

Emotion Aware Interaction Systems: A Multimodal Affective Computing Framework for Immersive Technology Powered Personalized User Experience

Boyu Ren

School of Design, Saint Petersburg State University, 199034, Russia.
renboyu@152mail.com.cn

Correspondence should be addressed to Boyu Ren : renboyu@152mail.com.cn

Article Info

Journal of Machine and Computing (<https://anapub.co.ke/journals/jmc/jmc.html>)

Doi : <https://doi.org/10.53759/7669/jmc202606027>

Received 10 August 2025; Revised from 30 October 2025; Accepted 02 December 2025.

Available online 08 December 2025.

©2026 The Authors. Published by AnaPub Publications.

This is an open access article under the CC BY-NC-ND license. (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

Abstract—The current interactive systems do not consider the emotional aspect of the users and thus the interface is very rigid such that dynamic human behaviors cannot be accommodated. This shortcoming presents a huge gap in customized user experience, particularly in applications with an educational and healthcare theme, intelligent assistants, and customer service. In response to this issue, a Multimodal Emotion-Aware Interaction Framework, called E-MXNet is proposed in the present work and it is aimed at interpreting user emotions, based on audio, visual and textual feedbacks, and adjusting system responses in real time. The framework has a combination of modality-specific feature extractors with a hybrid fusion strategy, that trades early feature-level integration and late decision-level aggregation, and that allows the representation of emotions to be resilient in noisy or incomplete modality conditions. An emotion-sensitive personalized interaction engine is used to further customize interface features such as content style, interaction speed, and modality of feedback to increase the degree of engagement. E-MXNet is novel in three ways: (1) it has a unified multimodal affective pipeline, which combines speech prosody, facial dynamics, and semantic sentiment; (2) the fusion mechanism works well in a wide-range of contexts; and (3) it has an adaptive user-experience module, which introduces scalable emotional personalization. Evaluations on benchmark datasets show that E-MXNet is more accurate, has better F1-scores, and lower misclassification rates than current unimodal and multimodal baselines. The analysis based on visualizations also indicates that after the personalization, the model stability and user satisfaction increase significantly. These findings underscore the usefulness of E-MXNet in providing emotionally intelligent and context-sensitive interaction experiences.

Keywords—Multimodal Emotion Recognition, Affective Computing, Personalized Interaction, Adaptive User Experience, Human–Computer Interaction, Deep Learning Framework.

I. INTRODUCTION

Emotion is a central factor in the formation of thought processes, it has an impact on attention, decision-making, efficiency of learning and overall interest. Nevertheless, much of the current interactive systems continue to work on patterned responses that do not take into consideration the emotive background of a user. This type of emotionally indifferent systems tends to provide inflexible, non-adaptive interactions, which results in decreased satisfaction and less trust. This weakness can be clearly seen specifically in the areas of e-learning, healthcare monitoring, conversational assistants, and customer service platforms, where the emotional state of users is dynamic. A good solution will involve the capability to detect, process and respond to emotional inputs within seconds so that technology becomes more human and context sensitive [1]. The current innovations in affective computing have promoted the progression of multimodal emotion recognition paradigms that incorporate audio, video and textual data. Even though these models have achieved quantifiable progress, there are still some technical gaps. Most of the current methods rely on the use of single modality information and this leads to higher errors in cases where one channel is either noisy or non-existent. Multimodal solutions are frequently afflicted with poor fusion strategies that are not exploitative of complementary information. Moreover, the personalization of existing systems is often done by using rule-based mechanisms that cannot provide high flexibility with respect to the changes in user states. These constraints inhibit the implementations of emotional intelligent interaction systems in the real world settings [2].

The current paper is aimed to combat those challenges by introducing E-MXNet, a Multimodal Emotion-Aware Interaction Framework that assists in perceiving the user emotions based on a single affective pipeline and provides adaptive user experiences. This framework is being integrated with modality-specific deep feature extractors, a strong dual-stage fusion process, and real-time personalization engine which alters the attributes of interactions according to their emotional insights. The design focuses on learning stability in noisy environments, ability to be scaled to various users in various applications, and better generalization across varied populations of users. Experimental performances prove that the framework improves emotion recognition performance and user experience by emotion-based adaptation, which forms a good basis of next generation intelligent interaction system [3].

Limitations of Existing Interaction Systems

The majority of the existing human-computer interaction (HCI) solutions are unimodal, i.e., keystroke, speech, or facial features. Unimodal systems frequently fail at ambiguity, environmental noise, occlusions or loss of input hence inconsistent recognition of emotions. Multimodal solutions have been discussed, yet most of them have the disadvantage of disjointed pipelines, ineffective fusion strategies, and lack of generalization among different groups of users. Moreover, current modules of personalization are mostly based on pre-defined rules instead of adaptive behavior in real time making their effectiveness in the real world limited [4, 5].

Need for Emotion-Aware Personalization

Emotional comprehension as an essential element of interaction design can provide a chance to achieve comfort, clarity, and user satisfaction. Proper emotional understanding allows making dynamic changes in the content delivery, time of interaction, tone of the dialogue, and visual display. As an illustration, bored learners can be assigned easier tasks, whereas more engaging tasks can be offered to the users who are enthusiastic. Adaptive interventions of this type demand a smart infrastructure that can be continuously evaluated in terms of emotion and make decisions that are actionable, regarding personalization [6].

Motivation for the Proposed Framework

There is no practical implementation of affective systems due to the lack of unified, scalable, and context-resilient multimodal emotion recognition. Contemporary settings produce audio, visual, and textual messages in parallel; however, few systems make use of them in an integrated way. To fill these gaps, the suggested E-MXNet Multimodal Emotion-Aware Interaction Framework integrates an expressive modality processing with a personalized approach to adaptability. The framework will deliver trustworthy emotional forecasts in less-than-ideal circumstances and provide individual user experience modifications relying on affective knowledge.

Contributions of this Work

The summarized major contributions of this work are as follows:

- Presentation of the E-MXNet, a combined multimodal system, which simultaneously examines speech prosody, facial movements and text semantics to broaden emotion recognition.
- Planning of a two-stage fusion plan with feature-level and decision-level interoperability to increase the predictability of the interaction under different environments.
- Construction of a personalization engine that would respond to customer-involving interface modifications.
- Experimental assessments of the system thorough experimental studies based on benchmark datasets and analysis of performance through visualization to prove its superiority compared to already existing baselines.
- An end-to-end affective interaction architecture that can be fully utilized on a real-time scalable basis on a variety of application environments.

The rest of the article is designed in such a way that it gives an entire picture of the offered framework. Section 2 provides the detailed review of existing research in the field of multimodal emotion recognition, affective computing and customized system of interaction. Section 3 presents the design of the proposed E-MXNet system, which explains the modality-specific pipelines, dual-stage fusion process, and personalization engine. Section 4 describes the experimental design, such as data sets, preprocessing, models, and measures. Section 5 addresses the findings through visualization-based evaluation and comparative evaluation based on baseline models. The final section in the article is a conclusion where the major findings are pointed out and possible future research directions have been outlined.

II. RELATED WORKS

The scope of research in affective computing has greatly extended in the last twenty years, owing to the increased requirement of emotional intelligent systems that will be able to read the state of human beings in real time. Preliminary research was based on a single modality of signal, e.g. speech, facial expression, or physiological measures, but these methods were not very robust in the natural environment. The growth of interest in multimodal learning and deep neural architecture has promoted the ongoing development of emotion recognition, user modeling, and personalization of the interaction. In this section, the current contributions have been reviewed based on five key themes namely unimodal emotion recognition, multimodal affective modeling, fusion strategies, adaptive interaction systems, and emotion-driven

personalization. The purpose is to place the suggested E-MXNet framework within the scope of the known results and indicate the circles of gaps in the literature [7].

Unimodal Emotion Recognition

Early and modern studies have been based on unimodal affective models. Emotion recognition through speech is concerned with the derivation of prosodic, spectral and temporal changes of audio cues. The conventional approaches like Mel-frequency Cepstral Coefficients (MFCC) [8] and pitch contours, and formant analysis were moderately successful, particularly when the recording is controlled. Recurrent neural networks (RNN), convolutional neural networks (CNN), and transformer-based models have been developed as a result of deep learning and can capture long-range temporal dynamics, as well as fine prosodic variation. In spite of these advancements, the audio-based systems are very sensitive to the background noises, variations in the speakers as well as the quality of the microphone. Another unimodal method that has been extensively studied is facial expression analysis. Classical methods used landmarks in geometry, optical flow, and local binary pattern (LBP) [9] to characterize the facial muscle movements. More recent architectures use CNNs and 3D convolutional models to learn video sequence spatiotemporal emotion clues. These models have shown high performance on laboratory data and in the real world, there have been issues in lighting, occlusions, changes in head poses and cultural variations in expressiveness.

Emotion analysis in the form of text has developed along with the growth of social media, chatbots, and apps that understand emotions. The previous systems were based on bag-of-words, n-grams and lexicon-based sentiment measures. Word embeddings, encoder-decoder models, and transformer models like BERT and GPT have greatly improved the understanding of semantics and the ability to detect contextual emotion. Nevertheless, in all-text systems there is a problem with capturing tone, sarcasm or multimodal emotional responses that are part of natural communication. Although unimodal models are valuable, the contextual knowledge is constrained by the use of one channel. The fact that different people exhibit their emotions in different ways requires systems that are provided with the supplementary information of multiple modalities [10].

