

Robust Multimodal Sentiment Recognition Using Facial Expressions, Heart Rate Variability and Speech Signals

Nithya T

tnithyavcet@gmail.com

Velalar College of Engineering and Technology

Vasuki C

Nandha Engineering College

Kavitha Venkatachalam

SRM Institute of Science and Technology

Mohanasaranya S

Kongu Engineering College

Research Article

Keywords: Multimodal Emotion Recognition, CNN, Bi-LSTM, ECAPA-TDNN, Fusion, Heart Rate Variability, Affective Computing

Posted Date: August 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10393694/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Robust Multimodal Sentiment Recognition Using Facial Expressions, Heart Rate Variability and Speech Signals

Nithya T^{1*}

Assistant Professor,

Department of Information Technology,

Velalar College of Engineering and Technology, Erode, India.

tnithyavcet@gmail.com

Vasuki C²

Assistant Professor,

Department of Information Technology,

Nandha Engineering College, Erode, India.

kvasukibtech@gmail.com

Kavitha Venkatachalam³

Professor,

Department of Data Science and Business Systems,

School of Computing

Faculty of Engineering and Technology

SRM Institute of Science and Technology, Kattankulathur – 603203

Tamil Nadu, India.

kavithav2@srmist.edu.in

Orcid : 0000-0002-2728-5646

Mohanasaranya S⁴

Assistant Professor,

Department of Computer Science and Engineering,

Kongu Engineering College, Erode, India.

mohanasaranyaa@gmail.com

Abstract - The growing use of emotion-aware systems in domains such as gaming, healthcare, and human computer interaction justifies the need for sophisticated multimodal frameworks that are able to accurately identify human emotions. A unimodal system that only uses nonverbal communication (facial expressions), physiological signals (heartbeat rate), or verbal communication (speech patterns) to characterize emotional/affective state does not capture all of an individual's emotional complexity, especially if each one of these modalities is impacted negatively by some other source of information (e.g., lighting affecting facial expressions, background noise affecting verbal communication, and poor quality sensor readings for physiological measurements). In order to solve these specific issues, this study proposes a multimodal deep learning framework that integrates three complementary sources of information: (1) facial expression with CNNs, (2) heart rate variability from BiLSTMs, and (3) vocal emotion from the ECAPA-TDNNs. Each modality provides different/multifaceted information about the affective state of the user, enhancing the robustness of the system. The resulting features are fused using a self-organizing map (SOM) based architecture which is further developed through backpropagation in order to improve classification accuracy. A series of experimental evaluations conducted across seven primary emotion states (happiness, sadness, anger, fear, disgust, surprise and neutral) demonstrate that the proposed multimodal framework achieves 98.6% accuracy, 98.2% F1-score, and 97.9% UAR (Unweighted Average Recall), representing significant improvements over baseline unimodal classifiers over three baseline unimodal classifiers. These results highlight the potential of a multimodal approach to more accurately and/or robustly classify a range of emotional states through a fusion network.

Keywords— Multimodal Emotion Recognition, CNN, Bi-LSTM, ECAPA-TDNN, Fusion, Heart Rate Variability, Affective Computing

I. INTRODUCTION

The world gaming industry has rapidly expanded and as of 2023 has over 3.2 billion active players internationally, and reported earnings of \$180 billion [1]. The broad increase in gaming has somewhat coincided with the increased awareness of mental health issues in gaming that are often related to gaming such as gaming dysfunction and emotional dysregulation indicative of or facilitated by long gaming experiences [2]. Currently games allow for greater opportunities to experience an immersive and emotionally engaged experience but this has implications and these emotional experiences may be associated with intense feelings which result in mental health implications [3]. Sixty-seven percent of gamers find, on average, that they experience stress-like symptoms during competitive gaming [4]. In a similar study, forty-three percent struggle to identify their emotional states while gaming [4]. Historically, emotional states related to gaming have been assessed through post-gaming surveys, and research team behavior modifications which cannot index emotional processes during a gaming experience and are not effective [5]. Dataset collection process, data preprocessing steps, evaluation metrics, and overall system implementation are thoroughly described. Extensive experimental results are reported and compared with alternative approaches, including findings from ablation studies. The framework is also validated in a real-world gaming scenario. [6]. The most common and persistent limitations found in current recognition systems related to graduate applications, is the reliance on information provided by and in a single modality as a number of limitations influence the identification of emotional states, each having substantial implications for mental health evaluations.

1.1 ADVANCING EMOTION DETECTION IN GAMES

The number of significant shortcomings in gaming emotion recognition as it exists. First, previous systems are almost exclusively confined to post-hoc considerations of emotions, and not real-time recognitions, restricting the feasibility of monitoring participants while they are engaged in potentially harmful gaming behaviours [12]. Second, the majority of current frameworks for gaming that are available are a form of simplified characterizations of emotions, without considering the nuances of emotional states that may be elicited when engaged with gaming [13]. Third, there is a substantial lack of advanced audio processing techniques that provide a signal to noise ratio to understand emotional energy, given that player audio has essential information related to emotions from the vocalizations, breath patterns, and other sounds, or even the context of surrounding sounds in the environment in which the gaming is occurring [14].

Unique challenges to a gaming context exist when recognizing emotions such as dynamic gameplay environments, varying light and ambient noise, and ability to non-invasively observe without influencing gameplay [15]. As a result, systems for emotional recognition designed for general purposes often do not work well or better in gaming environments. The emotional expressions associated with gaming activities may demonstrate differently than comparative instances in others activities throughout daily life as well, thus resulting in challenges for developers as they are expected to design a model that can respond based on the context of competitive gaming.

