Int. J. Advance Soft Compu. Appl, Vol. 18, No. 1, March 2026
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Emotion-Aware Adaptive User Interfaces for Enhanced User Experience Using Multi-Modal Deep Learning

Ahmed Alshehri[1],

[1] Department of Information Technology, Faculty of Computing and Information, Al-Baha University, Al-Baha, Saudi Arabia
e-mail: a.alyehyawi@bu.edu.sa

Abstract

Emotion-conscious computing is decisive in the further development of human-centered digital interaction, but the direct role in improving the User Experience (UX) has not been studied in detail. This paper introduces an Emotion-Aware Adaptive User Interface (EAAUI) system, which uses multi-modal deep learning to enhance usability, engagement and cognitive efficiency by adapting to emotions in real-time. The suggested solution combines facial expressions, speech prosody, and physiological cues with the help of the hybrid deep learning structure that consists of Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory networks (BiLSTMs), and Transformer-based attention systems to provide strong emotion detection and fusion. Emotions sensed dynamically are the motivators of adaptive interface changes, such as simplifying layout, modulating colors, and providing personal feedback. The framework is tested on benchmark datasets (AffectNet, RAVDESS and DEAP) and controlled user study with 30 people. The experimental findings show that the emotion recognition performance is good with an accuracy of 89.6% and an F1-score of 0.88. According to the UX view, the adaptive interface made task completion (21% less) and error reduction (18% less) significantly faster, user engagement (26% more) higher, and the System Usability Scale (SUS) score was 82.5. The results validate the claim that the operational effect of emotion-aware adaptive interfaces is beneficial to user experience, as they provide viable implications to UX-based applications in the educational, healthcare, and intelligent interactive systems.

Keywords: *User Experience (UX); Emotion Recognition; Multimodal Deep Learning Fusion; Adaptive User Interface; User Study*

1 Introduction

Emotion-aware computing is moving quickly to become the foundation of the next generation of human-computer interaction (HCI), because there is a growing need for human-computer systems that are intelligent and task-centric, but also emotionally aware [1]. As digital technologies become more integrated into everyday life, the need for emotionally aware systems extends into educational, healthcare, gaming, and intelligent assistant systems, among others [2]. In education, emotion-aware systems can understand moods associated with confusion, boredom, or engagement and adapt content accordingly

Received 28 July 2025; Accepted 27 December 2025

or provide personal assistance [3]. Healthcare emotion-aware systems, either in mental health contexts or related to assistive technologies, need to detect moods related to stress or anxiety, in order to guide appropriate intervention or improve patient outcomes. In gaming and entertainment personal experiences, there can be improved immersion or user satisfaction based on adjusting narrative flow, challenge level, or audio or visual feedback based on emotional state [4]. In the context of intelligent assistants, emotion-aware systems have been shown to produce user experiences that result in higher levels of trust.

Emotions affect how users process information, make decisions, and interact with the digital environment [5]. The nature of a user's affective state, whether positive states such as interest, excitement, or curiosity or negative states such as frustration, confusion, or anxiety, affects cognitive processing, memory, and learning efficiencies [6]. Positive affective states enhance cognitive processing ability, memory retention and learning efficiency while negative affective states detract from performance, increase cognitive load, and lead to disengagement [7]. Therefore, emotional awareness as part of the user interface design is a necessity to provide truly human centered computing experience. Notwithstanding this reality, most configuration interfaces today remain devoid of the ability to perceive or adapt to the user's emotion. These interfaces are therefore fixed, inflexible, and/or lacking in sensitivity to the affective flow in a human interaction that may lead to either a person-centric or frustrating experience for the user, especially in high-stakes or high-affect situations. One fundamental challenge to developing real-time emotion-aware adaptive interfaces is the accurate, continuous, and non-intrusive detection of emotional signs and interfacing modifications based on that sign recognition [8]. Conventional systems for detecting emotional state utilize unimodal approaches (e.g., facial expressions detected by webcams; speech prosody detected by microphones); however, unimodal approaches all have significant limits in capturing the full depth and range in human emotion [9]. For example, people can intentionally suppress facial expressions or vary widely in presentation of facial expressions based on cultural norms. One could produce in tandem a similar argument for the use of vocal tone or prosody metrics such as pitch or cadencing, as these may or may not capture emotion (e.g., chronic illness or fatigue) or overall affect [10]. Considering these inherent differences across unimodal approaches, this accordingly engenders variability that often increases the chance of misinterpreting emotional state and reduces the general robustness of the system. The issue of emotion detection is immense and overarching challenge. To ideally address this, working through the design of a system to use multimodal data is a vastly more efficacious approach to get at the entire nuance of emotional state of the interaction for the user.

Conversely, multi-modal fusion also introduces a whole lot of uncertainty as it relates to recombining asynchronous data streams, sampling rates of data, adding noise and/or artifacts, and often most importantly the low-latency needed for a range of applications. Also, multimodal models are much more computationally expensive, and need dedicated-deep learning architectures to articulate the models for practical deployment, especially with more real-time implementations [11]. From a design perspective, in successfully transforming the detected behavior to meaningful adaptation in the design of the interface need to effectively balance the ethical responsive and usability of the experience.

With the corresponding challenges and opportunity, the research will directly focus on two, unique aims. The first aim is to design a framework for an emotion recognition model based on a deep learning method to fuse the multi-source authentically created from facial,

audio, and physiological signals in user context, making the detection to be robust and real-time. Second, we aim to design a user-interface that translates the detected user behavior into programmatically generated visual representation, informed by visual design research and literature, and responds interactively to the user's emotions. This enables an immersive, user-centered experience that promotes personalization and socially intelligent design. To direct support these two aims, the research will lend three contributions. We propose a novel multi-modal deep learning architecture that leverages a convolutional neural network (CNN) for analyzing facial expressions; long short-term memory (LSTM) networks for analyzing audio and EEG signals; and an attention-based fusion approach to effectively integrate detection across the aforementioned modalities. We build an Emotion-Aware Adaptive User Interface (EAAUI) prototype that enables modifications of visual elements, interaction cues, and feedback mechanisms, contingent upon the detected emotional response. Lastly, we evaluate the efficacy of the proposed system through testing on standard datasets (AffectNet, RAVDESS, and DEAP) and in a structured user review process with 30 participants, which integrates objective performance measures, while exploring user satisfaction through subjective user feedback.

