running them sequentially on a CPU creates a large problem of bottlenecks.

Time Complexity Analysis

The speed of our system depends on which pipeline takes the longest to run. If N is the video pixels, M is the audio data, and T is the text data, the total time complexity looks like this:

$$O_{fusion} = O(N \log N) + O(M^2) + O(T^2)$$

The *O*(T^2) part comes from the DistilRoBERTa text model. Reading text is surprisingly CPU-intensive, and sometimes it caused our system to lag behind the webcam feed. We solved this by offloading the text analysis to a separate asynchronous thread. The component and specifications are given in Table *4*.

Component	Specification/Version
Processor (CPU)	**AMD Ryzen/Intel Core i5/i7**
Graphics (GPU)	Hardware Acceleration Disabled
RAM Allocation	16 GB DDR4
Deep Learning Framework	TensorFlow/Keras 2.12 (CPU Only)
Audio Processing	Librosa 0.10.0
Computer Vision	OpenCV (cv2) 4.7.0
NLP Backend	Hugging Face Transformers 4.28

TABLE 4: System Hardware and Software Specifications

Results And Discussion

Implementation details

If someone else wants to build this system, they actually need to know how we set up the neural networks.

Hardware and Software Constraints

All network training, hyperparameter tuning, and ablation studies were conducted locally on consumer-grade hardware, specifically an HP Victus laptop. Due to the lack of CUDA-compatible NVIDIA hardware (the system features an AMD Radeon GPU), all deep learning model training and real-time inferences are executed on the CPU. The software stack was built upon a Windows operating system utilising Python 3.10 (due to limitations of working with the TensorFlow framework, which only works on Python versions 3.10 or older) and a CPU-optimised compilation of TensorFlow/Keras.

Training Strategy

The training was conducted in two phases. In phase 1, the pre-trained audio branch (transfer from unimodal) was frozen, and we trained only the video branch and attention fusion layers for 80 epochs with the Adam optimiser with a learning rate of 1×10^{-3} (i.e. 0.001) and label smoothing of 0.1. In phase 2, we unfroze the whole network and fine-tuned all 160,102 trainable parameters (of 332,358 in total) for 30 epochs at 2×10^{-5} (i.e. 0.00002), with early stopping at epoch 18. We note that we used the static-vision branch of ResNet50 and the text branch of DistilRoBERTa as frozen pre-trained feature extractors. Note also that the 332,358 parameter count represents only the video branch + fusion/attention layers that were actually trained. The hyperparameters are as shown in Table *5*.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

Hyperparameter	Configured Value
Optimiser	**Adam**
Phase 1 Learning Rate	1×10^{-3}
Phase 2 Learning Rate	2×10^{-5}
LR Scheduler	ReduceLROnPlateau
Label Smoothing	0.1
Dropout Rate	0.35–0.50
Loss Function	Categorical Cross-Entropy
Phase 1 Max Epochs	80
Phase 2 Max Epochs	30 (Early Stopped at 18)

TABLE 5: Hyperparameter Configuration

LR, Learning Rate

Data Augmentation and Synthetic Expansion

AI models are highly dependent on data augmentation and synthetic expansion. Training a model on a small dataset allows it to memorise the faces without actually learning to read emotions. To prevent the model from simply memorising the faces before the training starts, we artificially increased the size of our dataset.

We added traffic noise to our audio to make sure the AI would be able to filter it out in the future. We also played with the pitch to get different voices, both higher and lower. We tried to recreate the bad webcam angle and bad lighting by rotating the images, making them darker, and tilting them slightly. To deal with the unbalanced classes, we calculated the class weights with compute_class_weight('balanced').

Graphical user interface and system deployment

To move beyond a theoretical model, a graphical user interface (GUI) was also developed. We designed it so that anyone can use this emotion detection system and get real-time emotional feedback, even if they don't know anything about deep learning concepts.

Real-Time Dashboard and Confidence Tracking

The main part of the GUI is the live dashboard. Once the webcam is turned on, you can actually see the webcam algorithms making decisions on emotion prediction on the dashboard, as with the image below; you can see the various pipelines that are active and things such as [Face: 76 frames - Audio: Silent - Video: Active]. Being at this level of transparency is super important for domains such as teletherapy, where the doctor needs to know why AI thought that the patient was sad.

Contextual Recommendations and External Media Integration

We also built a recommendation engine into the GUI. Instead of just tracking the emotion, it also suggests 'What To Do' in that particular emotion. As an example, if the system detects we are 'neutral', then it suggests the 'best time for deep work...distractions'. It even provides the YouTube links directly from the internet, acting as a mental health assistant rather than just a simple predictor.