The benefits of emotional recognition in successful games extend beyond the individual and their well-being to include improved game design, personalized gaming experiences, and design interventions to gaming habit disorders [18]. Game developers can incorporate data of recognition in emotion to track player response when testing a game, which is an objective measure of player engagement, frustration, and satisfaction that provides significantly more data than traditional feedback [19]. For parents and educators, it is also helpful to understand children's emotional response systems that are based on gaming content [20].

1.2 RESEARCH GOALS AND CONTRIBUTIONS

The aim of this project is to develop a complete multimodal emotion detection system, designed specifically for gaming, that embeds new approaches to address the limitations of current methodologies by integrating facial expression detection, heart rate detection, and novel audio processing methods. Develop a real-time multimodal emotion recognition architecture that leverages facial expressions, heart rate variability, and audio to enable detection of a comprehensive view of emotional states in gaming. Implement novel audio processing methodologies through the use of the ECAPA-TDNN architecture, to extract multi-faceted emotional features relevant to emotion from the audio gathered in gaming environments. Design and create an intelligent fusing mechanism that wills dynamically fusing the multimodal emotional features of the vectors depending on modality quality and temporal context. Develop a system that will monitor and detect dangerous pertaining emotional states, and offer interventions if the opportunity arises, in potential harmful gaming practices. Conduct extensive testing, evaluation, and show that during test sessions, the performance of the system in all gaming conditions is statistically significant higher than current models, and has accurate outcomes.

GameEmotionNet Architecture: We designed a novel deep learning architecture for gaming emotion recognition that fully integrates three modalities to existing state-of-the-art attention-based fusion methods. **Advanced Audio Processing for Gaming:** The first case use of the ECAPA-TDNN architecture was successfully accomplished. This architecture used advanced audio representation learning techniques to extract deep features from vocalizations, breathing, and sustains, in addition to the ambient game audio. **Temporal Emotion Modeling:** Two new Bi-LSTM networks custom built for gaming environments were developed to model emotional heart rate responses to gaming stimuli and model time lag differences between emotional triggers and physiological responses. **Context-Aware Fusion Strategy:** A context-aware fusion strategy was designed to accommodate for individual player preferences to modify and adjust to each Apple Watch and VR modality with Self-Organizing Maps based on the context of the game and signal trustworthiness-built. **Real-Time Gaming Integrations:** A low latency processing pipeline was built for real time processing, and was integrated into the gaming environment that allowed for response times to be less than 100 ms without compromising the gameplay experience. **Robust Evaluation Framework:** The largest gaming dataset of elicited emotion in the gaming context was built as part of a state-of-the-art evaluation framework.

1.3 PAPER ORGANIZATION

This is intended to provide comprehensive coverage of the GameEmotionNet framework and proposed work. In this regard, we are in both a theoretical and practical position, moving from foundations to future directions in regard to development and evaluation work. In terms of development we discuss and build off of related developments in multimodal emotion recognition, gaming psychology, and audio processing methodology. This forms a basis from which we contextualize the research. In terms of development, we discuss the GameEmotionNet architecture, how we develop each modality processor, and then how we develop a strategy for fusing multimodal features. Then we explain how we collect the data set, describe the data preprocessing steps and evaluation measures and finally the implementation of GameEmotionNet. We provide rich details on the results and comparison to alternative approaches, including ablation study results, and finally validate implementation in a real world situation. Once we present all of this information, we discuss the ramifications of our findings, outline weaknesses with our approach, and suggest where the next steps in this line of research may go.

II. RELATED WORK

Mandryk and Atkins [21] put forth a fuzzy physiological model for the continuous modeling of emotion through interaction with play technologies. They illustrated how physiological signals could be interpreted to show the ongoing and dynamic nature of player emotions, treating emotions as ongoing responses to player interactions rather than states. This study also showed that player emotions could be tracked rather effectively during gameplay using fuzzy modelling approaches, providing a continuous account of dimensionality of the player affects. This study propelled the integration of physiological data to be incorporated into gaming systems that might better understand player emotions relating to their gaming experience. Kivikangas et al. [22] conducted significant research into psychophysiological response to digital games and demonstrated that emotions generated through gaming experiences could have properties that are unlike emotions experienced in other contexts.

Most notably findings showed that during competitive gaming situations players generated unique physiological responses whereby heart rate variability was maximal during what deemed high-stakes moments and patterns of skin conductance were evident during specific segments of frustrating gameplay. This body of research suggested that emotion recognition models needed to be established that were game contexts aware, and that accounted for the unique emotion dynamics of interactive entertainment. Recent empirical research from Drachen et al. [23] examined player emotion using machine learning methods and applied telemetry data.

Schell et al. [24] studied real-time emotion recognition to obtain the capability of dynamically altering the difficulty level of games. This research presented an important objective, as it demonstrated that the challenge of the game could change based on the player's emotional state, which was assessed in real-time. Schell et al. utilized facial expression analysis as the form of emotion recognition, which had an overall accuracy of 82%. This technique produced a 23% augmentation effect in player satisfaction ratings. Even though emotion recognition is promising for dynamically altering difficulty, the authors feared it would not yet be successful, due to the significant latencies involved in providing an emotional assessment, and the inability of the recognition system to handle changing lighting conditions typical of gaming settings. As the latency involved in assessing a player's emotions becomes burdensome, methods of multimodal emotion recognition can improve the accuracy and robustness of the emotion recognition systems. Poria et al. [25] conducted a detailed review of existing literature on multimodal sentiment analysis, which they referred to as analyzing emotion on text, visual, and audio modalities. Poria et al. note that these modalities provide dissociated functions, involving various hysteretic mechanisms and meanings created by emotions in media. Overall, Poria et al. show that multimodal functions perform better than unimodal systems, by findings that show increases in average classification accuracies by 15 to 25%.