2 Related Work

The use of multi-modal deep learning has changed the development of emotion-aware user interfaces, enabling systems to better recognize and respond to emotional signals from users. A study by [12], presented Screen2Words, a multimodal framework that takes UI screens and summarizes them as coherent text; this demonstrates the potential value of aggregating visual and textual modalities. A more sophisticated development was proposed by [13], who developed a deep multimodal architecture to create latent user representations, driven by user interactions, to adaptively change based on emotional states. A study by authors [14] used physiological signals (EEG) and speech data to improve emotion recognition (AER), even in the presence of incomplete information. More recent work highlights design for emotion-aware interactions: authors [15] reported on multimodal deep learning for intelligent interfaces that dynamically adapt to users' emotional states, while a study by [16] demonstrated that integrating facial and text based cues increased user satisfaction with the empathy of the virtual assistant. Studies [17, 18] specifically on modeling physiological signals and on the application of Large Language Models (LLM) for affective fusion, further enhance the focus on robustness in the presence of noisy modalities. Studies [19, 20] that pointed to advances in multimodal emotion recognition due to deep learning. Additional studies that were bias-aware [21] and gender-aware [22] adaptive models were introduced and reported to personalize information in real time. Models that used temporal-attention [23] and transformers [24] did value-added methods for learning emotion in multimodal situations; taken together all indicated that multimodal deep learning is foundational for intelligent, emotion-adaptive user interfaces, during interaction.

3 Theoretical Foundations

Theoretical foundations will serve as a basis for our design of an emotion-aware adaptive user interface framework. Specifically, establishing psychological models of emotion, examine multimodal affective signals which behave and operate in a complementary fashion. To provide the basis for examining emotion, as well as create a discussion of current state of the art deep learning models to process, analyze, and fuse data streams.

3.1 Models of Emotion

Models of emotion provide the conceptual foundation for automated emotion recognition systems. Two leading models would be Ekman's Six Basic Emotions and the Valence–Arousal (V-A) Model.

3.1.1 Ekman's Six Basic Emotions

Paul Ekman's model specifies six basic emotions that humans universally recognize: happiness, sadness, anger, fear, surprise, and disgust. For each of these basic emotions, a person's facial expressions carry a predictable set of recognizable facial action units (i.e., behaviors) regardless of one's cultural background. Due to the phenomenon of biomechanical movement of facial muscles, and the ability to label and categorize the face in images in databases related to emotions, these six unique, separate, and discrete canonical categories serve as the theoretical basis for many emotion recognition systems based in computer vision technology. Although the labels in the schematic are internally consistent, the strength of Ekman's schema lies in its labeling and classification framework for recognizing discrete categories of emotion from observable facial expressions . Nevertheless, the reality of complex micro- and macro-expressions, and the subtle gradations of affective states that can be invoked through "real" interactions, may render these discrete categories less satisfactory in the long-run [25].

3.1.2 Valence–Arousal Model

The V–A model represents emotion as something that exists in continuous (i.e. two-dimensional) space. In rated emotion, the V dimension represents the degree to which an emotion is positive or negative, and the A dimension represents the degree to which an experience is intense, activating (e.g., high valence, high activation = euphoria, and low valence, low activation = sadness). Compared to classical models, the V–A Model yields subtle representations of affect. The V–A Model also promotes personalized and specific context-sensitive adaptation in an interactive system. V-A models work well when emotional signals must be tracked over an extended timeline, or when access to multi-modal information across physiological information and behaviors is included [26].

4 Multi-Modal Data Streams

Human emotions are expressed in various modalities including facial expressions, vocal intonation and physiological changes. A multi-modal approach utilizes each modality optimally in order to increase reliability and context considerations when recognizing emotions.

4.1 Visual Modality (Facial Expressions)

Facial expressions are perhaps the easiest indicator of emotional state and have the most interpretability in nonverbal behaviour. Important features from visual signals consist of eye closure speed, eye brow position, curvature of the mouth, or any number of facial action units (FAUs) identified in the Facial Action Coding System (FACS) [27]. Visual features are extracted through CNNs from either static images or video sequences. Visual expressions cannot be captured tangentially, nor is it available with occluded faces, or lighting conditions [28]. Cultural expressiveness may also contribute to challenges in

feature extractions or visual analysis and requires robust visual feature extraction and temporal smoothing to increase continuity of face data.

4.2 Audio Modality (Speech Prosody)

Speech can convey affect through para-linguistic features within speech such as pitch, tone, rhythm or loudness i.e., an elevated pitch and irregular rhythm may suggest excitement or stress, compared to a monotone consistent pitch suggesting boredom or sadness. Mel-frequency cepstral coefficients (MFCCs), spectral features and energy profiles are typically extracted from audio streams and modeled with temporal topology form an audio deep learning architecture [29]. Audio provides remarkably helpful secondary modality data when faces are neutral or ambiguously positive or negative.

4.3 Physiological Modality (EEG, HRV, GSR)

Physiological data provide objective, non-observable indicators of emotional states and are unlikely to be subject to conscious cognitive manipulation. Physiological based signals (Table 1) are typically obtained via wearable sensors or head mounted devices and can add value to observed behaviour data particularly when the data source may be more at stake, e.g. clinical event [30].

Table 1: Common Physiological Signals

Signal	Description
Encephalogram (EEG)	assisting with brainwave activity useful to recognize valance or arousal,
Heart Rate Variability (HRV)	generally measuring the autonomic nervous system and emotional regulation
Galvanic Skin Response (GSR)	measuring changes in skin conductivity directly related and activated by the physiology of activated sweat glands suggesting arousal.