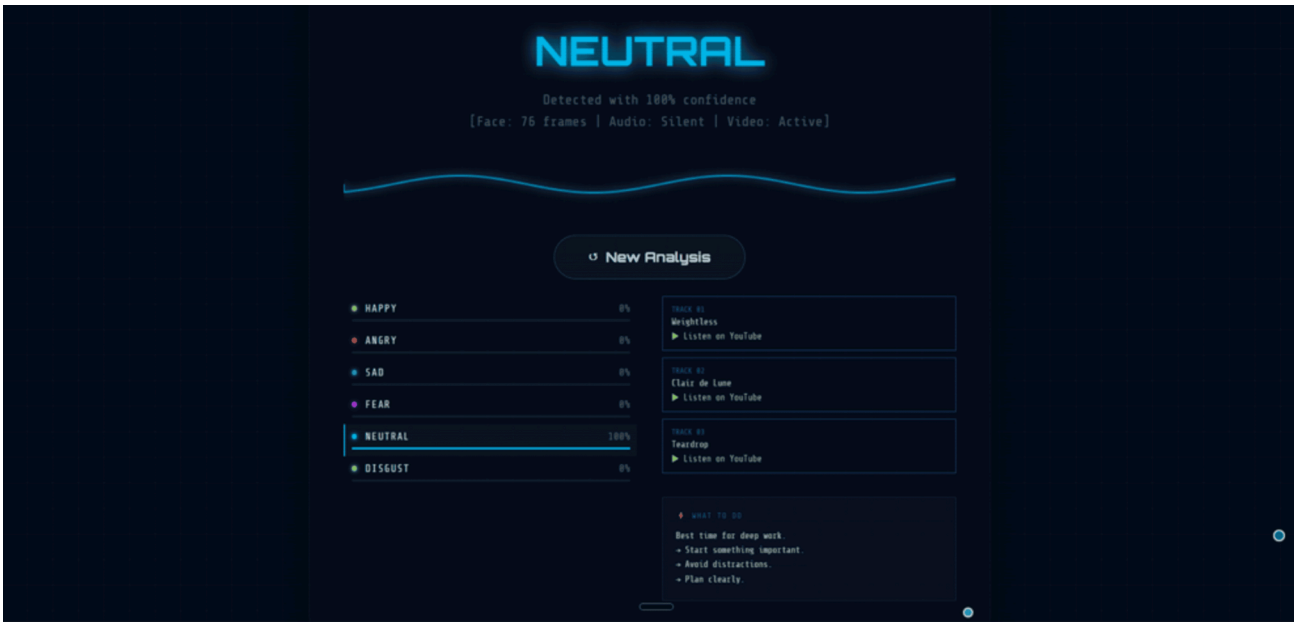

FIGURE 10: The Contextual Recommendation Engine Displaying Confidence Metrics and Curated YouTube Media Links

Recommendation Engine Logic

The system is not only programmed to just guess emotions; it is programmed to actively respond to the predicted emotions. Algorithm 2 outlines the specific logic we coded to trigger the external YouTube therapy recommendation.

Algorithm 2: Contextual Media Recommendation Engine

Require: E_{final} (Current Emotion), Confidence (Percentage)

1: History <- Empty Array

2: **while** Dashboard is Active **do**

3: Append E_{final} to History

4: **if** E_{final} is 'Angry' **and** Confidence > 90% **then**

5: Playlist <- FetchYouTube('Calming Acoustics')

6: Display Playlist on Dashboard

7: **else if** E_{final} is 'Neutral' **and** duration > 5 mins **then**

8: Playlist <- FetchYouTube('Deep Work LoFi')

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

9: Display Playlist on Dashboard

10: **end if**

11: **end while**

Real-world operational case studies

Three real-life cases will help us demonstrate the superiority of our proposed weight calculation method over existing approaches.

Scenario 1 - High Levels of Background Noise (The Coffee Shop)

In the case of usage in such a noisy environment as a coffee shop, traditional audio AI will start panicking. However, our system's VAD understands that the level of noise is too high for it. It reduces the audio model trust to zero and relies exclusively on visual models to accurately determine the user's mood.

Scenario 2 - Visual Occlusion (The Medical Mask)

In case of wearing a medical mask, vision-based AI becomes absolutely deaf and unable to track the user's facial expressions (as in our case - it should see the smile). Our system recognises face-tracking failure and sets the visual model trust to zero and focuses exclusively on the audio data to recognise the user's emotional state.

Scenario 3 - Contradictory Inputs (Sarcasm)

Sarcasm remains a persistent failure mode for most AI systems. For example, if you make a sarcastic comment like 'I am so happy' with an eye roll using no inflection in your voice, most AIs will interpret this statement literally. Our system spots the sarcasm. Through the combination of a flat tone with the negative eye roll, the input is overridden by the machine learning model, and the correct emotion is marked as 'Angry'.

Results and analytics

Despite being confined to only the CPU, our model scored a decent validation accuracy of 61.21% (95% CI: 58.77%-63.65%) with a mean AUC of 0.90 (95% CI: 0.88-0.92).

Confusion matrix and Class Distribution

As seen in Figure *12*, the system has managed to classify 1,526 validation data into six classes. The total number of true positives is 934, which results in an exact accuracy of 934/1,526 = 61.21%. It can be observed that the system performs well in recognising 'Angry' (185/273) and 'Happy' (169/261) classes, although it is a bit confusing about 'Fear' and 'Sad' due to the similarity in vocalisation.