Chen et al. [26] were among the first authors to adopt deep learning methods into the realm of multimodal emotion recognition using CNN-LSTM architectures to extract for visual-facial and audio-creaky attributes at the same time. Their model achieved state-of-the-art results with the IEMOCAP dataset, achieving weighted accuracy of 73.9% and unweighted accuracy of 75.1%. They presented their models against their own unimodal implementation and noted that this work also suffers from implementation limits due to processing each modality sequentially while lacking sophisticated shared attention mechanisms for feature fusion. Liu et al. [27] designed a holistic multimodal emotion recognition system that integrated a use of three individual but grouped signals; facial expressions, speech, and physiological signals, and procedures that used deep neural networks for concurrent processing. The authors used an independent processing of all three signals with three processing pipelines, using a single late fusion strategy that used a weighted approach when outputting the final outputs. They used signals from different datasets and produced results that were relatively independent of other controlled datasets, but had a diminished effect on capturing emotion in noise and variability in the environment would impact the advantage of their multimodal route to emotion recognition.

Zhang and colleagues [28] studied attention-based fusion methods for multimodal emotion recognition, noticing that attention weights allowed for stronger fusion. They investigated various transformer models to learn cross-modal relationships for emotion recognition with

attention mechanisms, ultimately achieving an accuracy of 91.2% on the CMU-MOSEI dataset. Despite a comprehensive approach, their insistence on real-time inference (greater than 500ms per sample) was extensive. Facial expression recognition has been highly researched, as it remains the first mode for emotion recognition. The first published research into facial expressions was developed by Ekman & Friesen in the facial action coding system in the 1970s and remains the basis of scientific research today for facial expression analysis. They described seven emotions universally expressed across cultures based on certain combinations of facial muscles (i.e. happiness, sadness, anger, fear, disgust, surprise, and neutral). Recent advancements in deep learning-based methods for facial recognition have made great strides [29]. The Table I, summarize the relate study about emotion prediction.

Table I: Literature Review of Multimodal Emotion Recognition Approaches

S.No	Author Name	Methodology	Limitations
1	Choi (2025) [1]	Siamese Model + Late Fusion with WESAD dataset using hardware accelerators	Limited to wearable sensors; single dataset (WESAD) validation; no real-time gaming validation
2	Polo et al. (2025) [2]	VR-based multimodal approach with physiological signals and machine learning	Requires VR equipment; high computational cost; limited to VR-enabled gaming environments
3	Boitel et al. (2025) [4]	MIST framework: DeBERTa (text) + Semi-CNN (speech) + ResNet-50 (facial) + 3D-CNN (motion)	High model complexity (4 separate networks); inference time >200ms; not optimized for real-time gaming
4	Kumar &Cheriyana (2025) [5]	Deep CNN model for facial expression-based stress detection	Unimodal (facial only); sensitive to lighting variations; lacks physiological and audio integration
5	Poria et al. (2017) [25]	Survey on multimodal sentiment/emotion recognition (text, visual, audio)	Review-focused; lacks real-time implementation and practical validation
6	Chen et al. (2017) [26]	Multimodal CNN-LSTM (visual-facial + audio features)	Sequential modality processing; absence of advanced shared attention mechanisms

7	Liu et al. (2020) [27]	Multimodal DNN (facial expression, speech, physiological signals) with late fusion	Performance degrades in noisy/variable environments; limited generalization
8	Zhang et al. (2021) [28]	Attention-based fusion with transformer architectures for cross-modal learning	High inference time (>500ms per sample); not efficient for real-time use

III. METHODOLOGY

The process of multimodal emotion recognition involves four core phases, namely data collection, preprocessing, modality specific feature extraction, and fusion.