5 Deep Learning Techniques

The intricacies involved with multi-modal affective data (high dimensionality, temporal dependencies, cross-modal differences in variability, and noise) necessitate the complexity of machine learning models that can identify subtle and complex patterns [31]. Basic machine learning models that leverage pre-engineered features (e.g., intensity histograms, prosodic descriptors, or statistical properties of EEG signals) have achieved some limited success, though the more common outcome is that they do not generalize well across contexts and/or users [32]. For these reasons, deep learning has become the primary modeling approach in affective computing, as it can automatically learn hierarchical feature representations from raw or minimally processed data, thereby reducing the need for domain-specific feature engineering. By stacking multiple layers of nonlinear transformations, deep learning models can learn low-level features, such as image edges or frequency shifts in audio signals, and progressively combine them into higher-level semantic concepts, including smiling, vocal distress, or EEG-based arousal.

5.1 Convolutional Neural Networks (CNNs)

Convolutional neural networks (CNN), represent an essential foundational component of modern visual recognition systems and are especially well-capable of extracting spatial and structural characteristics from two-dimensional datasets [33]. For emotion recognition, CNNs are used to analyze facial images or video frames to discover expressions. The CNN's convolution filters are organized into layers that automate feature extraction. Initial layers differentiate edges or textures while future layers detect configurations of the face (e.g., raising of the eyebrows, compressing the lips, movements of the cheeks), features commonly associated with emotional states. Pre-trained CNN architectures, such as VGGNet, ResNet, and EfficientNet have served as commonly used backbone models for emotion recognition studies. These models are pre-trained on large-scale image datasets. Fine-tuning allows the backbone model to retain general visual features while retraining model weights to recognize features of an affective expression. In addition, with the added temporal consideration, 3D CNNs and CNN-LSTM based systems have been developed that leverage a CNN backbone to observe the timed progression of expressions over time. These models are able to observe quickly evolving micro-expressions or subtle affective signals embedded consecutively in motion sequences.

5.2 Long Short-Term Memory Networks (LSTMs)

Although CNNs are strong for learning spatial features, emotional signals generally unfold in a dynamic series of temporality and build on the memory of previous states [34]. Specifically, LSTMs, which are a specialized form of RNNs, were created to satisfy the requirement of memory through the addition of memory cells and gating (i.e., input, forget, and output). That way, LSTMs can keep and/or discard information from previous steps over time and help to reduce the common issue of vanishing gradients seen in standard RNNs [35]. In affective computing, LSTMs are effective for tasks involving audio signals and for modeling physiological data, as in this work. For instance, the variations of prosody, or how one forms questions by changing pitch or loudness, are temporal by nature, or occur in a series of events, that can be used in calculating levels of emotional intensity, or proof of stress. EEG signals can also be used to capture dynamic fluctuation probabilities on frequencies (i.e., alpha, beta, gamma) during cognitive load and emotional load influence, and LSTMs can capture these dependencies then subsequently used to infer user experientially arousal or cognitive load. Bidirectional LSTMs (BiLSTMs are RNNs with bi directional capabilities, where LSTMs can use both past and future information to classify the emotional cadence of the representations.) have also been used in elaborating emotion recognition processes through mastering temporal relations in sequential tasks [36]. BiLSTM also allows the model to incorporate and rely on the temporal information of experience on a class-level. This helps the LSTM classify the emotional responses accurately.

5.3 Transformer Architectures

Transformers, initially designed for natural language processing, have changed the landscape of multi-modal affect recognition with novel architectures based on self-attention, as opposed to recurrence. Self-attention provides the transformer model with the ability to assess relationships among all input components in parallel, enabling the simultaneous processing of long-distance dependencies more efficiently than RNN approaches. In multi-modal systems, this architecture expands to cross-modal attention,

allowing for the assessment of relationships in visual, auditory, and physiological data in tandem. For example, a sudden increase in pitch in the vocal input and a frown in the visual channel can indicate anger or frustration at the same time. The Multimodal Transformer (MulT) and Perceiver IO illustrate this process through their ability to jointly align asynchronous modalities in an event, while facilitating two-directional information sharing across asynchronous streams of input — very common in both multi-modal emotion recognition [37].

The Transformers also scale up appropriately when datasets become larger, and they are highly transferable by adapting pretrained models such as BERT for language, or Vision Transformer for visual data for analysis of cognition. Their unique strengths in bridging the modalities of different coupling types makes them perfect for emotion-aware adaptive interfaces, where all modalities can work together in the capture of an individual radius of emotion in contrast to tasking to just one.

5.4 Hybrid Deep Learning Architectures

Given the complementary strengths of CNNs, LSTMs, and Transformers, we propose a hybrid architecture for the proposed framework. While CNNs are used to exploit spatial feature extraction from facial images, LSTMs capture the temporal dynamics between audio and EEG sequences, and transformers provide cross-modal attention while harnessing and fusing heterogeneous features into a single representation. This unique multi-stage pipeline allows the system to simultaneously capture local spatial cues, sequential dependencies, and cross-modal interactions. In this way, the proposed framework is sufficient for context-sensitive emotion recognition in a robust manner that drives real-time adaptive user interface changes.

6 Proposed Framework

This framework combines multimodal processing of affective signals with real-time interface adaptation. The architecture of the system is modular and hierarchical to enable a level of scalability, interpretability, and responsiveness to be realized. In Fig.1 (Hybrid Deep Learning Architecture), the end-to-end data flow is illustrated from raw input of sensors to emotion-based adaptation of the interface.

6.1 System Architecture

The architecture consists of six primary layers: Input Layer, Preprocessing Layer, Feature Extraction and Multi-Modal Fusion Layer, Emotion Classification Layer, and UI Adaptation Module. Each section serves a different purpose of processing input low-level sensor data, leading to an overall high-level adaptive behavior.