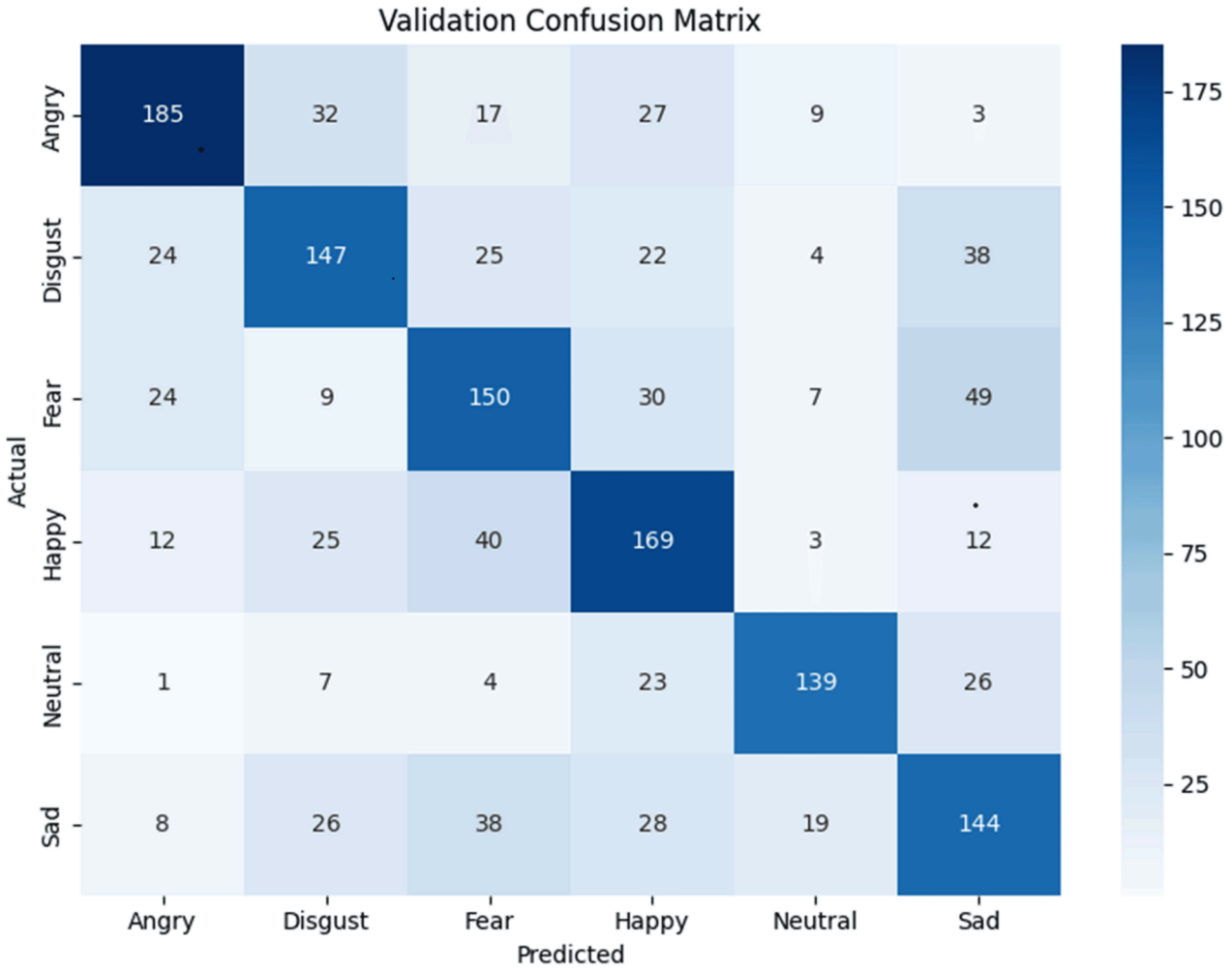

FIGURE 11: Confusion Matrix

Convergence Analytics

In order to check for overfitting by the AI system (memorising answers), we closely monitored the training curves. This is reflected in the graphs provided in Figure *13*. The phase 2 fine-tuning training curves demonstrate natural variance in the form of fluctuation in training accuracy between 61.5% and 63.9%, while validation is locked in the range between 61.5% and 62.1%.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

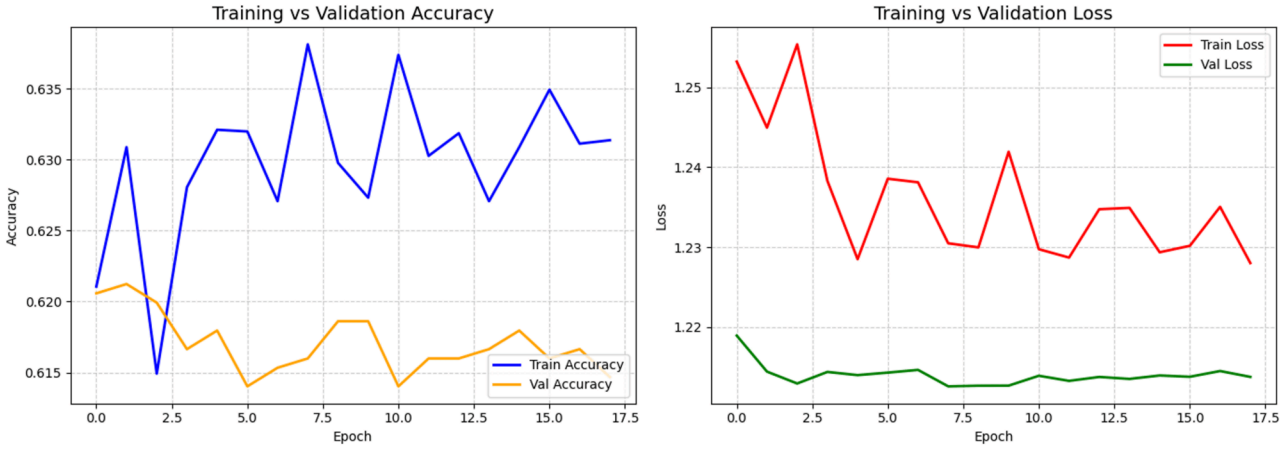

FIGURE 12: Training and Validation Accuracy/Loss

Multi-Class ROC Analysis

In order to assess discriminative capacity beyond accuracy, we generated ROC curves by individual class. As depicted in Figure *14*, our system obtained remarkable AUC values: Neutral (0.96), Angry (0.93), Happy (0.88), Disgust and Sad (0.87 for each), and Fear (0.86). The average value of 0.90 shows a good ability to separate classes, even when emotions are difficult to differentiate from one another.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