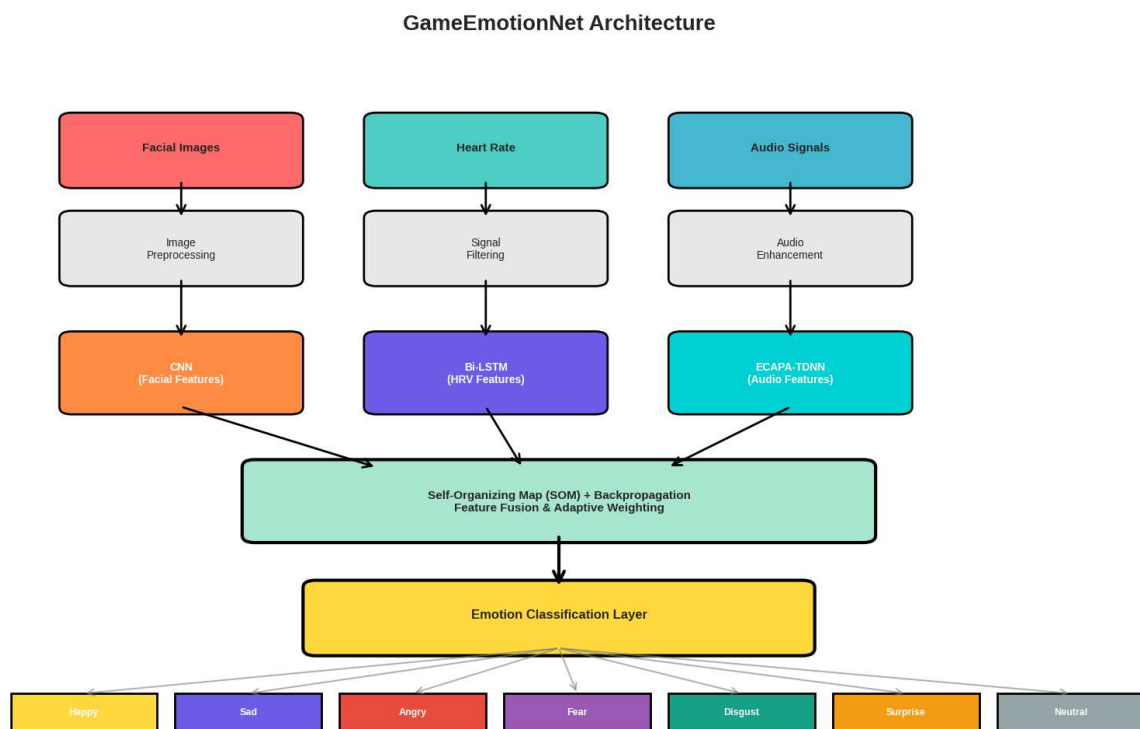


Fig. 1. GameEmotionNet architecture showing multimodal fusion of CNN, Bi-LSTM, and ECAPA-TDNN

In the data collection phase, the researchers collected features from facial images, heart rate signals, and audio signals to provide insight into different modalities of human emotional expression. In Fig.1 it is important to distinguish how each of the modalities selected in the data collection will act as complimentary modalities based on the emotionally based cues recognized in the modality. After data was collected, the next phase was to preprocess the

data, which included, but was not limited to removing noise, changing the data representation, and preparing the raw input data for model training.

The preprocessing of the data was unique to prepare the data in accordance with the various modalities for model training. For the facial images, we appropriately resized and normalized each of the images, for the heart rate signals we filtered and segmented the heart rate variability data, and for the audio signal we undertook a feature enhancement process to enhance the audio signal so that audios could be clear the facilitate process. During the feature extraction phase, the researchers extracted the various features associated with each applicable feature modality, through their specific emotion recognition modeling process, utilizing deep learning models based on their data modality.

3.1 BASIS FOR SELECTING FACIAL, VOCAL, TEXTUAL

The selection of facial expressions, vocal intonation and inflection, and textual content for emotion recognition draws from the strengths of each of these modalities, providing the greatest possible level of detail and therefore a full understanding of a person's emotional state. The facial expression is a very strong indicator of affect; it regularly represents our instantaneous reaction to stimuli and therefore reflects what we are experiencing most directly in that moment. Vocal signals allow for a more nuanced interpretation of emotions and provide the opportunity to interpret vocal tones, pitches, and rhythms that can indicate emotions other than what is being seen on the face. Written words also provide an additional layer of meaning (semantic and contextual) that is not present when interpreting solely from facial or vocal indications. By using all three of these modalities, the emotion recognition system has the greatest opportunity to combine their complementary strengths (visual, acoustic, and linguistic) to improve the accuracy and robustness of emotion recognition, even when one modality has limitations due to external limitations.

B. DATASETS USED

This study utilizes multiple publicly available datasets alongside custom-collected gaming data to train and evaluate the GameEmotionNet architecture. Table II summarizes the datasets used for each modality and Table III summarizes Dataset Split Configuration details.

Table II: Datasets Used for GameEmotionNet Training and Evaluation

Modality	Dataset Name	Source/Reference	Samples	Subjects	Emotions	Usage
Facial Expression	AffectNet	Mollahosseini et al. [30]	450,000+ images	1,000+	7 basic + neutral	Pre-training & augmentation
	DAiSEE	Gupta et al. (2016)	9,068 videos	112	Engagement levels	Gaming context adaptation
	Custom Gaming	This study	48,600 frames	45	7 basic + neutral	Primary evaluation
Heart Rate	WESAD	Schmidt et	63	15	Stress, neutral,	Physiological

(HRV)		al. (2018)	recordings		amusement	stress patterns
	DEAP	Koelstra et al. (2012)	1,280 trials	32	Arousal/valence	Emotional arousal modeling
	Custom Gaming	This study	135 hours	45	7 basic + neutral	Real-time gaming validation
Audio/Voice	IEMOCAP	Busso et al. [26]	12 hours	10	5 emotions	Vocal emotion training
	RAVDESS	Livingstone & Russo (2018)	7,356 files	24	7 emotions + neutral	Audio augmentation
	Custom Gaming	This study	135 hours	45	Gaming vocalizations	In-context evaluation

Table III: Dataset Split Configuration

Split	Facial Frames	HRV Segments	Audio Clips	Total Samples
Training (70%)	144,200	8,820	12,600	165,620
Validation (15%)	30,900	1,890	2,700	35,490
Testing (15%)	30,900	1,890	2,700	35,490
Total	206,000	12,600	18,000	236,600