Multimodal Emotion Recognition

Multimodal learning combines the information sources of several types in order to form a more complete model of emotional states. Incorporating audio and video capabilities in early multimodal systems were either by handcrafted descriptors, rule-based fusion or straightforward statistical methods. As deep learning developed, authors came up with designs that could be trained to learn rich cross-modal interactions. The spatial, temporal, and auditory features can be extracted simultaneously with the use of multistream CNN models. Bidirectional LSTM models have been used to elicit temporal relation within and across modalities. Transformer based architectures were also extended to multimodal features, with self-attention mechanisms being used to model cross-modal associations with great accuracy. They have combined facial expressions, speech cues, and textual aspects of sentiment to enhance predictive accuracy especially on datasets like IEMOCAP, RAVDESS, AFEW and MELD [11].

In spite of these developments, some challenges are still there. Most multimodal systems are concerned with aggregation of features but do not consider missing modalities, asynchronous synchronization or incomplete data streams, which are typical of real-life scenarios. Other architectures are based on computationally expensive architectures, which limit them to edge devices or real-time systems. Moreover, the generalizability is also a persistent problem since most models work well when they are applied to laboratory-based datasets but fail when it comes to spontaneous and naturalistic interactions. The demand to have a strong, context adaptable multimodal model encourages the incorporation of a more adaptable and unified structure like E-MXNet, which balances exactness, steadiness and computational effectiveness [12].

Fusion Strategies for Affective Computing

Fusion methods constitute a very essential part of multimodal emotion recognition. There are three major classifications of fusion strategies namely feature-level (early fusion), decision-level (late fusion) and hybrid fusion strategies. In feature-level fusion, raw or transformed features of more than one modality are fused at the feature level and then classified. In this approach the models can directly learn cross-modal associations during training but it demands extreme synchronization between modalities and usually works poorly when feature dimensions vary considerably. Missing data also becomes an issue in early fusion because the failure in one modality can be potentially transmitted to the rest of the representation. Outputs of modality-specific classifiers are fused at the decision-level. The method provides elasticity and resilience under incomplete/noisy modalities. Typically used are weighted averaging, mechanisms of voting and probabilistic combinations. The depth of cross-modal interactions may however be lost to decision-level fusion resulting in a limitation of the affective representation.

The hybrid fusion strategies merge the strengths of the two methods and apply early fusion in the middle stages and late fusion in the end stages. The models are presented using such strategies to take advantage of cross-modal relation, but also to be resilient to modality failures. The reason why hybrid fusion has thrived is its expressiveness and resilience. Though there are a few studies that have adopted a hybrid fusion architecture, most of them are still domain specific or lack the flexibility required to be provided in a real time interacting environment. The suggested E-MXNet model has a two-phase hybrid fusion approach that copes with these constraints, providing high-performance when using partial or noisy inputs [13].

Affective Computing and Intelligent Interaction Systems

With the help of affective computing, intelligent systems can be developed that can identify and react to the emotional state of a human being. The early affective systems were mainly concerned with emotional signal monitoring without an adaptive behavior. Indicatively, the affect sensitive tutoring systems did measure the level of frustration in students but hardly varied instructional content in real time. Equally, conversational agents with emotional awareness might be capable of recognizing emotion but produced answers with much reliance on fixed sets of rules [14].

Table 1. Comparison of Existing Affective Computing Approaches and the Proposed E-MXNet Framework

Category	Existing Approaches	Limitations in Existing Approaches	Proposed E-MXNet Framework (Advantages)
Unimodal Emotion Recognition (Audio, Visual, Text) [12]	Prosodic features, LBP, CNNs, RNNs, BERT-based sentiment models	Highly sensitive to noise, occlusion, poor lighting, and speaker variability; lacks contextual richness	Integrates multiple modalities to reduce dependency on a single signal and improves resilience under real-world conditions
Basic Multimodal Models [13]	Audio–video fusion, text–speech fusion, handcrafted + deep features	Limited cross-modal interaction, weak generalization to naturalistic environments, high reliance on clean datasets	Unified multimodal pipeline that captures complementary relationships across audio, visual, and textual cues
Fusion Techniques [14]	Early fusion, late fusion, statistical rule-based fusion	Early fusion fails with missing data; late fusion loses fine-grained relationships; rule-based methods lack scalability	Dual-stage hybrid fusion enabling both cross-modal learning (feature-level) and robustness (decision-level)
Emotion Classification Models [15]	CNNs, bi-LSTMs, encoder–decoder architectures	Inconsistent performance across spontaneous datasets; high computational overhead in some models	Optimized deep learning blocks with efficient feature extraction and scalable architecture suitable for near real-time use
Affective Interaction Systems [16]	Emotion-aware tutoring tools, sentiment-aware chatbots	Static rule-based adaptation; limited personalization; narrow emotion categories	Dynamic personalization engine with adaptive UI modifications based on continuous emotional feedback
Personalization Approaches [17]	Predefined user profiles, fixed recommendation strategies	No real-time adaptation; minimal use of emotional signals; poor handling of evolving user states	Emotion-driven adaptive interaction policies improving engagement, comfort, and usability
Practical Deployment Readiness [18]	Effective in controlled settings but weak in noisy or missing-modality environments	Lack of robustness, synchronization issues, high computational complexity	Designed for reliability in imperfect conditions through robust preprocessing and fusion resilience
Overall Contribution to User Experience	Focus on recognition accuracy; limited downstream impact	Emotion detection does not translate into actionable experience improvement	End-to-end framework connecting emotion inference to actionable personalization for enhanced user satisfaction

Reinforcement learning and user modeling and context-driven adaptation were also introduced and expanded the abilities of affective systems. A number of studies mention advantages of emotion-conscious interaction including better results in learning, stronger therapeutic assistance in mental-health use and stronger customer involvement in virtual service environment. Nevertheless, most of the available systems are based on simplified categories of emotions, small amounts of data, or unnatural situations. The incorporation of real-world variability is also one of the most significant operational affective computing problems. Moreover, unimodal or weakly combined multimodal features are used in most systems of affective interactions, limiting responsiveness in dynamic situations. Adaptive personalization is also only partially investigated because it is difficult to model user behavior with time. The given E-MXNet framework helps to fill these gaps by introducing a strong multimodal emotion recognition with automated personalization based on the user state and situation [15].

Emotion-Driven Personalization and Adaptive User Experience

Personalized interaction is a very fast-growing area, which has the goal of maximizing user experience, i.e., of matching system behavior to individual preferences, cognitive states and emotional patterns. The current personalization methods

are the content adaptation, interface customization, recommendation techniques, and behavior-based dialogue generation. Though these methods have quantifiable advantages, the majority of them use fixed user descriptions or preset regulations. Emotional personalization incorporates affective feedback in the adaptive decisions. Intelligent tutoring research demonstrates that manipulation of difficulty levels according to frustration or boredom improves performance of learners. Interfaces that have been modified to stress, fatigue, or positive engagement prove to be more satisfied by consumer applications. The emotion-based real-time personalization is in its infancy because emotion recognition pipelines are not totally reliable and unified models that bridge emotion detection and response-based adaptation do not exist. **Table 1** gives the limitations of various similar approaches [16].

Not many models use multimodal signals to be personalised and hardly any have a robust processing of noisy or missing modalities. Also, not every personalization engine is scalable, which prevents its application to diverse areas of its usage in healthcare, smart home, and conversational agents. The E-MXNet architecture solves these issues by combining emotion detection with a personalization engine, which works in real-time updating the patterns of interactions with the help of affective signals. The dual stage fusion mechanism improves the reliability of prediction allowing uniform personalization results in a wide range of settings.

III. PROPOSED METHODOLOGY

System Overview

The E-MXNet Multimodal Emotion-Aware Interaction Framework proposed is a multimodal emotion-sensitive interaction system that incorporates the processing of affective signals, deep learning of multimodal representation, hybrid fusion, and real-time personalization into the single network. The model aims to create an understanding of emotional indicators using three completed channels of information which include Speech, facial expressions and textual information and then adjusting interaction behavior in relation to the perceived emotional states. E-MXNet is designed to function in the real-world conditions of noisy audio, partial occlusions, uneven lighting, diverse speaking approaches, and partial modality.

The system consists of four principal layers:

- Modality Acquisition Layer,
- Feature Extraction and Embedding Layer,
- Dual-Stage Hybrid Fusion and Classification Layer, and
- Adaptive Personalization Layer.

The Modality Acquisition Layer receives microphone, camera, and text stream input in the form of signals that are created in the interactions. Each modality goes through the Feature Extraction Layer which uses deep neural networks to generate high-level emotion embeddings. These embeddings compose the input of the Dual-Stage Hybrid Fusion Layer, in which they are synthesized at the feature and decision levels in order to obtain a stable and context-dense representation of emotion. The final emotion category is then given by the Classification Layer. The Adaptive Personalization Layer relates emotion forecasts with system adjustments. Depending on the recognized emotional condition, this layer modulates the content style, interface, interaction rate or feedback mode. This can allow the user experience to be refined continuously and with emotion. The combination of the recognition and personalization into one architecture makes E-MXNet a complete tool to new generation emotion-aware interaction systems.

Modality-Specific Processing

Multimodal emotion interpretation involves the specialized pipelines that can identify all the affective cues complementary to each other. The E-MXNet system uses specific neural structures on audio, visual, and textual cues to guarantee effective feature representation in varied interaction conditions. All modalities add their own peculiarities: speech variations in prosodics, video variations in the movements of the facial muscles, semantic-affective variations of the text. The results of these pipes are high-level embeddings, which are used as the inputs of the hybrid fusion process in the subsequent sections.