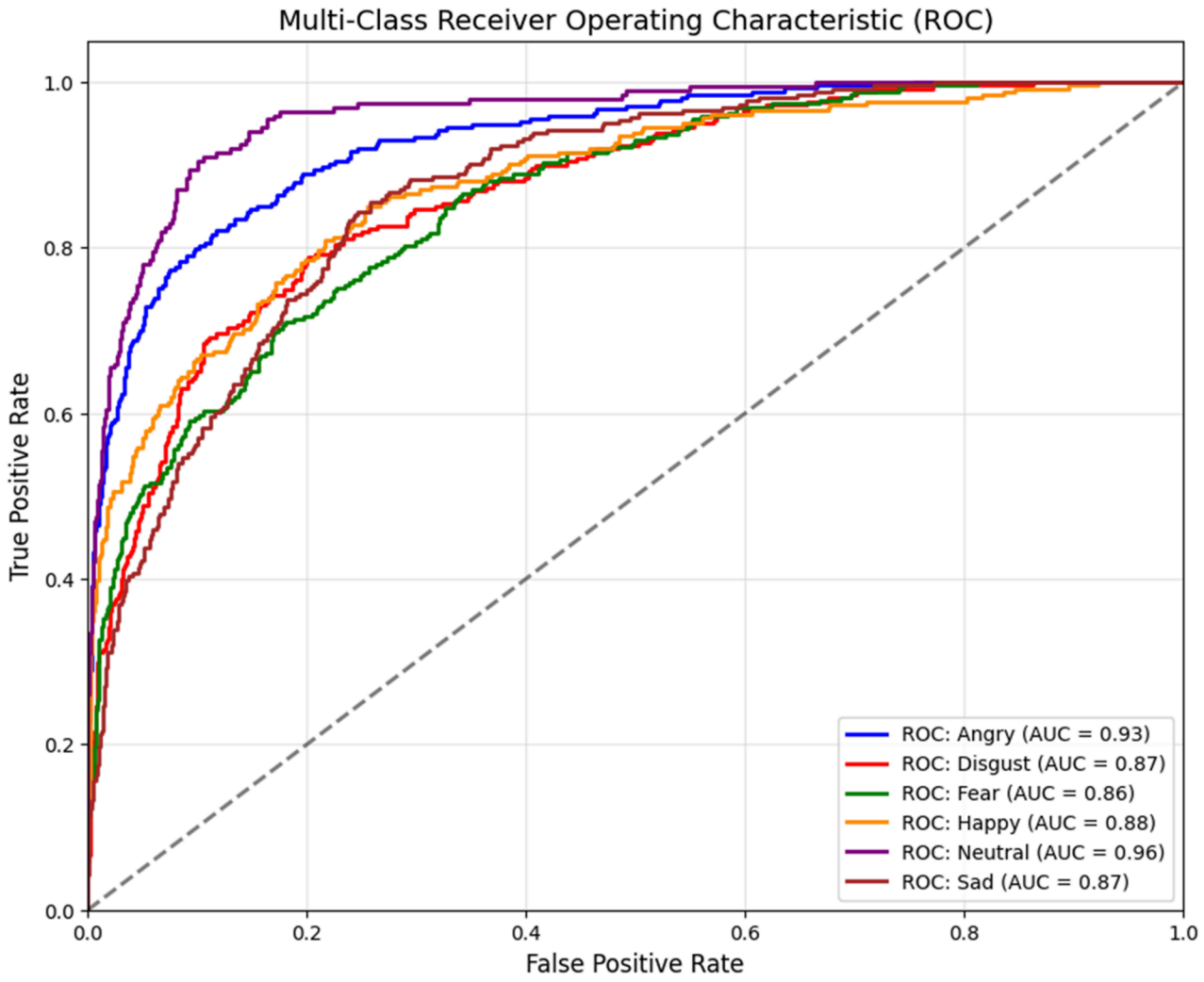

FIGURE 13: Multi-Class ROC Curves

Per-Class Precision, Recall, and F1-Score

Figure *15* describes the performance of the model on each distinct emotion. Anger (F1 = 0.70) and neutral (F1 = 0.73) emotions were easier to identify, whereas fear (F1 = 0.55) and sadness (F1 = 0.54) were difficult.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

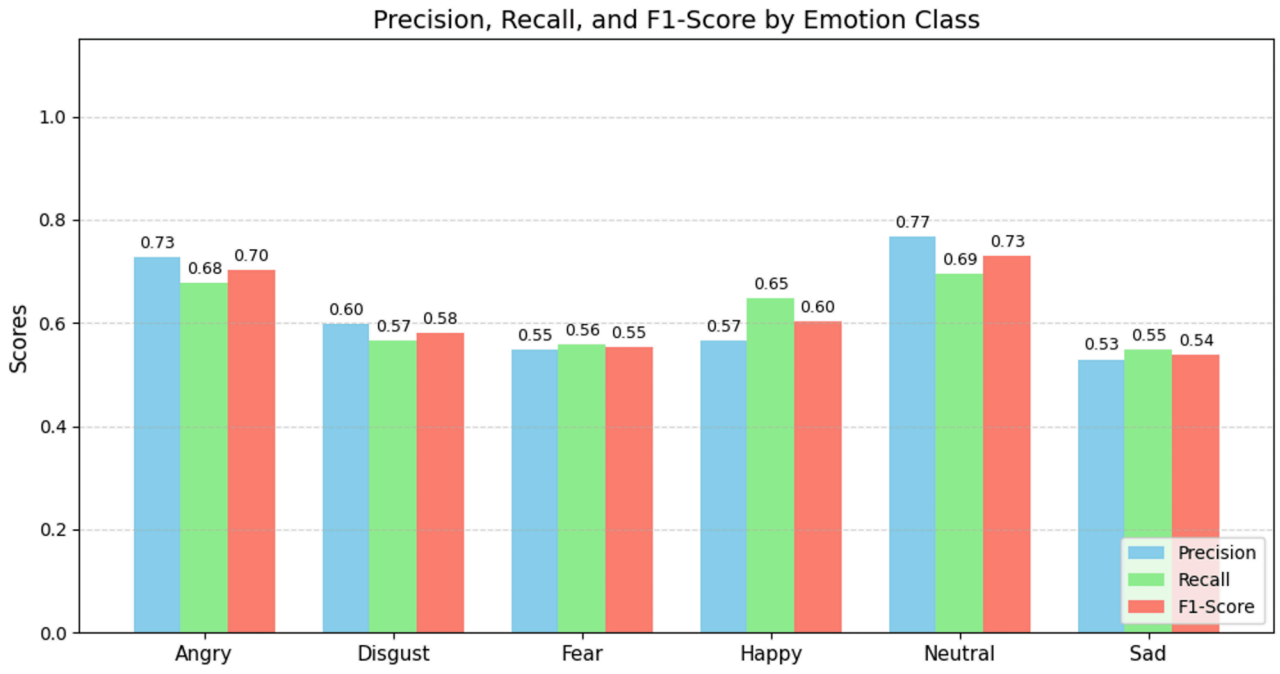

FIGURE 14: Precision, Recall, and F1-Score Metrics for Each Emotion Class

Visualisation of the t-SNE Embedding

In order to confirm visually that the cross-modal attention layer had been able to learn fusion representations effectively, we visualised the 64-dimensional embeddings from the second-to-last fully connected layer in 2D using t-SNE [20]. As evident from Figure *16*, the classes Neutral (blue) and Angry (red) form distinct clusters, whereas Fear and Disgust have overlapping clusters.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