6.2 Input Layer

The Input Layer begins the proposed Emotion-Aware Adaptive User Interface (EAAUI) process by collecting multi-modal data indicative of a user's emotional state. The Input Layer captures signals from visual, auditory, physiologic, and behavioral modalities in order to provide richer information about emotions. Visual input is recorded with either a stereo webcam or a depth camera that tracks facial expressions, micro-movements, and gaze behaviors, all of which signal emotions, such as happiness, confusion, or frustration. Audio input is recorded with an on-device microphone, where acoustic features such as

tone, pitch, prosody, and rhythm are analyzed and also abstractly signal emotional arousal and feelings. Physiologic input is recorded with wearable sensors that measure brain activity, skin conductance, and heart rate variability through the user experience and they provide objective measures of internal affective states. These behavioral indicators represent additional signals of user affect that are relevant during user discussion, and measure typing speed, cursor movement, or touch pressure as secondary measures of user engagement and stress, during the studies. Importantly, all of the data streams are synchronized for timing to accurately process across modals. Combined, these inputs provide an extremely thorough picture of the user's internal emotional state, and these data are intended to inform all processing, classification, and responsive interface behavior.

6.3 Preprocessing Layer

The preprocessing layer clears, standardizes, and aligns all input signals in time for feature extraction and deep learning analysis. The preprocessing ensures that heterogeneous data from the multiple sensors are converted into aligned and standardized representations for potential fusion. For the visual modality, the preprocessing stage may include face detection, alignment, and cropping of a face using an algorithm such as Haar cascades or MTCNN, after which it is subjected to illumination correction, and frame resampling at a fixed rate (30 fps, for example, so that both modalities are temporally aligned). For that audio stream, background noise suppression is carried out using spectral subtraction, and then the signal is segmented into short overlapping frames of 25 ms for the extraction of Mel-Frequency Cepstral Coefficients (MFCCs), which captures the emotional characteristics of speech. In the physiological modality, we apply the band-pass filter to EEG and related biosignals between 0.5 and 45 Hz to attenuate lower frequency drift and eliminate higher frequency noise. We then perform removal of artifacts using Independent Component Analysis (ICA), and apply z-score normalization to standardize the amplitude. Finally, the synchronization component manages the alignment of events across modalities, using interpolation and resampling, accounting for any slight temporal discrepancies (using timestamps). In summary, all modalities are aligned in time at the output of the preprocessing layer. The outcome of the preprocessing layer is a set of cleaned, normalized and temporally aligned tensors for multimodal feature extraction and emotion classification.

6.4 Feature Extraction and Multi-Modal Fusion Layer

This model transforms raw sensor data into compact, discriminative feature representations. Each modality is processed by a dedicated deep learning sub-network optimized for its data characteristics. The visual branch employs a fine-tuned CNN to extract spatial and structural features from facial regions, capturing texture variations, muscle dynamics, and action units. The resulting facial feature vector is represented as $F_v \in \mathbb{R}^{d_\alpha}$. The audio branch utilizes a Bidirectional Long Short-Term Memory (BiLSTM**)** network to model temporal dependencies and emotional prosody within Mel-Frequency Cepstral Coefficient (MFCC) sequences, generating an audio embedding $F_\alpha \in \mathbb{R}^{d_\alpha}$. Similarly, the physiological branch processes EEG and heart rate variability (HRV) signals through a Stacked LSTM or 1D-CNN**,** encoding temporal and frequency-based features into $F_p \in \mathbb{R}^{d_p}$.

Once modality-specific embeddings are generated, they are fused to create a unified affective representation. In early fusion (Eq. 1)**,** features are concatenated:

$$F_{fusion} = [F_v; F_\alpha; F_p] \quad (1)$$

allowing the model to jointly learn inter-modal relationships. Late fusion aggregates modality-wise emotion probabilities as given in Eq. 2.

$$P(E) = \sum_{m-1}^{M} \omega_m P_m\,(E) \quad (2)$$

where ω_m indicates the confidence weight of each modality. The investigation utilizes a Transformer-based hybrid fusion method, which utilizes both self-attention and cross-attention mechanisms to align contextual dependencies and process the information together. This method is thoroughly robust to noise and missing modalities, ensuring that the multi-modal emotional representation is coherent.

6.5 Emotion Classification Layer

This Layer transforms the fused multi-modal representation into final affective outputs. The feature vector F_{fusion} is passed through a fully connected neural network with a softmax layer for discrete emotion recognition or regression heads for continuous affect estimation. For categorical prediction, probabilities are computed Eq. (3):

$$P(E_i|\, F_{fusion}) = \frac{e^{zi}}{\sum_{k=1}^{C} e^{zk}} \quad (3)$$

where E_i represents the i^{th} emotion among C classes (e.g., Ekman's six emotions).
For continuous affect modeling, valence and arousal values are estimated as given in Eq. 4.

$$\hat{V}, \hat{A} = f(F_{fusion}; \emptyset) \quad (4)$$

enabling fine-grained emotional interpretation. This layer bridges deep feature learning with real-time affect inference, forming the foundation for adaptive interface responses.

6.6 UI Adaptation Module

The UI Adaptation Module represents the last stage of this approach. It takes the predicted state of emotion and alters the user interface. The emotion which has been detected is sent to the Sustained User Interface Adaptation Engine where it will alter aspects of the interface based upon the pre-set rule-based or reinforcement-learning based adaptation mechanisms. For example, the system can use color adaptation (i.e., muted cool tones when stressed to bright colors if the user appears to be engaged) layout adaption (reduce cognitive overhead if the user is likely in a state of frustration) and feedback adaptation (e.g., supportive and encouraging messaging if the user appears confused or showing low motivation) to adapt. If a conversation system, audio feedback can also be adapted (i.e., rate or tone of voice) to reflect empathy. The UI Adaptation Module as given in Fig.1, contributes to the maintenance of the emotional feedback process by constantly observing user reactions and refining the interface response with the hope to keep the user engaged, comfortable or in a steady emotional state.