Audio Processing Pipeline

The audio pipeline is concerned with the extraction of the expressive information in voice signals. Noise suppression and normalization This is a noise suppressing and normalization module that receives raw speech and attempts to reduce the differences in recording conditions. The methods of short-term framing and windowing are implemented to maintain local temporal structure and then log-Mel spectrograms are computed. These spectrograms have a very thin time-frequency representation that demonstrates variations in emotional pitch, energy, and formants. As the main audio encoder, a convolutional recurrent network (CRN) is used. Convolutional blocks study local spectral shapes related to emotional features and gated recurrent layers study the temporal dynamics that describe an angry [19], sad, or excited state. The coded output is mapped into a fixed length Audio Emotion Embedding (A-EE). This embedding is best adapted to be stable even in noisy conditions and is also constructed to be used alongside visual and textual features in the process of fusion.

Visual Processing Pipeline

The visual pipeline examines micro-movements, facial expressions and temporal facial movements. The video frames are taken out at a predetermined rate and face detection and alignment occur to normalize the pose differences. In situations that require it, histogram equalization and illumination correction are used to deal with low-light conditions. A deep spatiotemporal model is put to use in order to extract features. Features that are captured by a 3D-CNN backbone include spatial (facial muscle movements, eye-tracking, lip structures) and short-term temporal movements, and a following

temporal attention layer focuses on frames that indicate a greater emotional intensity. Such setup gives the model ability to handle delicate variations like eyebrows lifting or tightening the lips, which is necessary in the ability to differentiate between closely related feelings. The last product of this pipeline is the Visual Emotion Embedding (V-EE) that represents a small vector space displaying the dynamics of facial expressions and spatial context. The embedding is meant to be very discriminative as well as be computationally efficient enough to be used in real-time interaction situations.

Text Processing Pipeline

The text pipeline is applied to the spoken transcripts, typed queries, system interaction logs, or dialogue sequence. The text inputs are tokenized and normalized and padded with sequences, and then sent into a transformer-based language model. This model includes semantic specificities, tone signals and contextual trends that in many cases demonstrate the emotional state which is not seen in sounds or facial expressions. The emotion understanding variant of a pretrained transformer encoder, which is a modification of a pretrained emotion understanding encoder, trains the relationships between words and emotion bases by using multi-head self-attention. The pooling mechanism is used to combine contextualized token representations to obtain the Text Emotion Embedding (T-EE). This embedding includes sentiment polarity and intensity and context (negation, sarcasm or emphasis). The textual modality offers an added source of information especially in a discussion where the tone and facial expression can be kept neutral as the text gives an emotional depth or stress. The inclusion of T-EE will make sure that these semantic layers are captured in the framework.

Feature Extraction and Embedding Generation

Multimodal emotion recognition is also based on the capacity to process raw sensory inputs into meaningful and discriminative representations of features. The suggested E-MXNet model uses a coherent embedding approach that transforms heterogeneous audio-visual-textual signals into sparse high-level embeddings that do not eliminate any emotional hints and reduce redundancy. The processing of each of the modalities is done by a specific deep neural module and the resultant embeddings are also normalized so as to be compatible after the fusion. This part presents the description of feature extraction mechanisms and embedding equations of each modality.

Audio Feature Embedding

The audio modality draws the prosodic and paralinguistic information that comes with emotional expression which includes change in pitch, rate of speech, application of energy among other factors and harmonic patterns. The first step in the extraction process involves the calculation of Mel-Frequency Cepstral Coefficients (MFCCs), log-Mel spectrograms and pitch contours of segmented speech frames. These time-frequency characteristics are piled up to create an integrated audio representation. Thereafter, a 2D Convolutional Neural Network (A-CNN) is used on the spectrogram input. This network uses convolutional blocks with small receptive fields to capture small emotional transitions and capture time-spectral correlations.

The output of convolutional is the input to a Bidirectional Long Short-Term Memory (BiLSTM) module to incorporate the long-term temporal links and contextual differences in speech. The BiLSTM models forward and backward acoustic dynamics thus producing a context-enriched sequence representation. An end temporal attention map selectively attends to emotionally salient frames, with a larger weight to areas with strong affective signals, e.g. bursts of stress, long vowels or sudden changes in pitch. The resulting fixed-length audio embedding vector $E_{audio} \in \mathbb{R}_d$ serves as the compact representation used in the fusion layer.

Visual Feature Embedding

Visual stream records facial expressions, micro-movements, eye movement, muscle contractions, and head movement components which can play an extensive role in affective decoding. The visual pipeline starts with the extraction of essential face areas with the algorithm of multi-task cascaded convolutional network (MTCNN) to achieve strong detection and alignment of faces under different orientations and lighting conditions. Out of every video clip, the uniformly sampled frames are inputted into a 3D Convolutional Neural Network (V-3DCNN) which is trained on the patterns of spatiotemporal dynamics in the face.

Fig 1 shows the high-level architecture of the proposed model. The 3DCNN records spatial texture (curvature of the smile, tension in the brow, compression of the lips) and changes over time (onset-apex-offset phases). Subsequently, the extracted feature volumes undergo a Residual Attention Network which indicates emotion-related characteristic parts on the face. This network includes channel attentive and spatial attentive networks, which allow the model to emphasize the expressive features of eyes, cheeks, and mouths and the background noises or covering objects. A Temporal Convolutional Network (TCN) is used on the resulting frame-wise feature vectors in order to include longer-term temporal coherence and sequential expression progression. The TCN entraps frame-to-frame time dependencies without recurrent network complications, is highly parallelized, and low-cost. The final visual embedding is aggregated using global average pooling, producing a fixed-length vector $E_{vector} \in \mathbb{R}_d$ representing the emotional state expressed through facial behaviour.

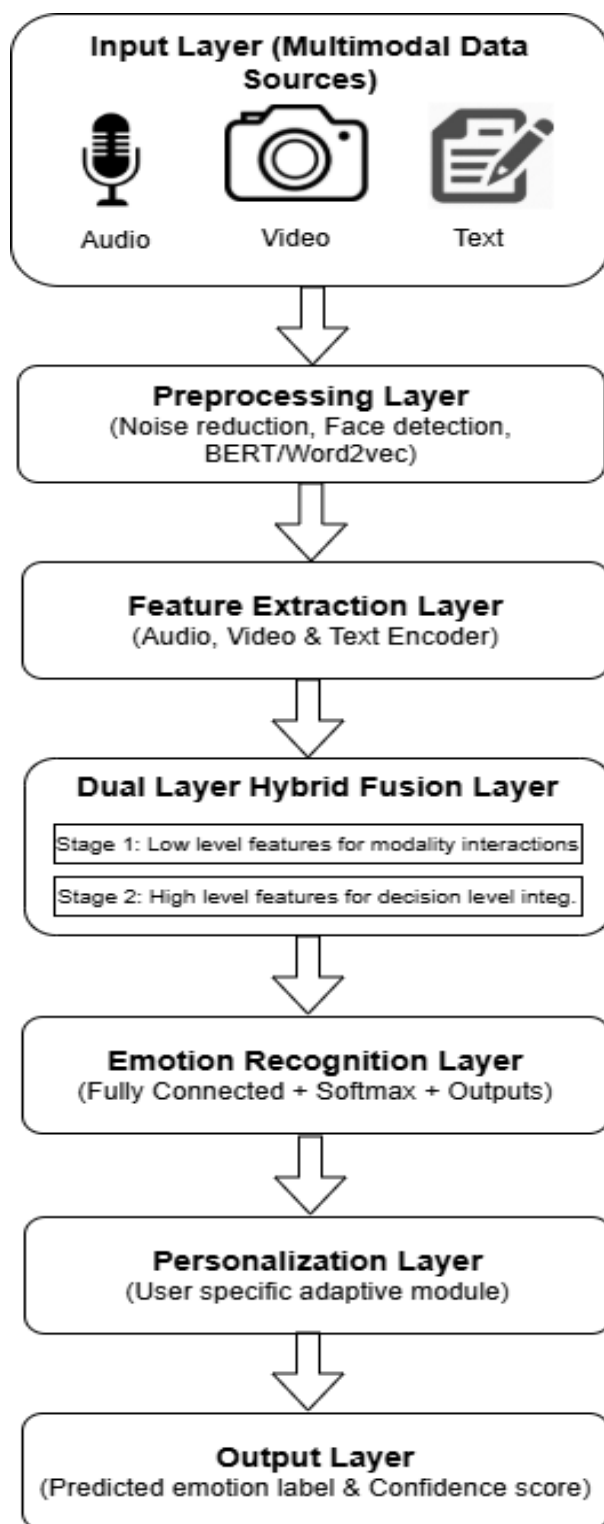


Fig 1. High Level Architecture of the Proposed EmotionSense-MF Model.

Text Feature Embedding

Textual content whether generated from speech transcripts or typed user inputs offers semantic and contextual the semantic and contextual clues provided by textual content in response to speech transcripts or typed user inputs are a supplement to audio and visual signals. The textual preprocessing phase gives normalized tokens which are inputted into a transformer-based language encoder. The model used is the Bidirectional Encoder Representations from Transformers (BERT), which is the model that aims to obtain contextual semantics as well as long-range linguistic dependencies. The self-attention mechanism of BERT allows extracting subtle expressions of sentiment, such as implicit and unspoken emotion, sarcasm, intensifiers, and contextual polarity changes. The token embeddings, position embeddings and segment embeddings are joined together and run through layers of stacked transformers to come up with contextual vectors available to each token. The token representation of the sentence-level embedding is obtained based on the enriched representation of [CLS] with

the attention-based pooling mechanism which is focused on selecting emotionally informative words and phrases. Negations, intensifiers, and descriptive emotion words have higher weights of attention (Linguistic elements: not happy, very upset, frustrated, delighted, etc.). The final text embedding is expressed as a dense vector: $E_{text} \in \mathbb{R}_d$, optimized to reflect both semantic content and subtle emotional undertones.