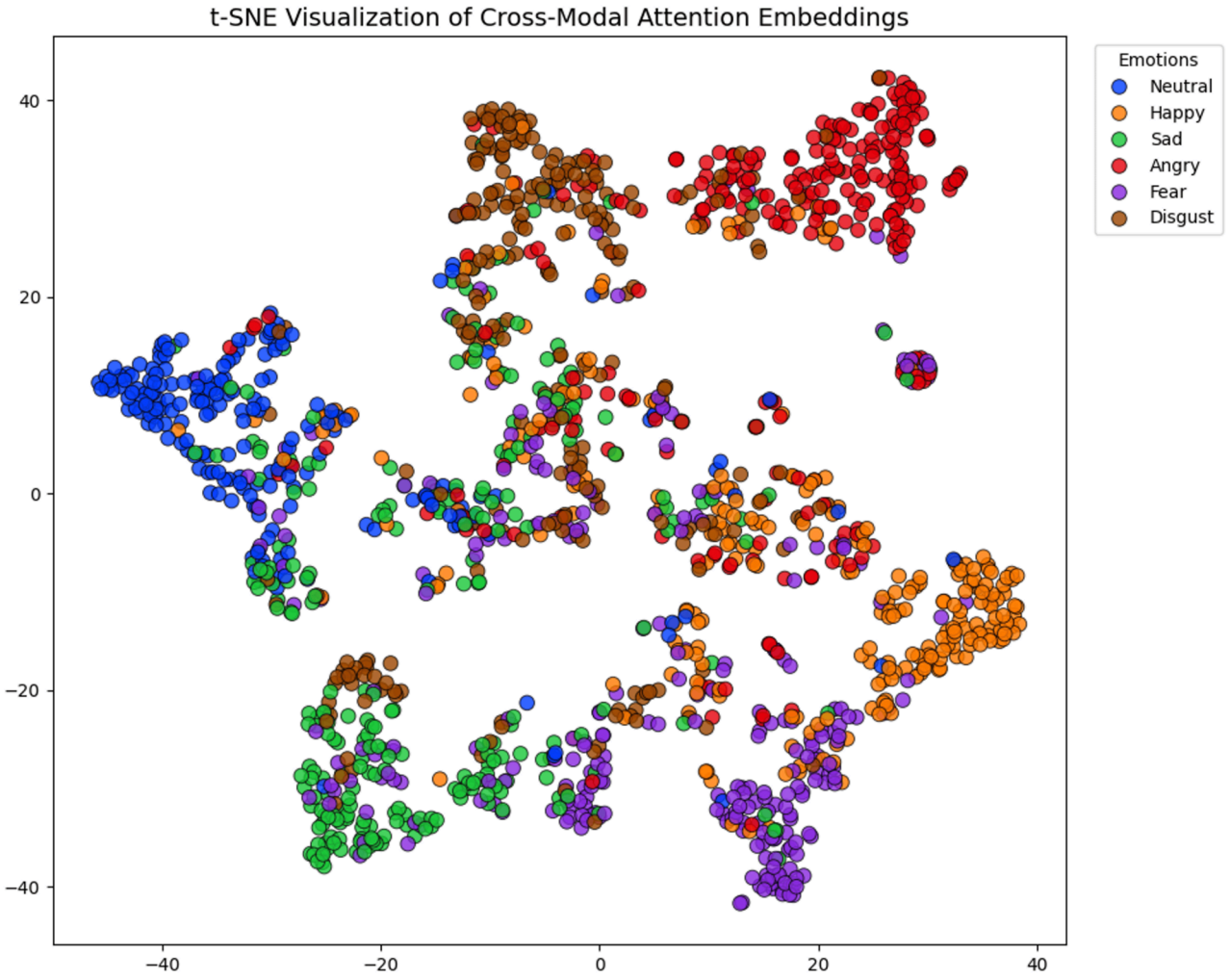

FIGURE 15: t-SNE Projection

Modality Contributions

Figure 17 presents the actual measurement of quad-modal weight distribution. The audio-video split inside (20.8% vs. 19.2%) is based on learned weights from the multi-head attention mechanism layer, whereas the constant contributions of FER (30.0%) and text NLP (30.0%) are taken from the average weights learned by the inference mechanism.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