Algorithm 1: End-to-End Emotion-Aware Adaptive User Interface (EAAUI) Framework

Algorithm 1: End-to-End Emotion-Aware Adaptive User Interface (EAAUI) Framework

Input: Visual frames X_v, Audio stream X_a, Physiological signals X_p, Previous UI state U_{t-1}
Output: Updated UI state U_t

Step 1: Multi-Modal Preprocessing
$X'_v \leftarrow$ FaceAlign + Normalize(X_v)
$X'_a \leftarrow$ Denoise + MFCC(X_a)
$X'_p \leftarrow$ BandPassFilter + Normalize(X_p)
Synchronize $\{X'_v, X'_a, X'_p\}$ using timestamps

Step 2: Feature Extraction
$F_v \leftarrow CNN(X'_v)$;
$F_a \leftarrow BiLSTM(X'_a)$;
$F_p \leftarrow LSTM/1D-CNN(X'_p)$

Step 3: Transformer-based Fusion
$F_{fusion} \leftarrow TransformerCrossAttention([F_v, F_a, F_p])$

Step 4: Emotion Inference
If Discrete Model:
$z \leftarrow W_e \cdot F_{fusion} + b_e$
$P(E_i|F_{fusion}) = e^{z_i} / \sum_{k=1}^{C} e^{z_k}$
$E_t \leftarrow \arg\max_i P(E_i|F_{fusion})$
Else:
$[\hat{V}_t, \hat{A}_t] \leftarrow f(F_{fusion}; \theta)$
Apply smoothing (EMA or median filter)

Step 5: UI State Mapping
If Discrete Emotion: $s_t \leftarrow MapDiscrete(E_t)$
Else: $s_t \leftarrow Quadrant(\hat{V}_t, \hat{A}_t)$

Step 6: Hysteresis for Stability
If TimeInState(s_t) < τ and Δs small: $s_t \leftarrow s_{t-1}$

Step 7: Policy Selection
If Rule-based: $A_t \leftarrow \pi_{rules}(s_t)$
Else (RL): $A_t \leftarrow \arg\max_a Q(s_t, a)$
$Q \leftarrow Q + \alpha[r + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a)]$

Step 8: Apply UI Adaptation
If ColorAdapt: ApplyPalette(cool) if Stress; vivid if Engagement
If LayoutSimplify: ReduceWidgets(); IncreaseWhitespace(); HighlightPrimaryCTA()
If FeedbackAdjust: ShowHints(); EncouragingCopy(); AdjustTone()
If VoiceTone: SetTTS(rate↓, pitch↓) for empathy; rate↑ for motivation
$U_t \leftarrow ComposeUI(U_{t-1}, A_t)$

Step 9: Closed-Loop Evaluation
$R_t \leftarrow EngagementMetrics(clicks, errors, time, self-report)$
Log(s_t, A_t, R_t); Return U_t

Fig.1: Emotion-Aware Adaptive User Interface (EAAUI) Framework

7 Dataset and Experimental Setup

7.1 Datasets

To investigate the performance and robustness of the developed this framework, multiple publicly available benchmark datasets that represented various affective modalities (e.g., facial expressions, speech prosody, and physiological signals) were incorporated.

7.1.1 AffectNet AffectNet [38] is one of the largest facial expression datasets, with over one million images manually labeled for the seven basic emotions (happiness, sadness, anger, fear, disgust, surprise, and neutral), and also the valence-arousal dimensions. The AffectNet dataset was used to train and finetune the visual branch (CNN) of the framework for extracting spatial facial features and micro-expressions related to emotional states.

7.1.2 RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song): RAVDESS consists of 24 actors who engaged in speech and song that expressed emotional states of calm, happy, sad, angry, fearful, surprised, and disgusted. The audio subset was used to train the BiLSTM for learning the temporal and prosodic variations that speech-based emotion cues can express.

7.1.3 DEAP (Database for Emotion Analysis using Physiological Signals): DEAP is a dataset made up of EEG signals, peripheral physiological signals (GSR, heart rate, and respiration), all collected while participants watched affective videos. The dataset has self-reported valence, arousal, dominance, and liking scores. The dataset was used to train the physiological LSTM branch of the system to decode internal affective responses beyond what can be seen through behavior.

Together, the datasets provide a complete representation of visual, auditory and physiological modalities needed to support a realistic multi-modal learning and evaluation environment.

7.2 Preprocessing

To ensure consistency, reduce noise, and time align across modalities, each data source went through a defined preprocessing pipeline. With the visual modality, facial photographs were identified from the video data, cropped, and then standardized in a common position using Multi-task Cascaded Convolutional Networks (MTCNN). The images were resized to 224×224 pixels, normalized to [0, 1] and then augmented using horizontal flip, brightness adjustment, and random rotation for robustness. For the audio modality, speech audio was extracted and mixed to mono audio and then downsampled to a 16kHz frame rate. The sound bar was minimized through spectral subtraction techniques, and the speech audio was broken into 25ms frames with 10ms of overlap. Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from each frame, followed by computing the first and second order derivatives from each coefficient to include potential temporal prosody. For the physiological data (e.g., EEG), after noise reduction through bandpass filtering methods, all recorded data were broken into "chunks" and underwent Independent Component Analysis (ICA) for artifact removal so that sound and/or speech recorded was comparable across subjects. Reporting for all physiological measures were normalized by z-scores for inter-subject comparison. Finally, the time domain of all modalities was synchronized from the original timestamps to an interpolated time frame rate, ensuring that facial modality, audio/speech modality, and physiological measures reflected the same apparent emotion, thereby enabling coherent and temporal multi-modal fusion and end-to-end learning across the entirety of the dataset.

7.3 Training Configuration

All experiments were conducted using a consistent training configuration to ensure transparency and reproducibility. This approach was implemented in PyTorch 2.0, which supports hybrid CNN-, LSTM- and transformer-based models. The training took place on an NVIDIA RTX 3090 GPU (24 GB VRAM) with an Intel Core i9 CPU and 64 GB RAM under Ubuntu 22.04 LTS. The model was trained for 100 epochs using a batch size of 64 utilizing the Adam optimizer with a learning rate of 0.0001. Discrete emotion classification loss was handled using cross-entropy loss, while valence-arousal regression was calculated using mean-squared error (MSE) loss. The learning rate was also smoothed out with a

cosine-annealing scheduled learning rate for convergence. Each modality network (CNN vision, a BiLSTM network for Audio, and LSTM/1D-CNN network for physiology networks) were pre-trained on AffectNet, RAVDESS, a DEAP dataset and then jointly fine-tuned in the transformer-based model in the fusion model. To curb overfitting, a dropout and early stopping methodology was used. The end-to-end differentiable pipeline optimized all modalities together, allowing for synchronized optimization while learning cross-modality emotion for real-time adaptive user interaction. To ensure experimental reproducibility, all experiments were conducted using fixed random seeds, and identical, standardized preprocessing pipelines were applied across all datasets and modalities, including normalization, temporal alignment, and artifact removal. These measures enable reliable replication of the reported results.