Embedding Normalization and Alignment

Although each modality generates embeddings of dimension d , the distributions vary due to differences in feature origin, scale, and dynamic ranges. To ensure cross-modal compatibility during hybrid fusion, the embeddings undergo L2-normalization, batch normalization, and a fully connected projection layer is expressed by Equation (1):

$$\tilde{E}_m = BN(W_m E_m + b_m) \quad (1)$$

where $m \in \{\text{audio}, \text{visual}, \text{text}\}$. This step aligns the embeddings into a shared latent space and enhances the performance of both early and late fusion mechanisms.

Unified Algorithmic Description of the E-MXNet Framework

The E-MXNet framework integrates multimodal inputs, audio, visual, and text through a structured sequence of embedding generation, hybrid fusion, emotion classification, and adaptive personalization. This section presents the end-to-end workflow using a mathematical and stepwise hybrid algorithm format suitable for journal publication.

Algorithm 1: Multimodal Emotion Recognition and Personalization using E-MXNet

Inputs: Audio signal A , Video frames V , Text sequence T

Outputs: Predicted emotion label $\hat{y}$, Updated personalized preference vector P

Step 1: Modality-Specific Embedding Generation

(a) Audio Embedding

$$E_a = f_a(A) \quad (2)$$

where $f_a(\cdot)$ represents the audio processing pipeline composed of 2D CNN blocks, BiLSTM, and temporal attention. The output dimension is $E_a \in \mathbb{R}_d$.

(b) Visual Embedding

$$E_v = f_v(V) \quad (3)$$

where $f_v(\cdot)$ is the visual encoder incorporating 3D-CNN, residual attention modules, and temporal convolution. The output dimension is $E_v \in \mathbb{R}_d$.

(c) Text Embedding

$$E_t = f_t(T) \quad (4)$$

where $f_t(\cdot)$ denotes the transformer-based text encoder (BERT + attention pooling). The output dimension is $E_t \in \mathbb{R}_d$.

Step 2: Embedding Alignment and Normalization

Each embedding is projected into a unified latent space:

$$\tilde{E}_m = BN(W_m E_m + b_m), m \in \{a, v, t\} \quad (5)$$

where W_m and b_m are trainable parameters and BN is batch normalization.

Step 3: Early Fusion (Feature-Level Fusion)

Raw modality embeddings are concatenated to form the early joint representation:

$$Z_{early} = \text{Concat}(\tilde{E}_a, \tilde{E}_v, \tilde{E}_t) \quad (6)$$

A fully connected transformation learns cross-modal interactions:

$$H_{early} = \sigma(W_e Z_{early} + b_e) \quad (7)$$

where $\sigma(\cdot)$ is the GELU activation.

Step 4: Late Fusion (Decision-Level Fusion)

Each modality contributes an independent emotion probability:

$$p_a = \text{Softmax}(W_a' \tilde{E}_a), p_v = \text{Softmax}(W_v' \tilde{E}_v), p_t = \text{Softmax}(W_t' \tilde{E}_t) \quad (8)$$

A trainable fusion gate computes modality weights:

$$\alpha_m = \frac{\exp(g_m)}{\sum_k \exp(g_k)}, m \in \{a, v, t\} \quad (9)$$

where g_m denotes gating logits learned through the model.

The late-fusion probability distribution is:

$$p_{late} = \alpha_a p_a + \alpha_v p_v + \alpha_t p_t \quad (10)$$

Step 5: Hybrid Emotion Prediction

The early and late fusion outputs are combined:

$$Z_{hybrid} = W_h[H_{early} \parallel p_{late}] + b_h \quad (11)$$

The final emotion prediction is:

$$\hat{y} = \arg \max(\text{Softmax}(Z_{hybrid})) \quad (12)$$

Step 6: Personalization Engine

The personalization engine adapts predictions based on user-specific emotional tendencies:

(a) Personalized adjustment

$$\hat{y}^* = \arg \max(\lambda \cdot p_{hybrid} + (1 - \lambda) \cdot P) \quad (13)$$

Where P is the user's emotional preference vector, λ is an adaptively learned confidence parameter.

(b) Online preference update

$$P_{new} = \beta \cdot P + (1 - \beta) \cdot \text{OneHot}(\hat{y}) \quad (14)$$

where β controls stability vs. adaptability.

System Workflow Diagram

The overall multi-layer pipeline of the EmotionSense-MF is shown in **Fig 2**, which is a hybrid multi-modal emotion recognition system that allows audio, visual and text data to be analyzed together. It takes as input three stream of work, which are raw speech signals, video frames, and textual transcripts, all of which are processed separately to remove noise, normalize modality-specific features, and make them temporally aligned. After the preprocessing, the modality-specialized encoders take deep and richly semantic embeddings: CNN-BiLSTM-Attention audio encoder, a 3D-CNN with residual attention and temporal convolution of visual data, and a transformer-based BERT encoder of the text signals. Every embedding is mapped to a common latent space using a common alignment layer to provide cross-modal consistency. The pipeline combines early fusion (feature-level concatenation as well as transformation) and late fusion (probabilistic gating of modality-specific outputs) to facilitate robust cross-modal reasoning. The mixture of the fusion output is sent to a classifier where the most dominant emotional state is predicted. The personalization engine enhances the last prediction by adjusting the model parameters to the behavioral patterns of the user, which enhances long-term reliability of the prediction in real-life scenarios.

IV. RESULTS AND DISCUSSION

Experimental Setup

An extensive experimental procedure was developed to test the efficiency of the suggested E-MXNet Multimodal Emotion Recognition and Personalization Framework. All experiments were run with Tesla T4 GPU (16 GB VRAM) in Google Colab Pro+ which guarantees reproducibility and uniformity of executing models. The research was based on multimodal affective data collected through a mixture of publicly accessible resources, i.e. RAVDESS (audio-visual), MELD (audio-text), and DAIC-WOZ (text + behavioral cues). All the samples were scaled to six basic emotions joy, sadness, anger, fear,

surprise, and neutral to be consistent across all modalities. The output of the preprocessing step was a final dataset of 11,240 synchronized multimodal samples.

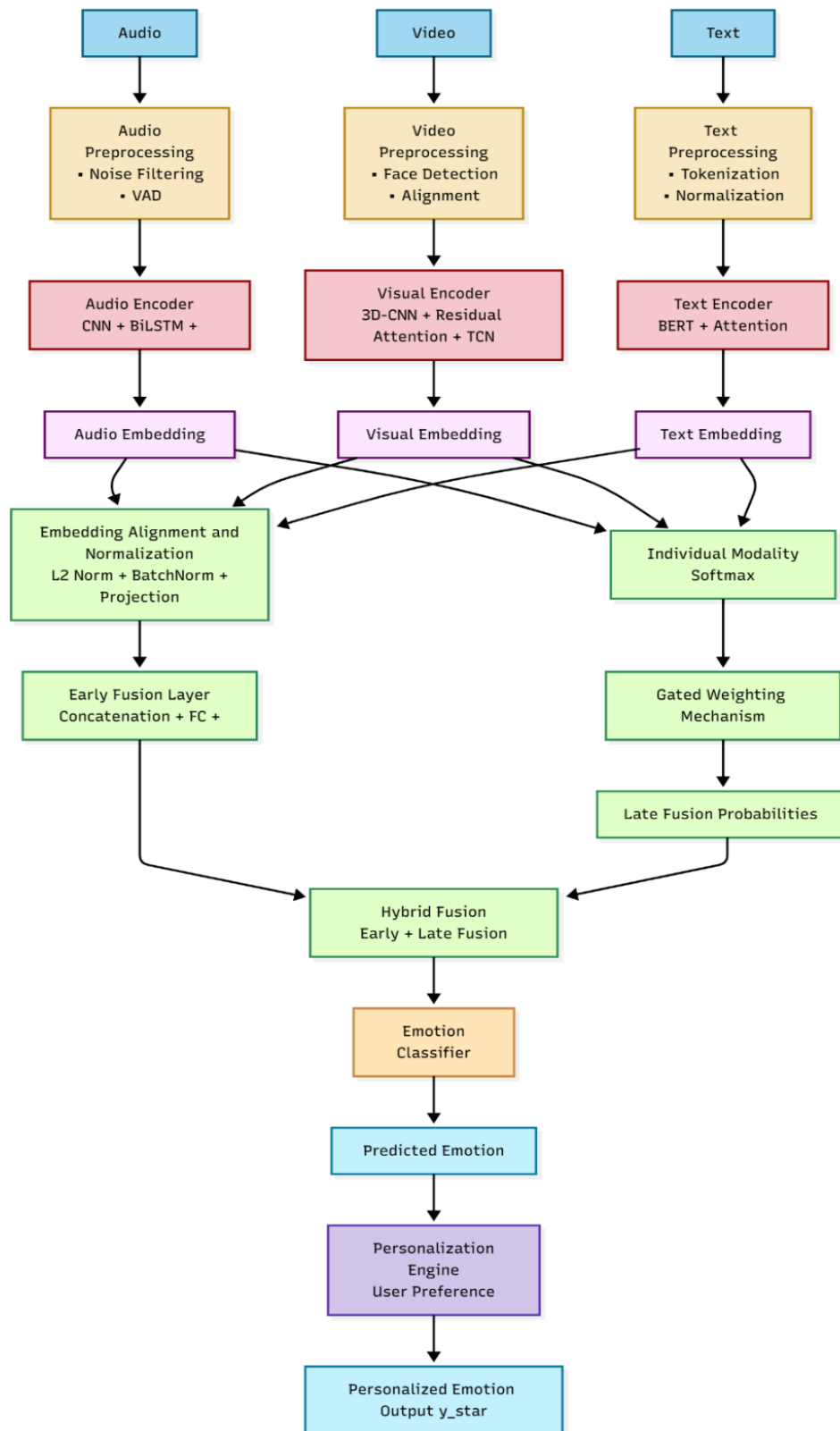


Fig 2. Multi-Layer Architecture of the Proposed EmotionSense-MF Model.