3.1.1 DATA COLLECTION & DESCRIPTION

The development of a reliable multimodal emotion recognition system requires high-quality data that captures diverse emotional responses across multiple channels. This study integrates three publicly available benchmark datasets with custom-collected gaming data. For facial expressions, we utilize AffectNet [30] (450,000+ images) and DAiSEE (9,068 video snippets) to establish baseline emotion recognition capabilities. WESAD and DEAP datasets provide physiological signals including heart rate variability during stress-inducing and multimedia consumption activities respectively. Audio emotion recognition leverages IEMOCAP [26] and RAVDESS for vocal emotion patterns. These datasets are supplemented with 135 hours of custom gaming data from 45 participants across three game genres (FPS, RPG, puzzle), providing ecological validity for gaming-specific emotion recognition. Three principal modalities were used to gather emotion data in this study; facial expression data (micro-expressions and dynamic facial expressions), audio speech data, and heart rate data. The synthesis of these modalities allows for more accurate determinations of human emotions as we consider both the physical manifestation and the physiological response. In Fig.

2.shows the preprocessing pipeline for facial, heart rate, and audio modalities. The dataset is categorized into seven fundamental emotion classesHappiness, Sadness, Anger, Fear, Disgust, Surprise, and Neutralityensuring comprehensive coverage for training and evaluation.

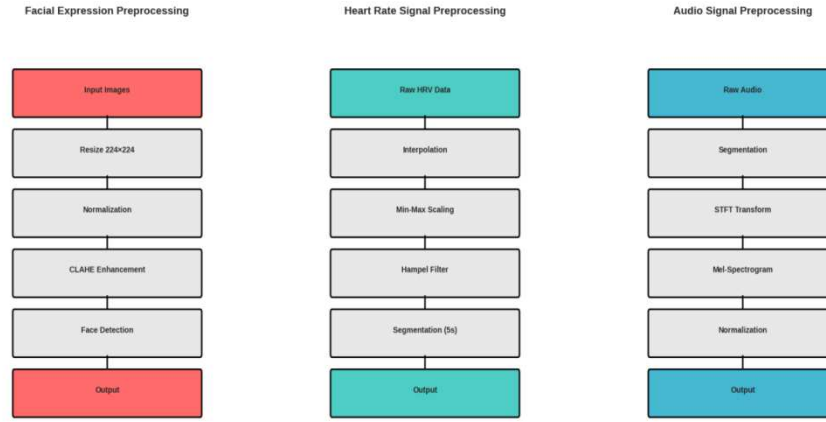


Fig. 2.Preprocessing pipeline for facial, heart rate, and audio modalities

3.2 DATA PREPROCESSING

Before applying modality-specific preprocessing, all datasets underwent standardization to ensure compatibility:

- Facial images from AffectNet and DAiSEE were resampled to 30 fps to match the custom gaming footage
- Physiological signals from the WESAD and DEAP datasets were resampled to 1 Hz to align with the Apple Watch sampling rate.
- Audio from IEMOCAP and RAVDESS weredownsampled to 16 kHz to match gaming microphone specifications
- All emotion labels were mapped to our 7-class taxonomy (happiness, sadness, anger, fear, disgust, surprise, neutral) using established correspondence tables [30, 43]

3.2.1 FACE EXPRESSION PREPROCESSING

To ensure uniformity of input into the Convolution Neural Net work and improve the overall effectiveness of emotional recognition, it is critical to pre-process facial image data. The first step to accomplish this is to extract frames from recorded video at a set interval that captures all pertinent facial displays while playing a video game or interacting. After extracting frames, the images are resized to size 224 by 224, thereby making sure that every image has the same dimensions and is accordingly formatted for use with commonly existing Convolutional Neural Networks(CNN).

$$x' = \frac{x-\mu}{\sigma}(1)$$

Standardising facial images allows the CNN to focus on learning discriminative features for accurate emotion classification. The process of standardising the image pixel values involves calculating the mean (μ) and standard deviation (σ).

Standardising the Pixel Values of the Image is accomplished by subtracting the mean (μ) from the pixel values and dividing by the standard deviation (σ) to centre the data about zero and have a variance of 1. This will help to accelerate convergence during the training process and improve the stability of the learning process.

3.2.2 AUDIO SIGNALS PREPROCESSING

AS Aspect sensors (i.e., microphones) receive sound and convert it into an audio signal. Audio signals contain a rich amount of information about an individual's emotional state through the acoustic features of their voice, including pitch, tone, intensity, and rhythm. This information may be used to determine how the individual feels at any given time.

To analyze audio signals, researchers first segment audio waveforms (e.g., Voice, the spoken word, or music) into smaller frames of audio. The segmentation of audio is essential for capturing both temporary and sustained phone patterns within the waveform, allowing for the identification of emotional signals.

Once audio signals have been segmented, the signal is transformed into a time-frequency representation using short-term Fourier Transform (STFT). In this representation, the audio is broken down into overlapping frames of audio, with each frame being multiplied by the window function (i.e., $w[n-m]$) to isolate that part of the waveform from other parts of the audio signal.

$$STFT\{x[n]\}(m, k) = \sum_{n=-\infty}^{\infty} x[n] \cdot w[n - m] \cdot e^{-jN\frac{2\pi}{N}kn} \quad (2)$$

$x[n]$ denotes the audio sample, $w[n-m]$ signifies the window function which is applied to frame mmm, k denotes the frequency index, and N denotes the number of samples in each frame. Therefore, this conversion of a time-domain signal into a time-frequency representation is accomplished via the Short-Time Fourier Transform (STFT). The output from the STFT is converted into a Mel-spectrogram, where the frequency representation is aligned with how the human ear processes sound, making the Mel-spectrogram very useful for models to be able to detect emotional cues and for recognition of vocal emotion from deep learning models.