7.4 Evaluation Metrics

To evaluate the recognition of emotion and functionality of the adaptive interface two categories of evaluation measures were used. The emotion recognition performance was assessed by the model's use of Accuracy, Precision, Recall, and F1-score to measure reliability of classification across multiple classes of emotion. Each measure was calculated per class and was averaged to support a balanced measure of these measures. Confusion matrices were explored to assess class-specific classification bias and misclassifications across modalities. Evaluation of user experience meant conducting a controlled user study in which participants interacted with this prototype. Overall usability and satisfaction were quantified through the System Usability Scale (SUS), while the perceived cognitive and emotional workload was quantified by the NASA Task Load Index (NASA-TLX) during the interaction of the user study. The subjective assessment complemented the quantitative measures, thus adding to our understanding of how real-time emotional adaptation increased engagement with frustrating emotions and performance for the task overall.

8 Results, User Survey, and Analysis

Experimental results of the proposed framework in the aspects of quantitative model performance, evaluating the adaptive interface, and qualitative feedback through user surveys. Taken together, the results demonstrate the effectiveness of the multimodal emotion recognition system and its real-time interface adaptation in improving user experience and engagement.

8.1 Emotion Recognition Performance

In order to evaluate the accuracy of the proposed multi-modal deep learning architecture, several model configurations were evaluated using various combinations of modalities. The results, presented in Table 2, indicate that including additional modalities clearly improves results.

Table 2: Model Evaluation Results

Model	Accuracy	F1-score
CNN (Face only)	72.4%	0.71
LSTM (Audio only)	65.8%	0.64
CNN + LSTM (Fusion)	84.3%	0.83
CNN + LSTM + EEG (Full)	**89.6%**	**0.88**

The fact that the unimodal CNN and LSTM models had moderate accuracy further highlighted the limitations of unimodal emotion classification. When facial and audio modalities were combined, there was a performance jump of more than 12%, which highlights the complementary role of the visual and auditory systems. With the inclusion of physiological features (EEG signals), both accuracy and F1-score improved to 89.6% and 0.88, respectively, reaffirming that multimodal integration provides a more comprehensive representation of emotional states. These results indicate that the proposed framework is effective when using the hybrid Transformer-based fusion model.

8.2 UI Adaptability Testing

To evaluate the benefits of emotion-aware adaptation, usability testing was conducted to compare the prototype to a traditional static interface. The same tasks were assigned to participants in both conditions, and three objective measurements of performance were collected: task completion time, error rate, and user engagement. The results revealed a performance advantage for the adaptive interface: errors decreased 18% (showing that users made more accurate, and therefore more confident, future actions); task completion time decreased 21% (showing that cognitive load was reduced and interactions were more fluid); and user engagement increased 26% (shown by longer session times and increased clickstream activity). These findings suggest that emotion-aware adaptations (i.e., adaptive color themes, minimalized layout to improve task workflow, and contextual guides) have helped the user maintain control over their cognitive focus while also maintaining feelings of comfort. Affective information through interaction Affectivity also was an attribute of EAAUI, with affective states being detected through user feedback. Interface elements could then be adjusted in real-time. Accordingly, elements contributing to perceived fluidity were incorporated into the system, resulting in a more personalized and interactive communication experience that balances cognitive task efficiency with emotional tranquility. These combined results presented large usability improvements as demonstration of the usability benefits associated with emotion-aware adaptation for longer engagement and satisfaction within human-computer interaction environments.

8.3 User Survey Design and Feedback

In addition to the quantitative data, a user study was carried out which explored subjective perceptions toward the EAAUI system. The user study conducted in this research involved voluntary human participation and was carried out in accordance with ethical standards for research involving human subjects. All participants provided informed consent prior to participation, no personally identifiable information was collected, and the study posed no risk to participants. The procedures complied with institutional and international ethical guidelines for human-centered research. A total of 30 participants ranging in age from 18 to 45 (with 50% of the participants from a technical background) completed the study. Participants used both the static and emotion aware versions of the interface to facilitate a within-subject comparison.

8.3.1 Survey Methodology

Participants completed an 8-item Likert-scale questionnaire (Table 3). In this survey (1 = Strongly Disagree, 5 = Strongly Agree) designed to assess responsiveness, engagement, human-likeness, and overall satisfaction.

Table 3: Survey Questions

Question	Statement	Avg. Score
Q1	The interface responded effectively to my emotional state.	4.2
Q2	I felt more engaged while using the adaptive UI.	4.5
Q3	The emotion-aware features made the interface feel more human-like.	4.1
Q4	The system helped reduce my frustration.	4.3
Q5	UI changes were appropriate and helpful.	4.0
Q6	I would prefer this UI over traditional ones.	4.6
Q7	My privacy was respected.	3.9
Q8	I am satisfied with the emotion-aware interface.	**4.4**

Overall user acceptance is high, as indicated by the means of rating responses, with a notably high ratings for engagement (Q2) and user preference (Q6). Conversely, privacy received the lowest mean score (3.9) indicating that privacy might be a concern for some users which might be alleviated in part, by incorporating features or mechanisms of transparency and control in future versions.

8.3.2 Qualitative Feedback

Participants provided open-ended comments, offering valuable insights into user sentiment:

- *"Felt like the system understood me."*
- *"Loved the calming color changes."*
- *"Needs better handling of neutral mood states."*

The total average for the System Usability Scale (SUS) was 82.5/100, which indicates that the usability is excellent, and the remaining 87% of the participants rated the system as "Good" or "Excellent." Overall, this study demonstrates how adaptive emotional feedback increases user satisfaction, promotes engagement, and results in a more meaningful human–computer interaction.