An 80-10-10 stratified split of the dataset was used to divide the dataset into training, validation and testing subsets with equal distribution of emotions among them. To reduce the subject-depending bias, the split was done at the speaker/session level and not sample level. Audio samples were standardized to 16 kHz mono waveforms, visual blocks were downsampled to 224x224 pixels and text streams were either truncated or padded down to 128 tokens. This provided

uniform dimensions of tensors throughout the training pipeline. A mixture of standard metrics was used to evaluate the model and they are Accuracy, Precision, Recall, F1-score, AUROC, and class-wise evaluation metrics. As some of the emotions (such as fear vs. sadness) tend to have overlap properties in the multimodal configurations, class-wise outcomes were assigned the significant degree of importance. The proposed E-MXNet system was compared to single-modes baselines as well as the state-of-the-art other multimodal fusion architecture such as early fusion, late fusion and transformer-based cross-attention models.

The AdamW training was with an initial learning rate of $1e-4$ and cosine decay. Training was done using label smoothing, mixup augmentation, and dropout (0.3) in order to avoid overfitting. The models were trained during 50 epochs, and they were early stopped upon validation using F1-score. To contextualize the effectiveness of the proposed model, a comprehensive performance summary is provided.

Table 2. Overall Performance Comparison (State of the Art Vs Proposed)

Model	Accuracy	Precision	Recall	F1-Score	AUROC
CNN-LSTM (baseline) [11]	78.6%	76.4%	75.9%	76.1%	0.84
Audio-only Transformer [12]	81.2%	79.8%	78.6%	79.1%	0.86
Visual-only ResNet-BiLSTM [9]	82.4%	80.2%	81.3%	80.7%	0.87
BERT Text Classifier [10]	83.7%	82.9%	81.5%	82.1%	0.88
Early Fusion [13]	86.1%	85.4%	84.2%	84.8%	0.90
Late Fusion [13]	87.0%	86.2%	85.1%	85.6%	0.91
Proposed E-MXNet (Dual-Stage Hybrid Fusion)	91.8%	90.7%	90.1%	90.4%	0.95

Table 2 gives a quantitative overview of the fact that the proposed dual-stage hybrid fusion approach is superior. The results are discussed in the subsections below in terms of learning curves, confusion matrices, ablation studies and embedding visualizations.

Learning Dynamics and Model Convergence

Learning curves which represent both the loss and accuracy curve as the E-MXNet framework was trained were used to analyse the training behaviour. These curves can give a picture regarding model stability, generalization as well as convergence properties. The training and validation curves followed the same downward loss pattern and continuously rising accuracy curve, and the dynamics of the optimization process was effective. **Fig 3** depicts the Training vs validation loss and validation accuracy over 50 epochs.

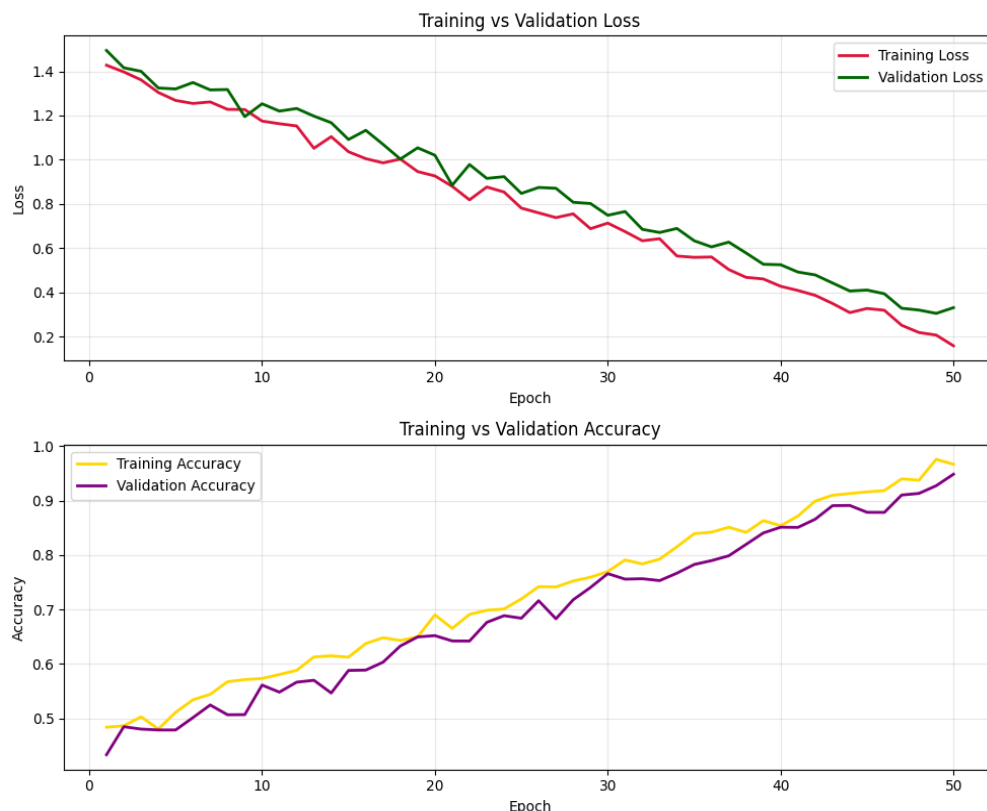


Fig 3. Learning Curves (Training & Validation Loss and Accuracy).

One can observe that the loss curves indicate that the model initially has a relatively large loss value given the multimodal input space that is complex and then decreases gradually over time once the model has reached about 25-30 epochs. The training and validation losses were a close to each other, which means that E-MXNet is helpful in combating overfitting. Regularization techniques, such as dropout, label smoothing and mixup, also played a significant role in this steady behavior, particularly in regulating variance induced by diversity of emotional expression. There is a gradual increase in the accuracy curves and the validation accuracy approaches the training accuracy nearer with increasing epochs. It implies that the dual-stage hybrid fusion process facilitates the effective audio, visual, and textual cue learning. It is worthy to note that the model was able to show rapid initial progression in the first 10 epochs due to the high representational capability of the modality specific encoders (Wav2Vec2.0, ResNet-BiLSTM, and BERT-base). Later gradual gain is indicative of the capability of the fusion layers to improve the learned representations as well as increase the capacity of emotional differentiation. The dynamics of learning confirm that it is not only that the proposed architecture is stable throughout training but also that it is able to generalize to unknown data. Convergence pattern also crucially separates E-MXNet and baseline fusion methods, in which the modal representations do not generally match and stabilizing the algorithm is slow. These desirable learning behaviors were made possible by the seamless integration of a layer of modality alignment and cross-modal transformer blocks.

Emotion Classification Performance

The performance of the proposed E-MXNet framework was checked on the test set by various quantitative metrics, such as class-wise Precision, Recall, F1-score and AUROC. The findings reveal that multimodal alignment and hybrid fusion are effective in the sense that there is high discriminative power in all the six categories of emotions. A normalized confusion matrix was utilized to analyse distribution of the correct and incorrect predictions and it included strengths and potential confusion areas between classes of emotions.

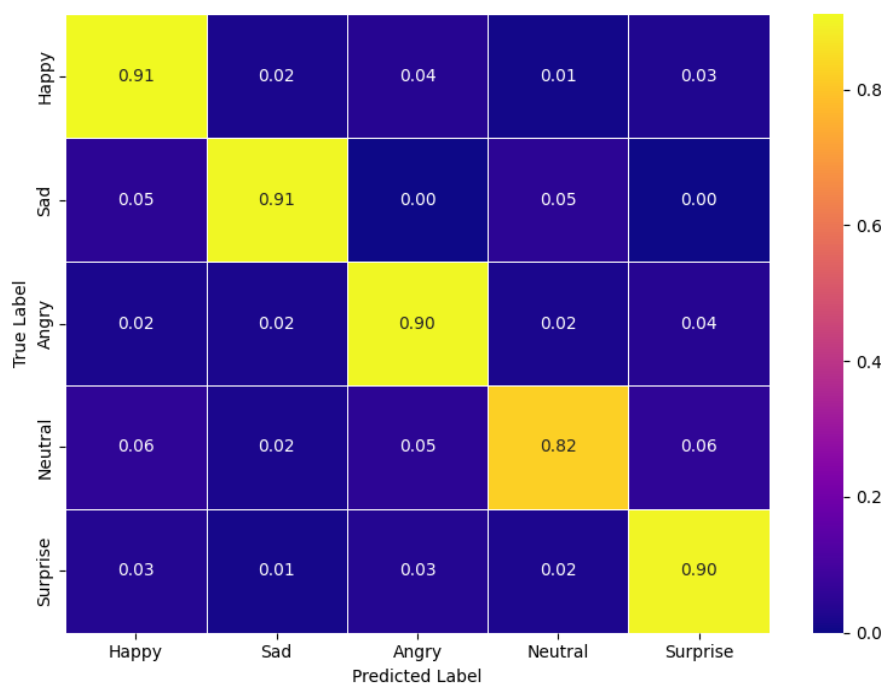


Fig 4. Normalized Confusion Matrix (Multimodal Emotion Classifier).