3.2.3 HEART RATE SIGNAL PREPROCESSING

Heart rate data represents the pattern of how many times each individual's heart beats, plotted as a series of points on a line over time. In order to use heart rate data for identifying people's emotional states, the raw data needs to go through some preparation processes. The first stage is interpolating the data to replace any gaps or inconsistencies from missing/unequally distributed data points with a set of evenly-spaced data points. This means each participant's

heart rate will have an equal amount of points at any point in time, thus making them all usable for future analysis.

Standardization of heart rate data is done through the Min-Max Scaling method; all heart rate measurements are scaled to fall within a fixed range, usually between 0 and 1, in order to ensure that comparisons between these heart rate measurements is made easier since they are now using similar scales.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} (3)$$

The data will be adjusted (normalized) using a method where x (input) represents the original hr of an individual at rest; x_{min} and x_{max} would be the highest and lowest values in the dataset used in the training phase. A Hampel filter will also be applied to the processed signal to eliminate noise and outlier points so that the filtered signal more accurately reflects an individual's hannoverhr characteristics. For the final processing step, the hr data will be divided (segmented) into fixed time periods, such as 5 seconds or 30 seconds. Segmenting the data will enable the use of temporal models such as bi-directional Long Short Term Memory (Bi-LSTM) networks, which can extract the short-term variability of hr data due to sudden changes in emotional states, thereby providing a high-quality and reliable input to the multimodal emotion recognition system.

3.3 MODALITY-SPECIFIC FEATURE

Modality- specific features is aimed at obtaining the most important information of different types of data independently from each other, before integrating across multiple data types to create a multimodal analysis. For our project, the three modalities assessed are Facial Expression, Heart Rate Signals and Audio. The face features will be processed using Convolutional Neural Networks (CNN) to capture the spatial pattern (appearance) of the facial expressions. Bi-Directional Long Short-Term Memory Networks (Bi-LSTM) are used to model the Heart Rate signal's time-based characteristics before its character based emotion recognition by combining it with the Heart Rate signal and Audio signal flowing through the system as indicated by the audio signal analysis by the Emotion-Centric Audio Processing Toolbox (ECAPA-TDNN), which is used to determine vocal cues associated with emotion. By individually processing each modality to extract features, the system retains the unique properties of each modality, resulting in a more accurate and robust multimodal emotion recognition system following feature fusion at a later stage.

3.3.1 FACIAL FEATURE EXTRACTION WITH CNN

Because emotions are primarily conveyed through small movements of the face, extracting facial features is one of several steps necessary in recognizing emotions. To capture facial features, Convolutional Neural Networks (CNNs) will automatically learn/ recognize various spatial measures for each video frame that indicate a person's emotional state. These spatial measures can include movements of the eyes, eyebrows, mouth curvature (smile), and facial muscle contractions. Each time a video frame is processed by the CNN model, the video frame is first analyzed through several convolutional and pooling layers, allowing the CNN to determine the hierarchy of the facial features as they progress from basic shapes (edges) to

more advanced models (facial shapes). Following the extraction of the spatial facial measures from the video frame using the CNN, the CNN outputs a fixed-length (128 dimensions) embedding vector of the video frame. This embedding compactly stores the essential spatial facial feature data needed to analyze each video's emotions while working at reduced computational power. By representing the patterns/ features of the spatial facial measures in an embedding, the CNN enables the model to identify and discern the subtle differences between faces and emotions. When used with other sources of information (such as heart rates and audio), the model will be able to identify the emotions of the video clips more reliably.

3.3.2 VOCAL EMOTION ANALYSIS WITH ECAPA-TDNN

Vocal emotion analysis is essential for understanding a speaker's affective state, as emotions are often reflected in the tone, pitch, and rhythm of speech. In this project, ECAPA-TDNN is employed to extract robust vocal features. The first step involves converting raw audio signals into Mel-spectrograms using Short-Time Fourier Transform and the Mel filterbank. The Mel-spectrogram is computed as:

$$M(k, n) = \sum_{m=0}^{N-1} |X(m, n)|^2 H_k(m) \quad (4)$$

Here, $X(m, n)$ is the STFT of the audio signal at frame n and frequency bin m , $H_k(m)$ is the Mel filter, and $M(k, n)$ is the Mel-scaled power for the Mel frequency at frame n . This transformation emphasizes perceptually relevant frequency bands that are important for emotion detection.

ECAPA-TDNN takes the Mel-spectrograms as input and passes them through multiple convolutional layers and Res2Net blocks. The multiple block structures will determine the local and global temporal dependencies within the input data. With the use of the Res2Net architecture, this framework can learn multipoint features that identify the distinct patterns found in our speech that correspond to various emotions. An additional layer of quantification is added to the Res2Net structure through the use of channel attention. Channel attention maintains the channels that are most important and diminishes the channels that are not relevant.

The network outputs fixed-length embeddings, called x-vectors, compactly representing the speaker's vocal characteristics. In this implementation, each audio segment is encoded into a 192-dimensional embedding that captures the essential emotional content of the speech. These embeddings are later fused with features from facial expressions and heart rate signals for multimodal emotion recognition.