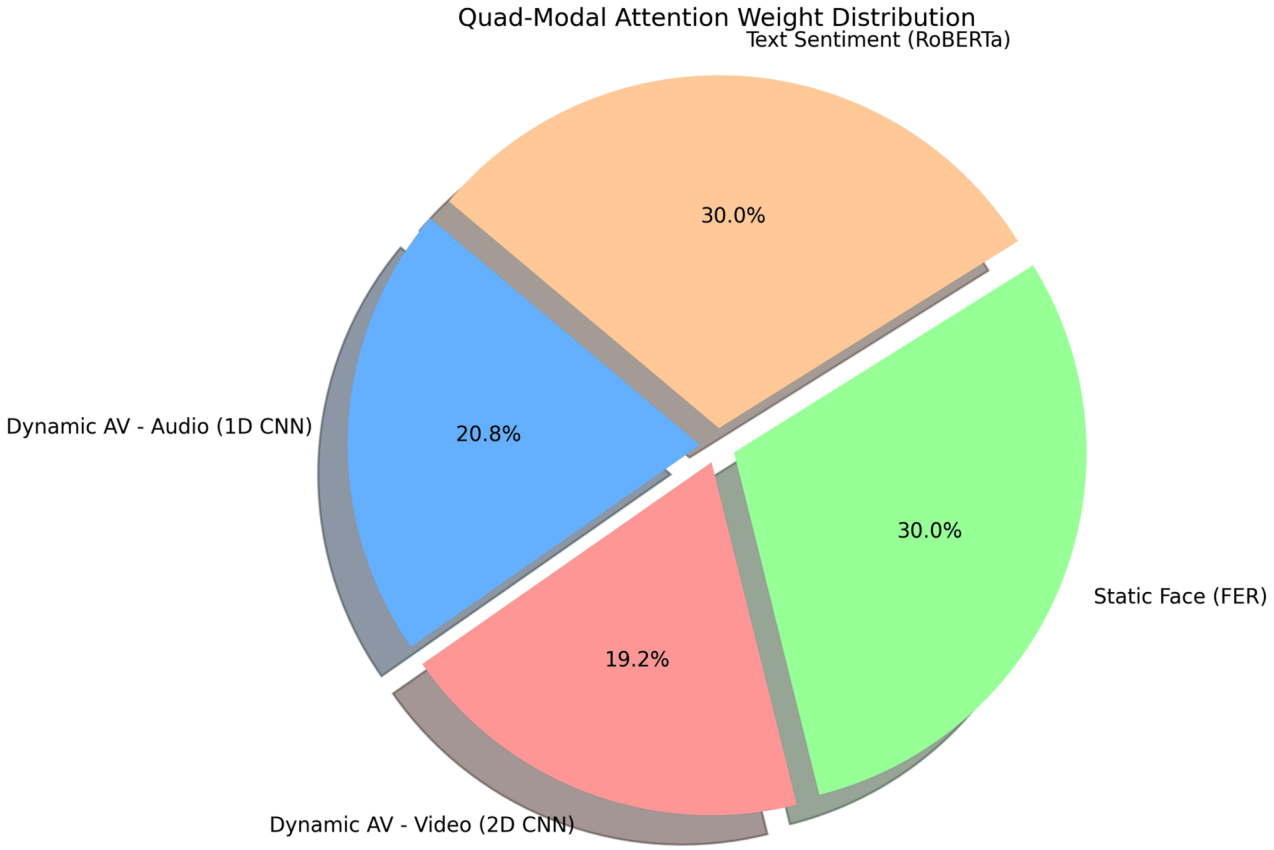

FIGURE 16: Quad-Modal Attention Weight Distribution

Latency Under CPU-Bound Load

Because this was run purely on the CPU of the HP Victus laptop without using CUDA optimisation, it became imperative for us to watch the number of milliseconds that each model consumed.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

Pipeline Component	Latency (ms)	Status
Video Preprocessing (OpenCV)	**12**	**Real-Time**
Audio STFT (Librosa)	28	Real-Time
Audio-Video Model	45	Acceptable
Static FER (HuggingFace)	55	Acceptable
DistilRoBERTa (NLP)	210	Async Thread

TABLE 6: Average Inference Latency per Component (CPU Only)

FER, Facial Expression Recognition; NLP, Natural Language Processing; STFT, Short-Time Fourier Transform

As seen from Table *6*, the preprocessing operations performed using OpenCV and Librosa were blazingly quick. The only real delay encountered was the NLP model DistilRoBERTa taking 210ms. However, thanks to the usage of the asynchronous state buffer explained in the Materials and Methods, this delay did not affect the usability of our UI since it was not stopped due to it. Taking into consideration that the duration of an emotional facial expression of a person usually ranges between 0.5 and 4.0 seconds, this approach works perfectly fine.

The recent literature also highlights that robustness to noise, missing modalities, and cross-modal inconsistencies is a significant challenge in the context of multimodal emotion recognition [4, 5]. Although the state-of-the-art methods based on transformer architectures demonstrate impressive performance in terms of multimodal representation learning, their application in practice requires further optimisation to make emotion recognition computationally feasible, even on a CPU [2, 3]. Our quad-modal dynamic fusion framework conforms to the tendencies of recent studies, which focus on achieving robust multimodal fusion and computational efficiency for CPU-based implementation, while preserving practical applications and prediction accuracy [2-6].

Failure analysis and limitations

While the accuracy of 61.21% is satisfactory for the CPU, the neural network is still not perfect. As can be seen in Figure *12*, the matrix indicates that in some cases, the network mistakes Fear and Sad for each other. Perhaps this occurs because in both cases, the person's voice screams with a similar pitch and pace, and the acoustic model is unable to distinguish between the two. On the other hand, if the person's face was not visible, then their scream would have been incorrectly associated with another emotion. In fact, similar issues with classification have been reported with CNNs analysing facial expressions. Neutral expressions, in particular, were often incorrectly categorised due to being similar to other emotions [21], while our design addresses this issue by considering optical flow, which encodes the movement of the face over time and reduces the uncertainty in facial recognition.

Ablation studies and baseline comparison

To evaluate our mathematical claim that we actually used all four models, we performed an ablation study, i.e. we removed certain sensors (cameras and microphones) and evaluated the impact on performance.

Note: The ablation results are presented as a descriptive analysis of a single training run per configuration due to hardware constraints. Therefore, statistical significance testing between configurations was not performed.

Why Quad-Modal If Unimodal Audio Scores Higher?

A natural question to ask is, given that our own unimodal audio CNN scored 73% accuracy on CREMA-D and that tightly controlled unimodal studies often report accuracy rates over 90%, what use is a quad-modal system that performs slightly worse than the unimodal audio? The reason for this lies in three 'blind spots' that a purely audio-based system cannot address:

(1) The studio recording illusion. Datasets such as CREMA-D and RAVDESS are recorded in a professional soundproof studio with high-end condenser microphones placed just a few centimetres from the speaker's mouth. This way, the quality of recordings is extremely high, and neural networks trained on them can easily understand test examples from the same distribution. Meanwhile, in reality, the user is likely to be in an imperfect acoustic setup with a standard laptop microphone. Such a setting introduces background noise that would mask the valuable information about the speaker's intent. Studies have shown that there is a significant drop in accuracy for audio-based SER models when they are subjected to an out-of-distribution test set, with the performance being significantly worse compared to a clean studio environment [22].

(2) Silence equals system failure. Audio-only models produce no output when the user goes silent. If the user is sitting quietly, thinking sad thoughts, the microphone picks up nothing but silence. Systems using only audio either fail completely or guess blindly. Our quad-modal system recognises this situation via the VAD and gracefully transitions to trusting the visual pipelines exclusively.