9 Discussion

This work shows that multimodal emotion processing with adaptive interfaces can result in better user satisfaction and usability. The accuracy of the emotion prediction is significantly increased by the multi-modal system, +17.6% as compared to 72% obtained on average considering a single-modality CNN-based emotional recognition system. Beside improved classification performance for emotion as results also prove that there is significant improvement of usability and user experience with the multi-modal system. In particular, such an approach resulted in a decrease of errors of 18%, a shorter completion time for tasks by 21%, and an increase of the user engagement by 26%. The Interface, created for this research study, successfully utilized the emotionally-derived data collected to adapt the user interface to meet the needs of the user. For example, when users appeared frustrated, this would simplify the interface, and when users appeared disinterested, it would prompt the user to re-engage with the application. This is in line with Norman's

work on Emotional Design where he states that developing interfaces that respond to users' emotions will not only provide better performance, but will develop a sense of trust and empathy between the user and the interface. In addition to the positive usability testing results, many of the survey respondents indicated they felt a greater sense of connection and engagement while interacting with the adaptive interface.

The proposed framework marks a transition from static usability models to a dynamic system capable of detecting and responding to users' emotions, enabling more personalized and empathetic interactions. By integrating emotional feedback to this Interface tailors user experiences in real time. For example, during experiments, the interface simplified layouts when frustration was detected and offered motivational cues when disinterest arose. These adaptive responses improved engagement, trust, and empathy between users and the system. The study's outcomes affirmed the framework's alignment with Norman's emotional design principles, showing measurable performance improvements and a System Usability Scale (SUS) score of 82.5. Users reported feeling better understood and more engaged with the content, underscoring the benefits of emotionally responsive design. However, developing such affect-aware systems presents challenges, including the high cost and calibration of sensors, as well as ethical concerns related to data privacy and user consent. To ensure responsible and sustainable implementation, these issues must be addressed across both academic and practical environments. Ultimately, framework represents a significant step toward next-generation interactive systems that treat emotion as a key input for personalization, thereby transforming human-computer interaction into a more adaptive and human-centered experience.

9 Conclusion and Future Work

The study presented a holistic framework that engages emotion recognition and interface adaptation through a multi-modal deep-learning pipeline. The combined CNN-LSTM-Transformer architecture was effective in integrating facial, auditory, and physiological measures for strong and real-time emotion classification. The results demonstrated that incorporating modality improved detection accuracy. Additionally, adaptive system responses—such as changes in color, layout, and feedback—enhanced user engagement and reduced cognitive workload. Overall, both quantitative experiments and user surveys suggested that incorporating emotional intelligence in the design of an interactive interface provided a more authentic and satisfying interaction experience. From a broader perspective within HCI, this framework is one step forward in describing emotionally congruent computing, where a computing system can dynamically attend to the changing states of its users. However, computational complexity, calibration of sensors, and user privacy are important future considerations. Future work will focus on lightweight, privacy-preserving models suitable to be deployed on edge and mobile devices incorporating federated learning and self-supervised representation learning capabilities to minimize reliance on labelled data in training regime. In addition, extending the adaptive user interface framework to incorporate cultural, contextual, and personality-based personalization can promote inclusivity and emotional authenticity in adaptive experiences.

Funding:

This work was not supported by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Ntoa, S. (2025). Usability and user experience evaluation in intelligent environments: A review and reappraisal. *International Journal of Human–Computer Interaction*, 41(5), 2829–2858. https://doi.org/10.1080/10447318.2024.2394724

[2] Sethi, S. S., & Jain, K. (2024). AI technologies for social emotional learning: Recent research and future directions. *Journal of Research in Innovative Teaching & Learning*, 17(2), 213–225. https://doi.org/10.1108/JRIT-03-2024-0073

[3] Sirisha, N., et al. (2025). Emotion-centric artificial intelligence–driven engagement systems for adaptive learning environments: Personalized knowledge acquisition and cognitive skill enhancement. In *Proceedings of the International Conference on Sustainability Innovation in Computing and Engineering (ICSICE)* (pp. 528–540). https://doi.org/10.2991/978-94-6463-718-2_46

[4] Sinlapanuntakul, W. P., Derby, J. L., & Chaparro, B. S. (2022). Understanding the effects of mixed reality on video game satisfaction, enjoyment, and performance. *Simulation & Gaming*, 53(3), 237–252.

[5] Steinert, S., & Dennis, M. J. (2022). Emotions and digital well-being: On social media's emotional affordances. *Philosophy & Technology*, 35(2), Article 36.

[6] Wang, Y. (2023). Affective state analysis during online learning based on learning behavior data. *Technology, Knowledge and Learning*, 28(3), 1063–1078.

[7] Chang, C. C., & Yang, S. T. (2024). Learners' positive and negative emotion, cognitive processing, and learning effectiveness in task-centered digital game-based learning. *Interactive Learning Environments*, 32(9), 5058–5077.

[8] Tajja, Y., Martin, L., & Herrmann, M. (2025). A concept for dynamic adaptation of intelligent user interfaces based on emotion and behavior. In *Proceedings of the Intelligent Systems Conference* (pp. 276–287).

[9] Udahemuka, G., Djouani, K., & Kurien, A. M. (2024). Multimodal emotion recognition using visual, vocal, and physiological signals: A review. *Applied Sciences*, 14(17), Article 8071. https://doi.org/10.3390/app14178071

[10] Spreadborough, K. (2022). Emotional tones and emotional texts: A new approach to analyzing the voice in popular vocal song. *Music Theory Online*, 28(2).

[11] Sar, A., et al. (2025). Multi-modal deep learning framework for early detection of Parkinson's disease using neurological and physiological data. *Scientific Reports*, 15(1), Article 34835.

[12] Wang, B., et al. (2021). Screen2Words: Automatic mobile UI summarization with multimodal learning. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)* (pp. 498–510).

[13] Tato, A., & Nkambou, R. (2023). Towards a multi-modal deep learning architecture for user modeling. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference (FLAIRS)* (Vol. 36).