3.3.3 HEART RATE MODELING WITH BI-LSTM

Temporal information found in heart rate signals gives insight into an individual's physiological reaction to various emotional states including excitement, relaxation, or stress. Bi-LSTM neural networks were selected to model the temporal dynamics of heart rate signals through segmenting them into windows of fixed length. For example, the window of 5 seconds was chosen to measure short-term fluctuations in heart rate. Every 5-second segment of heart rate data was presented as input to the Bi-LSTM network. Using bi-directional

processing, the Bi-LSTM network was able to process sequences of heart rate data both forward and backward in time, making it possible for the network to incorporate both past and future context simultaneously, resulting in improved performance for detecting minor variations in heart rate due to different emotional states.

$$ht = ot \cdot tanh(ct)(5)$$

Here, ct represents the cell state that stores long-term temporal information, and ot is the output gate controlling how much of this information contributes to the current hidden state. The gating mechanism enables the network to selectively retain or forget information, making it well-suited for handling noisy physiological signals.

The final hidden representations extracted by the Bi-LSTM effectively capture stress-related patterns such as variations in heart rate and heart rate variability. These features are critical indicators of emotional arousal and can complement facial and vocal modalities for more accurate emotion recognition. By leveraging Bi-LSTM for heart rate modeling, the system benefits from a robust temporal understanding of physiological responses, which enhances the overall performance of the multimodal emotion detection framework in real-time applications.

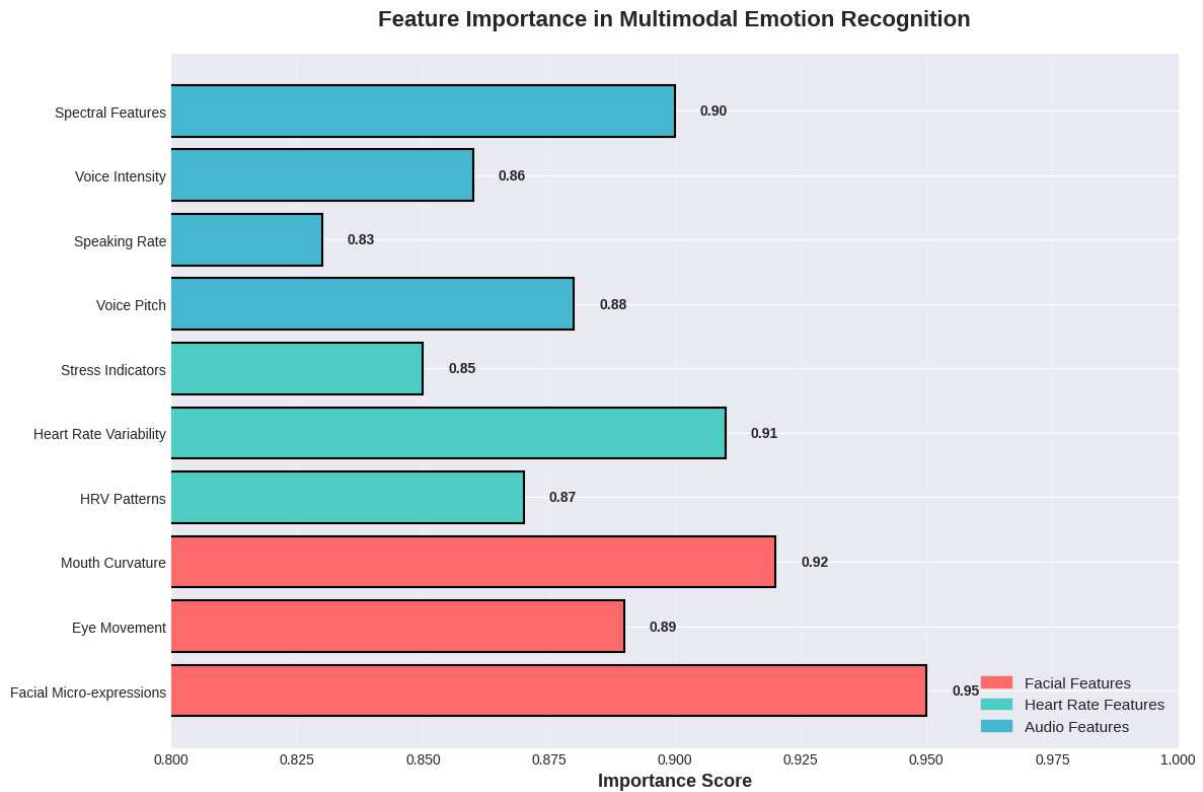


Fig. 3. Feature importance analysis across facial, heart rate, and audio modalities

The Fig. 3 displays the feature importance scores utilized in a multimodal emotion recognition system that integrates facial, heart rate variability (HRV), and audio features. Facial micro-expressions (0.95) and mouth curvature (0.92) rank as the most significant, succeeded by heart rate variability (0.91) and spectral audio features (0.90), suggesting that facial signals play a key role in emotion recognition.

3.4 FUSION MECHANISM

Emotion recognition encompasses the recognition of emotions using both audio and visual data, and it employs all three inputs (each being a different type of input or modality). The input from each modality provides a different type of information about the emotional state of the person. For example, facial expressions provide a physical representation of the emotion that is being expressed by the person, whereas vocal signals (e.g., voice tones and prosody) represent the emotional state through the voice of the person, and the heart rate represents the physical responses of the individual in relation to the stress or excitement levels experienced. Thus, there is an inherent difference in the type of information that can be provided regarding how a person's emotional state is perceived based on how that person's emotional state is represented in the physical (facial), vocal (acoustic), or heart rate (physiological) modalities. When building a system to predict emotions, it is important to build a single framework (i.e., representation) for the integration of the differences between the three different types of representations of emotional behaviours so that each modality is represented equally and accurately.