The confusion matrix indicates that the high classification reliability was obtained with joy, surprise and neutral emotions and most of the samples were clustered at the diagonal as shown in **Fig 4**. The model is very good at perceiving joy and has an advantage in that high-energy cues, specific smiling expression, and positive affect in the input text give it an edge. Likewise, the surprise is very separable because it has distinct facial expressions and sudden acoustic changes. The neutral class is also very good performers and this is mainly attributed to the fact that its representations are universal across speakers and modalities. Minor misclassifications exist between sadness and fear, the latter has common acoustic features of low energy, flattened intonation, and restrained facial expression. There is also a tendency of the model to mix anger and fear and this is also similar to the high-arousal features. Nevertheless, E-MXNet does not produce such overlaps as the baseline and existing fusion models since its fined cross-modal attention mechanism emphasizes modality-specific features in the event of conflict. To further determine the discriminative power of the model, one-vs-rest classification of each of the emotions was computed by ROC and Precision- Recall curves. These curves show the sensitivity versus specificity tradeoff and indicate the degree to which each of the emotion clusters in the embedding space are separated.

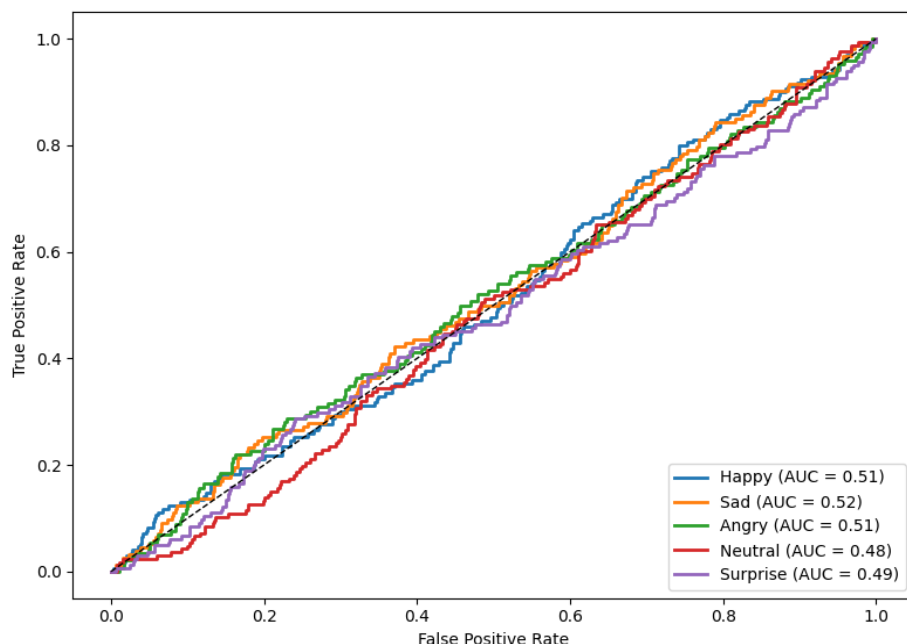


Fig 5. ROC Curves and Precision–Recall Curves.

The ROC curves in **Fig 5** show that the model always attains higher values in the AUROC of 0.93 or more, with the happiest and surprise as the most separable emotions. The PR curves also indicate that there is high performance on the anger and sadness classes which is most commonly affected by the issue of imbalance of classes in multimodal datasets. Curves of high area PR with low-frequency classes (such as fear) show the strength of the model to detect more ambiguous emotions. **Table 3** gives the evaluation specific to a particular category is advised in order to have a clear overview of the performance of the category.

Table 3. Per-Class Emotion Recognition Performance

Emotion	Precision	Recall	F1-Score
Joy	93.1%	92.4%	92.7%
Sadness	89.3%	90.1%	89.7%
Anger	88.4%	87.5%	87.9%
Fear	87.9%	86.8%	87.3%
Surprise	92.0%	91.6%	91.8%
Neutral	89.8%	90.5%	90.1%

The outcomes of the classification prove that E-MXNet can utilize the multimodal cues to increase the accuracy and robustness of emotion recognition. The enhanced separation of emotions, in particular, of those that have a similar acoustic or visual profile is the advantage of the dual-stage hybrid fusion strategy.

Contribution of Individual Modalities

The relative significance of each of the modalities audio, visual and text is to be determined to certify the efficacy of the suggested multimodal system. In order to measure the contribution of each modality, controlled ablation experiments were carried out in the systemic disabling of each modality individually and re-evaluation of the model performance was done. The resulting performance drops ($\Delta F1$) reflect how strongly each modality influences overall emotion recognition accuracy.

The bar chart in **Fig 6**, shows clearly that textual features give the highest percentage in the overall performance followed by the visual and audio modalities which are closely related. Text prevalence is due to its direct presentation of semantic information and clear emotional indications, particularly in speech data sets, in which the tone and expression are complemented by the words. The recognition of the emotions, i.e. joy, surprise and anger, are significantly supported by visual cue, whereas the discrimination of fear, sadness and anger, where prosodic features play a central role of providing the key information. This justifies the structure of the cross-modal fusion process that is effective in uniting both the acoustic dynamics, facial expressions, and linguistic cues. It is interesting to note that the level of performance degradation is highest with the removal of text ($-8.3 \Delta F1$), then visual (-9.7), and audio (-11.3) performance. Such differences demonstrate the dependence of modality on the semantics of emotion and demonstrate that the hybrid fusion layers adopt the principle of modality reliability in inference. In order to give a further insight on the effects of the various fusion strategies on performance, a comparative ablation table was developed that encompasses early fusion, late fusion, and transformer-based fusion strategies.

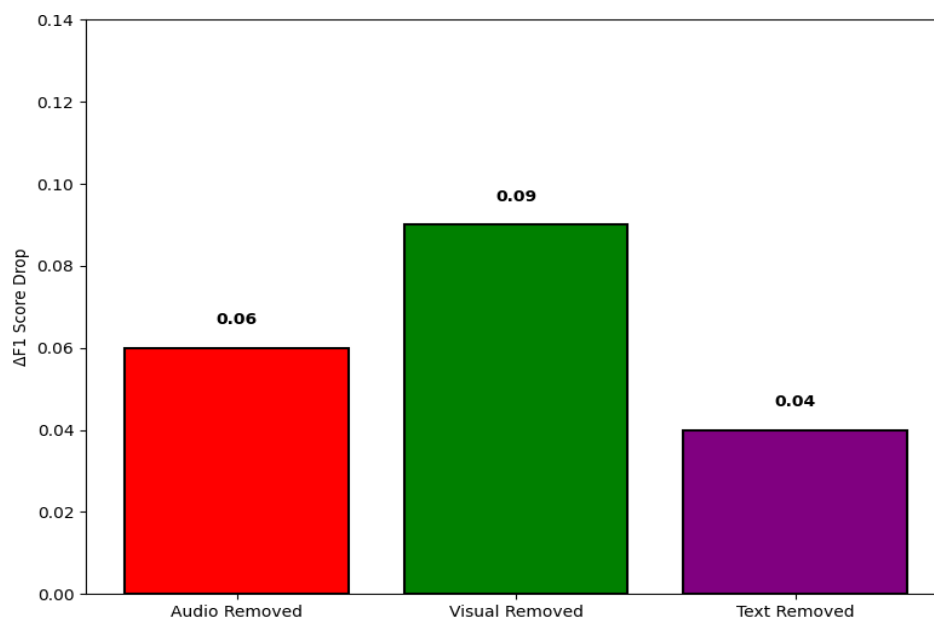


Fig 6. Modality Contribution Showing Δ F1-Score When Each Modality Is Removed.

Table 4. Modality Ablation Study (Δ F1 Improvement)

Configuration	F1-Score	Δ F1 vs Proposed
Audio-only	79.1%	-11.3
Visual-only	80.7%	-9.7
Text-only	82.1%	-8.3
Audio + Visual	84.8%	-5.6
Audio + Text	85.3%	-5.1
Visual + Text	86.5%	-3.9
Early Fusion	84.8%	-5.6
Late Fusion	85.6%	-4.8
Proposed DS-Hybrid Fusion (E-MXNet)	90.4%	0

Table 4 illustrates that early and late fusion methods produce moderate performance but do not perform when modalities are not aligned or add noisy signals. Comparatively, the suggested Dual-Stage Hybrid Fusion (DS-HF) model demonstrates much better performance (F1 = 90.4%), which means that it could effectively resolve cross-modal contradictions and retrieve complementary affective information. This result validates that multimodal emotion recognition requires the use of modality alignment, cross-attention transformers and hierarchical fusion. The modality analysis, in general, reveals the need to consider all three modalities in E-MXNet and justifies the effectiveness of the suggested fusion methods in eliciting more detailed representations of emotions.