[14] Tiwari, U., Chakraborty, R., & Kopparapu, S. K. (2025). Unifying EEG and speech for emotion recognition. *arXiv*.

[15] Duan, S., et al. (2024). Emotion-aware interaction design in intelligent user interfaces using multi-modal deep learning. In *Proceedings of the IEEE*

International Symposium on Computer Engineering and Intelligent Communications (ISCEIC) (pp. 110–114).
[16] Rajesh, S. G., et al. (2025). Enhancement of virtual assistants through multimodal AI for emotion recognition. *IEEE Access*, 13, 102159–102179.
[17] Woo, S., Zubair, M., Lim, S., & Kim, D. (2025). Deep multimodal emotion recognition using modality-aware attention and proxy-based multimodal loss. *Internet of Things (The Netherlands)*, *31*, Article 101562. https://doi.org/10.1016/j.iot.2025.101562
[18] Chen, X., et al. (2025). Emotion-aware design in automobiles. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–18).
[19] Dritsas, E., et al. (2025). Multimodal interaction, interfaces, and communication: A survey. *Multimodal Technologies and Interaction*, 9(1), Article 6. https://doi.org/10.3390/mti9010006
[20] Wu, Y., Mi, Q., & Gao, T. (2025). A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions. *Biomimetics*, *10*(7), 418. https://doi.org/10.3390/biomimetics10070418
[21] Drissi, M. (2024). More is less? A simulation-based approach to dynamic interactions between biases. *arXiv*. https://arxiv.org/abs/2412.17505
[22] González, P. V., et al. (2025). A gender-aware saliency prediction system for web interfaces. *Brain Informatics*, 12(1), Article 25.
[23] Shen, J., et al. (2025). Spiking neural networks with temporal attention-guided adaptive fusion. *arXiv*. https://arxiv.org/abs/2505.14535
[24] Theresa, W. G., et al. (2025). Multi-modal emotional analysis in customer relation management. *Scientific Reports*, 15(1), Article 26437.
[25] Ortony, A. (2022). Are all "basic emotions" emotions? *Perspectives on Psychological Science*, 17(1), 41–61.
[26] Yik, M., et al. (2023). On the relationship between valence and arousal in samples across the globe. *Emotion (Washington, D.C.)*, *23*(2), 332–344. https://doi.org/10.1037/emo0001095
[27] Mahmoud, A., Scott, L., Florkiewicz, BN. (2025) Examining Mammalian facial behavior using Facial Action Coding Systems (FACS) and combinatorics. *PLoS ONE*, 20(1): e0314896. https://doi.org/10.1371/journal.pone.0314896
[28] Yang, S., & Chong, X. (2021). Feature extraction technology for real-time video based on deep CNN. *Multimedia Tools and Applications*, 80(25), 33937–33950.
[29] Shanmugam, M., Ismail, N. N. N., Magalingam, P., Hashim, N. N. W. N., & Singh, D. (2023). Understanding the use of acoustic measurement and mel-frequency cepstral coefficient (MFCC) features for the classification of depression speech. In M. A. Al-Sharafi, M. Al-Emran, G. W. H. Tan, & K. B. Ooi (Eds.), *Current and future trends on intelligent technology adoption* (Studies in Computational Intelligence, Vol. 1128). Springer. https://doi.org/10.1007/978-3-031-48397-4_17
[30] Zakaria, T. M., Langi, A. Z. R., Nazaruddin, M. S., & Anshori, I. (2025). Artificial intelligence (AI) in neurofeedback therapy using electroencephalography (EEG), heart rate variability (HRV), and galvanic skin response (GSR): A review. *IEEE Access, 13*, 133078–133112. https://doi.org/10.1109/ACCESS.2025.3582805
[31] Zhu, X., Guo, C., Feng, H., et al. (2024). A review of key technologies for emotion analysis using multimodal information. *Cognitive Computation, 16*, 1504–1530. https://doi.org/10.1007/s12559-024-10287-z
[32] Saeidi, M., et al. (2021). Neural decoding of EEG signals with machine learning. *Brain Sciences*, 11(11). https://doi.org/10.3390/brainsci11111525

[33] Espino-Salinas, C. H., Galván-Tejada, C. E., Luna-García, H., Gamboa-Rosales, H., Celaya-Padilla, J. M., Zanella-Calzada, L. A., & Tejada, J. I. G. (2022). Two-Dimensional Convolutional Neural Network for Depression Episodes Detection in Real Time Using Motor Activity Time Series of Depresjon Dataset. *Bioengineering*, *9*(9), 458. https://doi.org/10.3390/bioengineering9090458

[34] Yuvaraj, R., Baranwal, A., Prince, A. A., Murugappan, M., & Mohammed, J. S. (2023). Emotion Recognition from Spatio-Temporal Representation of EEG Signals via 3D-CNN with Ensemble Learning Techniques. *Brain Sciences*, *13*(4), 685. https://doi.org/10.3390/brainsci13040685

[35] Su, Y., & Kuo, C.-C. J. (2022). Recurrent neural networks and their memory behavior: A survey. *APSIPA Transactions on Signal and Information Processing*, *11*(1), 1–2. https://doi.org/10.1561/116.00000123

[36] Pavlatos, C., Makris, E., Fotis, G., Vita, V., & Mladenov, V. (2023). Enhancing Electrical Load Prediction Using a Bidirectional LSTM Neural Network. Electronics, 12(22), 4652. https://doi.org/10.3390/electronics12224652

[37] Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Hénaff, O., Botvinick, M. M., Zisserman, A., Vinyals, O., & Carreira, J. (2022). *Perceiver IO: A general architecture for structured inputs and outputs*. *arXiv*. https://arxiv.org/abs/2107.14795

[38] AffectNet Dataset. (n.d.). Kaggle. https://www.kaggle.com/datasets/fatihkgg/affectnet-yolo-format

[39] RAVDESS Dataset. (n.d.). Kaggle. https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

[40] Koelstra, S., et al. (2012). DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1), 18–31. https://doi.org/10.1109/T-AFFC.2011.15