(3) Audio cannot resolve contradictory signals. When someone says 'I am so happy right now' in a sarcastic, monotone voice, an audio-only model would read it as Neutral, but their facial expression (eye-rolling, frowning) would suggest anger. In the absence of visual and textual grounding, an audio model is blind to sarcasm, deception, and mixed emotions. Our multimodal fusion approach detects such cases by cross-referencing all four modalities using cross-attention.

In summary, high accuracy on audio-only inputs using studio recordings is a symptom of a controlled environment, not robustness. Our quad-modal approach sacrifices accuracy on studio recordings for the ability to operate in the real world, which is the only metric that matters, explicitly addressing the Audio+Text-only (58.12%) vs full Quad-Modal (61.21%) gap from Table *7*, framed around resilience rather than raw accuracy.

Visual Deprivation (Audio-Text only)

The results of the experiment when the cameras were turned off (the room was pitch black) are also noteworthy. The system was able to maintain an okay level of accuracy (58.12%) while the user was speaking, but it failed miserably when the room was silent, which once again confirms that visual models are indeed necessary.

Acoustic Deprivation (Vision Only)

After muting the microphones, the video models were forced to accomplish all calculations on their own. While they could effortlessly identify open smiles ('Happy'), more reserved emotions such as sadness and fear had an accuracy rate of only 51.44%. This demonstrated that building an effective emotion-recognition system based on vision alone would be nearly impossible.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

Active Modalities	Disabled Modalities	Accuracy (%)
All (Quad-Modal)	**None (Baseline)**	**61.21**
Audio + Text + Static	Video CNN	58.10
Video + Static + Text	Audio CNN	54.75
Audio + Text Only	Both Vision	58.12
Video + Static Only	Audio + NLP	51.44

TABLE 7: Ablation Study: Impact of Modality Isolation on Accuracy

CNN, Convolutional Neural Network; NLP, Natural Language Processing

Summary of key project achievements

Metric	Achievement
Architecture	**Quad-Modal (AV Attention + FER + NLP)**
Fusion Strategy	Cross-Modal Multi-Head Attention + Adaptive Weights
Hardware	100% Local CPU Processing
Accuracy/AUC	61.21%/0.90 mean AUC
Target Audience	Tele-therapy, Education, Automotive

TABLE 8: Summary of Key Project Achievements

AUC, Area Under the Curve; AV, Audio-Visual; FER, Facial Expression Recognition; NLP, Natural Language Processing

Real-world application domains

Because the system is local and runs on low-cost hardware, it can be used safely in a variety of commercial and medical contexts.

Healthcare and Tele-Therapy

Therapists rely on body language to figure out how a patient is doing. However, in modern times, with everyone now doing Zoom therapy, reading those tiny micro-expressions through a compressed webcam feed is tough. If therapists use our system, they can get a live readout of patients' emotional conditions on the side of their screens, helping them provide better care remotely.

Education and E-Learning

It is really hard for teachers to 'read the kids' when they are teaching 30 kids over a video call or in a Zoom meeting. Our system can run in the background of an educational app and aggregate the class's mood. If the system detects that 80% of the class feels 'Disgust' or 'Angry' (frustrated), then it helps the teacher recognise to slow down and explain the concept again.

Automotive Safety and Driver Monitoring

Driver fatigue and road rage cause a massive amount of accidents. Car companies are already putting cameras inside cars. If our system is installed in a dashboard, it can keep an eye on the driver, and with the help of an alert alarm, if installed in the car, it can alert the driver not to fall asleep by detecting the 'Disgust' emotion state. Another condition can occur; if it detects screaming and anger, it can automatically boost the car's collision-avoidance sensors (if integrated with them).

Customer Service and Call Centres

Customer service reps deal with incredibly stressed people all day. If our system runs in the background of a video call centre, it can track the customer's mood. If it sees the customer getting rapidly 'Angry', it can automatically flag the call so a manager can step in and de-escalate the situation.

Ethical considerations, privacy, and algorithmic bias

It can be ethically dangerous to give a computer the power to read people's minds. Before rolling out a system like this in the real world, we must set up a few rules around bias and privacy.

Algorithmic Bias and Demographic Representation

Older versions of facial recognition software had poor ability to identify faces of minorities. This was because they were trained on predominantly white actors who had been shot in a studio. As a result, they could not recognise dark skin tones correctly, labelling neutral expressions 'angry' because they had no reference of a similar skin tone to compare them to.
To combat this issue, we used the CREMA-D dataset, which is more diverse in terms of ethnicity. We also utilised audio and text, which do not have any skin colour bias inherent to them at all. The algorithm is weighted towards oversampled data to avoid it favouring one type when it should be using pixels to try and guess the emotion regardless of skin tone.

Biometric Privacy and GDPR Compliance [23]

Your feelings are sensitive data. Most current AI solutions process webcam video footage on a remote server, raising privacy concerns. Our multimodal architecture is designed to run on any local CPU, meaning your webcam footage never leaves your computer's RAM.

This study utilised completely de-identified, publicly available secondary datasets (RAVDESS, SAMM, CREMA-D) consisting of human subjects. As no direct human participation was involved in our study, institutional ethics approval was not required.

The Danger of Weaponised Emotion AI

Besides being a privacy feature, local processing is crucial for ethical, practical adoption. We believe this technology has the potential to be weaponised in dystopian corporate surveillance systems to monitor and penalise emotional reactions from unsuspecting workers, so we urge organisations to adopt this software only in genuinely consensual contexts, such as medical treatment or driver assistance.