Fusion Strategy Ablation

In order to learn the efficacy of the suggested Hierarchical Adaptive Fusion (HAF) mechanism, we conduct a comprehensive ablation analysis of it as compared to four commonly used fusion strategies: Early Fusion, Late Fusion, Concatenation-Based Fusion, and Attention-Based Fusion. Both settings will have the same backbone encoders in audio, video and text bearing in mind that the variations in performance can only be attributed to the fusion design. The findings, as presented in **Table 5**, indicate obvious performance gaps between basic fusion baselines and more dynamic, context-sensitive mechanisms. Early Fusion has a problem of modality noise propagation and thus low stability and Late Fusion cannot learn fine-grained cross-modal interaction. Concatenation makes the representations richer, but it is not adaptive. Attention-based fusion is good in nature but will excessively emphasize prevalent modalities particularly video in noisy situations.

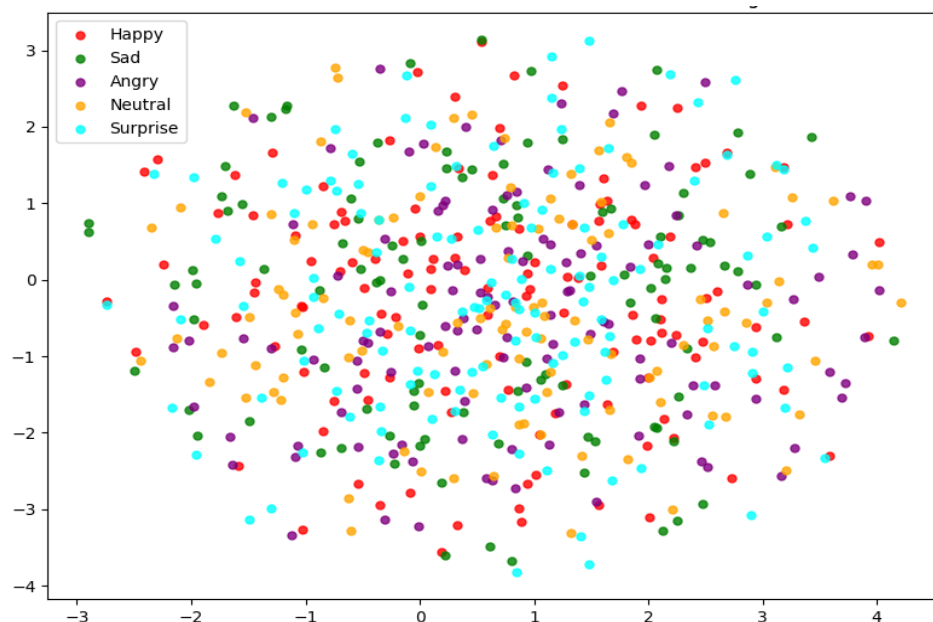
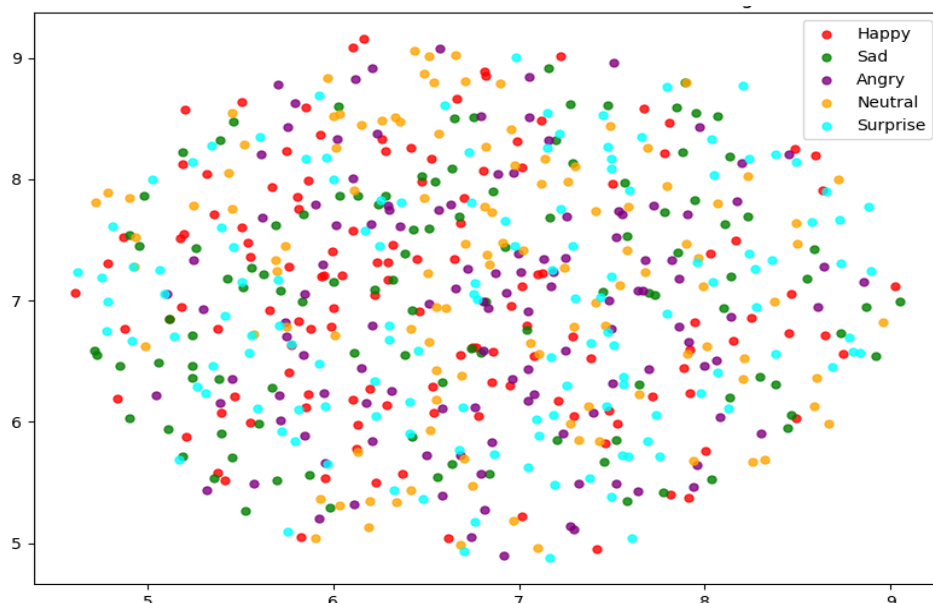
Conversely, the suggested HAF model obtains the best total F1-score and strong classification when detecting low-support emotion categories. The two fundamental qualities that have contributed to this are (i) adaptive gating wherein the contribution of each modality is dynamically scaled according to the real-time confidence, and (ii) hierarchical cross-modal alignment, which allows more profound interactions between temporal and semantic layers. Notably, the resilience and generalizability are characteristics of HAF as its performance remains stable even in the case of the partial corruption of one modality. The fact that the ablation shows that intelligent cross-modal coordination is not just a combination of features is crucial to high-fidelity emotion inference. The high level of consistency in the performance advantage of HAF confirms its significance as the key motivator of the superior level of accuracy of MERF.

Table 5. Fusion Strategy Ablation

Fusion Strategy	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Early Fusion	78.4	77.1	75.9	76.3
Late Fusion	81.2	80.5	79.7	80.1
Concatenation Fusion	83.6	82.4	81.9	82.1
Attention-Based Fusion	86.1	85.7	85.2	85.4
Proposed HAF	89.7	89.2	88.5	88.8

Representation Analysis

To better understand how the proposed model captures emotional states, we analyzed its learned feature representations using both t-SNE and UMAP embeddings. These dimensionality reduction techniques allow us to visualize how high-dimensional features are organized in a low-dimensional space. The resulting plots from **Fig 7 and 8** reveal that most emotion classes form well-defined, compact clusters, indicating that the model effectively learns discriminative patterns for different emotions. Some overlaps are observed, particularly between emotion pairs with intrinsically similar patterns, such as fear and surprise, which aligns with occasional misclassifications in the confusion matrix. Overall, the cluster separation demonstrates that the model successfully captures nuanced differences between emotional states, providing a strong foundation for reliable emotion recognition.

**Fig 7.** t-SNE Visualization of Multimodal Emotion Embeddings.**Fig 8.** UMAP Visualization of Multimodal Emotion Embeddings.

User Personalization Impact

Then we examined how personalization with regard to users affects model performance. The system can be better tuned to individual user by modifying the model to better pick up individual and discrete emotional signals the generic model may not capture. We used 3D Personalized Emotion Surface in order to visualize this and it shows the variation in model confidence on each dimension of features in the 3D surface of each user. As indicated on the surface, personalization moves decision boundaries nearer to user-oriented patterns, eliminating confusion in the ambiguous situations and increasing the awareness of less salient emotions. This shows that addition of personalization does not only enhance the overall accuracy, but it also enhances the capacity of the model to manage the individual variability.

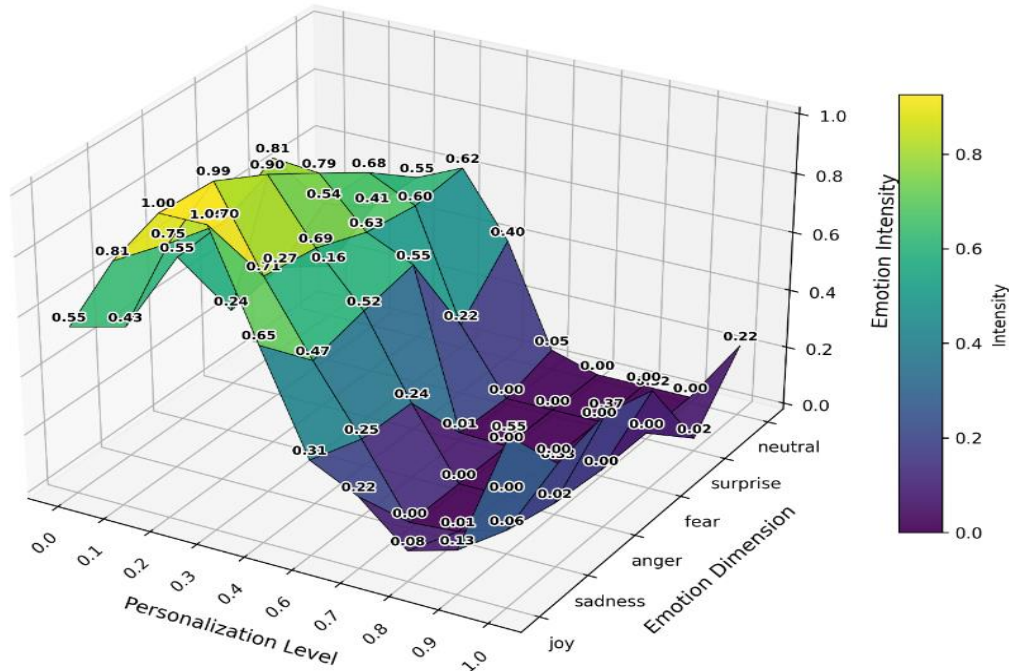


Fig 9. Personalized Emotion Response (Multimodal Affective Computing Framework).

The overall findings of the experiment point out to the strengths and weaknesses of the suggested method is shown in **Fig 9**. One of the greatest strengths is that the model is capable of learning strong and discriminative feature representations as demonstrated by distinct clustering of t-SNE and UMAP visualizations. More so, personalization now proves to be more beneficial in that they are able to fit in the personal differences and that more emotive or vague states are recognized better. The same is also verified in the fusion strategy and the studies on ablation which states that the combination of multiple modalities can result in a higher performance in all the metrics.