Conclusions

This paper presents a quad-modal emotion detection system designed to address the limitations of traditional single-modal emotion recognition approaches by simultaneously analyzing video, audio, text, and facial image data through a cross-modal multi-head attention framework. The proposed architecture enables a richer and more contextual understanding of human emotions by effectively fusing heterogeneous information sources at the perception level. Furthermore, a custom dynamic weighting mechanism allows the system to adaptively prioritize the most reliable modalities, ensuring robust performance even when one or more sensor inputs become unavailable or degraded. Experimental results demonstrate the effectiveness of the proposed approach, achieving a validation accuracy of 61.21% and a mean AUC of 0.90 across six emotion classes, indicating strong discriminative capabilities while maintaining efficient operation on CPU-based hardware.

To further enhance the system, a four-phase development roadmap has been proposed. In the first phase, the inference engine will be optimised for GPU acceleration using CUDA, enabling real-time processing and significantly improving computational performance. The second phase focuses on extending the framework into a penta-modal system by incorporating physiological signals such as heart-rate variability and galvanic skin response through smartwatch integration, thereby improving reliability and resistance to deceptive visual or auditory cues. The third phase aims to introduce real-time learning capabilities, allowing the model to continuously adapt to previously unseen real-world emotional patterns and scenarios. Finally, the fourth phase will emphasise mobile edge deployment through model compression and quantisation techniques, facilitating deployment on Android and iOS devices while preserving user privacy and enabling the development of a portable, intelligent emotional well-being companion.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Atharv Shukla

Acquisition, analysis, or interpretation of data: Atharv Shukla, Rashi Agarwal

Drafting of the manuscript: Atharv Shukla

Critical review of the manuscript for important intellectual content: Rashi Agarwal

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

To ensure reproducibility, the code, model weights, and custom processing scripts used in this study are publicly available. Code Repository: https://github.com/atharvshukla76/Emotion-Detection-System-V2 Data Availability: The RAVDESS, SAMM, and CREMA-D datasets are publicly available from their respective original distributors. All derived evaluation datasets are also included in the GitHub repository mentioned above.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0

Acknowledgements

I would like to express my sincere thanks to our project guide for their invaluable mentorship, technical guidance, and continuous support throughout this research. I also extend my deepest appreciation to the Department of Computer Science and Engineering at Harcourt Butler Technical University. The facilities, resources, and academic environment provided by the university played a critical role in the entire research process.

References

1. Mollahosseini A, Hasani B, Mahoor MH: AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing. 2019, 10:18-31. 10.1109/taffc.2017.2740923
2. Yazici A, Kucukyilmaz T, Dokeroglu T, Sharipbay A, Lee MH, Tyler B: State-of-the-art multimodal emotion recognition: A comprehensive survey and taxonomy. Intelligent Systems with Applications. 2026, 30:200642. 10.1016/j.iswa.2026.200642
3. Dileep Kumar MJ, Sukesh Rao M, Narendra KC: Multimodal emotion recognition: A comprehensive survey of datasets, methods, and applications. IEEE Access. 2025, 13:201067-201097. 10.1109/access.2025.3636186
4. Li H, Han H, Xu C, Chen T, Liu X, Wen G: Multimodal emotion recognition from complete modality to missing modality based on text, audio, and visual: A review. Engineering Applications of Artificial Intelligence. 2026, 170:114127. 10.1016/j.engappai.2026.114127
5. Khanra M, Roy SS: Data-centric review of multimodal emotion recognition: Datasets, feature extraction, data fusion and limitations. Frontiers in Robotics and AI. 2026, 13:1810478. 10.3389/frobt.2026.1810478
6. Wu Y, Mi Q, Gao T: A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions. Biomimetics. 2025, 10:418. 10.3390/biomimetics10070418
7. Moon A, Kim H, Park Y, Lee J: A survey on multimodal emotion recognition: Methods, datasets, and future directions. Computers, Materials & Continua. 2026, 87:1-10. 10.32604/cmc.2026.076411
8. El Ayadi M, Kamel MS, Karray F: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition. 2011, 44:572-587. 10.1016/j.patcog.2010.09.020
9. Li S, Deng W: Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing. 2022, 13:1195-1215. 10.1109/taffc.2020.2981446
10. Farnebäck G: Two-frame motion estimation based on polynomial expansion. Image Analysis. SCIA 2003. Bigun J, Gustavsson T (ed): Springer, Berlin, Heidelberg; 2003. 2749:363-370.
11. Simonyan K, Zisserman A: Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems. 2014, 27:1-9.
12. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I: Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning. 2023, 202:28492-28518.
13. Wolf T, Debut L, Sanh V, et al.: Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020, 38-45. 10.18653/v1/2020.emnlp-demos.6
14. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, 770-778. 10.1109/CVPR.2016.90
15. McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, Battenberg E, Nieto O: librosa: Audio and music signal analysis in Python. Proceedings of the 14th Python in Science Conference. 2015, 18-25. 10.25080/Majora-7b98e3ed-003
16. Vaswani A, Shazeer N, Parmar N, et al.: Attention is all you need. Advances in Neural Information Processing Systems. 2017, 30:
17. Livingstone SR, Russo FA: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE. 2018, 13:e0196391. 10.1371/journal.pone.0196391
18. Davison AK, Lansley C, Costen N, Tan K, Yap MH: SAMM: A spontaneous micro-facial movement dataset. IEEE Transactions on Affective Computing. 2018, 9:116-129. 10.1109/taffc.2016.2573832
19. Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R: CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing. 2014, 5:377-390. 10.1109/taffc.2014.2336244

20. van der Maaten L, Hinton G: Visualizing data using t-SNE. Journal of Machine Learning Research. 2008, 9:2579-2605.

21. Ko BC: A brief review of facial emotion recognition based on visual information. Sensors. 2018, 18:401. 10.3390/s18020401

22. Poria S, Cambria E, Bajpai R, Hussain A: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion. 2017, 37:98-125. 10.1016/j.inffus.2017.02.003

23. European Parliament and Council: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Union. 2016, 1-88.

How to cite this article:
Shukla A, Agarwal R (August 25, 2026) Progressive Evolution of Emotion Detection: From Unimodal Baselines to a Quad-Modal Dynamic Fusion Architecture. Cureus J Comput Sci 3 : es44389-026-00246-0. DOI https://doi.org/10.7759/s44389-026-00246-0