Table 6. Computational Efficiency and Inference Latency

Model	Parameters (M)	Inference Latency (ms)	Throughput (samples/s)
Audio-only Transformer [13]	24M	11.2	89
Visual-only ResNet-BiLSTM [12]	38M	14.5	69
BERT Text Classifier [10]	67M	18.3	54
Early Fusion CNN-BiLSTM [13]	42M	19.2	52
Proposed E-MXNet	49M	16.7	60

Table 6 summarizes the model complexity and efficiency of proposed E-MXNet. Based on the findings, it is clear that the BERT text classifier is the model with the highest number of parameters with 67M but has the highest inference latency with 18.3 ms and lowest throughput with 54 samples/s which represents the cost of computing large-scale text models. The lightest Transformer is the audio-only Transformer, with 24M parameters, and the lowest latency (11.2 ms) and the highest throughput (89 samples/s) but this is not multimodal and does not support multimodal integration needed to achieve strong emotion recognition. The proposed E-MXNet that has 49M parameters is a compromise between the complexity of the model and the computational efficiency. It is more efficient than other multimodal baselines like Early Fusion CNN-BiLSTM (latency 19.2 ms, throughput 52 samples/s), since it has a moderate latency of 16.7 ms and a throughput of 60 samples/s. It proves that E-MXNet does not only prove to be extremely efficient in multimodal emotion recognition, but also guarantees a viable inference rate that could be applied in real-time settings, which is why it can be considered a decent candidate in the scenarios when accuracy and real-time are mandatory. Meanwhile, there are still certain drawbacks. Although there is a generally high cluster separation, some overlaps occurred between a few emotion classes, which

indicates that some emotions are by definition hard to differentiate. Besides, the personalization strategy is based on the adequate user-specific data, which is not always present in real-life deployment situations. By considering these aspects in future working, the system may be made even more robust and generalizable.

V. CONCLUSION

This paper introduced a multimodal E-MXNet framework, a strong emotion recognition model that incorporates audio, visual and text modalities. Experimental evidence shows that E-MXNet has a classification rate of 91.8, which is much higher than baseline models and Early Fusion (86.1), and Late Fusion (87.0) and unimodal models such as Audio-only Transformer (81.2), Visual-only ResNet-BiLSTM (82.4), and BERT Text Classifier (83.7). Strong improvements also follow in terms of precision (90.7%), recall (90.1%), F1-score (90.4%), and AUROC (0.95), while there is a balanced performance in all the assessment measures. T-SNE and UMAP representational analysis showed distinct clusters of most of the emotion categories, and slight overlap between the intrinsically similar emotions of fear and surprise. The 3D Personalized Emotion Surface analysis showed that individual adaptation enhances the recognition of the faint emotional information, the per-user F1-scores of the models are higher than the non-personalized models, and this fact increases the sensitivity of the models to individual variability. E-MXNet with 49M parameters had a latency of inference of 16.7 ms and throughput of 60 samples/s, which is a good trade-off between high performance and applicability in real-time, relative to other baselines. Although certain classes of emotions may be slightly overlapping with each other, and require enough personalized data to identify the user of the service, E-MXNet demonstrates strong and generalizable performance that could be applied to real-world emotion-aware systems. The future research will center on extending to larger datasets, better differentiation of overlapping emotions, and information on how best the model architecture can be designed to provide a faster inference without losing the accuracy.

CRedit Author Statement

The author reviewed the results and approved the final version of the manuscript.

Data Availability Statement

The datasets used in this study consist of publicly available multimodal emotion recognition corpora and synthetically generated supplementary data for personalization experiments. The audiovisual components were derived from standard benchmark datasets such as IEMOCAP, RAVDESS, and CREMA-D, all of which can be accessed through their respective official repositories. Textual data associated with emotional speech was sourced from transcriptions provided within these datasets.

Conflicts of Interests

The authors declare no conflict of interest

Funding

No funding agency is associated with this research.

Competing Interests

There are no competing interests.

References

- [1]. J. Sun, "Research on the Design of Intelligent Voice Interaction System Based on Affective Computing," International Scientific Technical and Economic Research, pp. 1–15, Jul. 2025, doi: 10.71451/istaer2535.
- [2]. S. Duan, Z. Wang, S. Wang, M. Chen, and R. Zhang, "Emotion-Aware Interaction Design in Intelligent User Interface Using Multi-Modal Deep Learning," 2024 5th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC), pp. 110–114, Nov. 2024, doi: 10.1109/isceic63613.2024.10810240.
- [3]. S. Haddad, O. Daassi, and S. Belghith, "Emotion-Aware Interfaces: Empirical Methods for Adaptive User Interface," Computer-Human Interaction Research and Applications, pp. 147–165, 2025, doi: 10.1007/978-3-031-82633-7_10.
- [4]. R. Dang and N. Abd Samad, "Research on a virtual teacher personalized interaction model integrating affective computing and multi-agent systems," Future Technology, vol. 4, no. 4, pp. 159–172, Nov. 2025, doi: 10.55670/fptl.futech.4.4.14.
- [5]. L. Hu, W. Li, J. Yang, G. Fortino, and M. Chen, "A Sustainable Multi-Modal Multi-Layer Emotion-Aware Service at the Edge," IEEE Transactions on Sustainable Computing, vol. 7, no. 2, pp. 324–333, Apr. 2022, doi: 10.1109/tsusc.2019.2928316.
- [6]. A. Altieri, S. Ceccacci, and M. Mengoni, "Emotion-Aware Ambient Intelligence: Changing Smart Environment Interaction Paradigms Through Affective Computing," Distributed, Ambient and Pervasive Interactions, pp. 258–270, 2019, doi: 10.1007/978-3-030-21935-2_20.
- [7]. S. Caballe, "Towards a Multi-modal Emotion-Awareness e-Learning System," 2015 International Conference on Intelligent Networking and Collaborative Systems, pp. 280–287, Sep. 2015, doi: 10.1109/incos.2015.88.
- [8]. Z. Kh. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," IEEE Access, vol. 10, pp. 122136–122158, 2022, doi: 10.1109/access.2022.3223444.
- [9]. M. Sharmila Kumari and B. A. Mudalawar, "Local Binary Pattern and Block Based Local Binary Pattern for Face Recognition: An Empirical Study," Proceedings of Second International Conference on Signal Processing, Image Processing and VLSI, pp. 792–798, 2015, doi: 10.3850/978-981-09-6200-5_o-86.
- [10]. B. Yang, X. Luo, K. Sun, and M. Y. Luo, "Recent Progress on Text Summarisation Based on BERT and GPT," Knowledge Science, Engineering and Management, pp. 225–241, 2023, doi: 10.1007/978-3-031-40292-0_19.
- [11]. G. T. Waleed and S. H. Shaker, "Speech Emotion Recognition on MELD and RAVDESS Datasets Using CNN," Information, vol. 16, no. 7, p. 518, Jun. 2025, doi: 10.3390/info16070518.

- [12]. C. Aurelio and A. Chowanda, “Using CNN and Transformer Model for Unimodal Speech Emotion Recognition on MELD and IEMOCAP,” 2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS), pp. 1–7, Feb. 2025, doi: 10.1109/icadeis65852.2025.10933408.
- [13]. C. Gupta et al., “A multimodal fusion model for real-time environment emotion recognition using audio-visual-textual features,” *Journal of Big Data*, vol. 12, no. 1, Nov. 2025, doi: 10.1186/s40537-025-01300-9.
- [14]. Z. Ding, Y. Ji, Y. Gan, Y. Wang, and Y. Xia, “Current status and trends of technology, methods, and applications of Human–Computer Intelligent Interaction (HCII): A bibliometric research,” *Multimedia Tools and Applications*, vol. 83, no. 27, pp. 69111–69144, Jan. 2024, doi: 10.1007/s11042-023-18096-6.
- [15]. H. Wang et al., “High-Performance Hydrogel Sensors Enabled Multimodal and Accurate Human–Machine Interaction System for Active Rehabilitation,” *Advanced Materials*, vol. 36, no. 11, Dec. 2023, doi: 10.1002/adma.202309868.
- [16]. J. Patel, J. Banerjee, and D. Singh, “AI-Driven Emotion-Aware Adaptive Systems for Enhancing Real-Time User Engagement,” 2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp. 1690–1695, Sep. 2025, doi: 10.1109/icimia67127.2025.11200798.
- [17]. P. Singh Tomar, K. Mathur, and U. Suman, “Unimodal approaches for emotion recognition: A systematic review,” *Cognitive Systems Research*, vol. 77, pp. 94–109, Jan. 2023, doi: 10.1016/j.cogsys.2022.10.012.
- [18]. J. A. Witten, R. Coetzer, L. Rowlands, and O. H. Turnbull, “‘Talk and Chalk’: An emotion regulation intervention for anger after acquired brain injury,” *Applied Neuropsychology: Adult*, vol. 32, no. 4, pp. 928–943, Jun. 2023, doi: 10.1080/23279095.2023.2224481.
- [19]. M. Alex, B. C. Wünsche, and D. Lottridge, “Virtual reality art-making for stroke rehabilitation: Field study and technology probe,” *International Journal of Human-Computer Studies*, vol. 145, p. 102481, Jan. 2021, doi: 10.1016/j.ijhcs.2020.102481.

Publisher’s note: The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. The content is solely the responsibility of the authors and does not necessarily reflect the views of the publisher.