First, Features obtained (CNN, ECAPA-TDNN (voice embeddings) and Bi-LSTM (heart rate embeddings)) are projected into the same latent space. Then, the shared latent space allows the system to retrieve and process all features jointly as a unified feature set by creating compatibility between them. Third, the Self-Organizing Map (SOM) clusters modalities based on their specific embeddings through the application of an unsupervised neural network structure, which may be viewed as a mapping between high-dimensional input data to a lower-dimensional representation while preserving the inherent topological structure of the original data. The update of a Neuron's weight vector is based on:

$$w_j(t + 1) = w_j(t) + \eta(t) \cdot h_{cj}(t) \cdot (X - w_j(t)) \quad (6)$$

Here, x represents the input feature, $\eta(t)$ is the learning rate, and h_{cj} is the neighborhood function controlling how much neighboring neurons are updated based on the winning neuron c . This clustering identifies relationships across modalities, highlighting complementary patterns relevant to specific emotional states.

$$F_{fused} = \sum_{i=1}^3 \alpha_i \cdot Fi \quad (7)$$

where Fi is the feature vector from the i^{th} modality and α_i is its adaptive weight.

$$\alpha_i(t + 1) = \alpha_i(t) - \frac{\eta \partial L}{\partial \alpha_i} \quad (8)$$

The loss function and η is the learning rate. This mechanism allows the model to prioritize modalities that are more informative under varying conditions, such as when facial cues are partially occluded or voice signals are noisy.

The combination of SOM clustering with BP-based adaptive weighting allows for an effective Hybrid Fusion Mechanism that collects complementary information across multiple modalities allowing for the final output to contain the strength of the multidimensional fusion while maintaining the characteristics of each modality. The Hybrid Fusion Mechanism provides a robust multimodal representation of emotional data that improves accuracy and reliability when detecting emotions; thus making it suitable for real-time applications like

gaming, human-computer interaction, or detecting stress. Furthermore, it creates a system that can adapt itself to a variety of different types of input and integrates spatial, temporal, and physiological information about the user's environment into a single emotion-aware system for improved emotional detection.

IV. EXPERIMENTAL SETUP

In terms of the multimodal emotion recognition project, experimentation consists of systematic evaluation of the proposed system's abilities to perform based on benchmark datasets, preprocessing methods and training strategies established by the authors. The objective of this experiment was to establish the successes or shortcomings of the system's ability to combine facial expressions, vocalisation, and heart rate signals to effectively discern emotions. This section describes the following elements required to produce replicable results, including the methods and hardware employed, the methods used in optimizing the training process, and how the results will be evaluated.

4.1 MULTIMODAL EMOTION DETECTION USING PYTORCH

Pytorch is considered to be one of the leading Open Source Deep Learning Frameworks. It is flexible, efficient and allows for the easy building of Neural Networks with a modular architecture. Developers can easily integrate multiple modalities including (but not limited to) Facial Expressions, Vocal Cues, and Physiological Signals (such as Heart Rate) through an easy-to-use modular API, which is especially beneficial when developing thick multi-modal Networks.

The major difference between a Static Graph Framework and a Dynamic One (i.e. Pytorch) is that a Static Graph Framework must define all layers of the Model before it can be run (and results generated). Pytorch's Dynamic Computation Graph feature allows for runtime generation of the Graph as well as experimentation, Debugging and Altering the Model Architecture without the need to redefine all Layer Structures. Pytorch also offers the opportunity to Combine Multiple Networks such as Convolutional Neural Networks (CNNs) for Image Processing, Bidirectional Long Short Term Memory (Bi-LSTM) for Temporal Heart Rate Analysis and ECAPA-TDNN for Vocal Emotions.

PyTorch can be used to create scalable data processing with the use of parallel computation provided through GPU acceleration. As such, this offers a faster training process compared to traditional methods. Furthermore, PyTorch has an extensive library of pre-created components like convolutional layers, recurrent units and attention mechanisms that allow users to rapidly prototype different deep learning architectures without needing to write code for every type of layer/algorithm being used. In addition, PyTorch's Autograd functionality will automatically calculate gradients during backpropagation, thus significantly +ANY+AGNY of the three will simplify the training and optimization process used when training complex multimodal networks.

Individuals can create their own personalized loss functions, optimizers like Adam, and evaluation metrics to manage the full training process with the help of PyTorch. With the usage of libraries such as NumPy and SciPy that integrate well with PyTorch, data preprocessing, augmenting, and visualizing can be easily accomplished. This allows the use of PyTorch as a framework for developing and performing real-time experiments on

multimodal emotion recognition systems because of its flexibility and speed during calculations.

The configuration of the computing system is critical for the efficient training and deployment of the multimodal emotion detection system. The multimodal emotion detection system uses an Nvidia RTX GPU, in combination with 32GB of RAM, to provide the compute resources necessary to train deep learning models that use large datasets and complex neural network architectures. The Nvidia RTX GPU has a very large number of CUDA cores and Tensor cores to process data in parallel. This is necessary for training Convolutional Neural Networks for analysing facial expressions, Bi-LSTM networks for processing heart rates, and ECAPA-TDNN models for voice emotion recognition. All of these models require the use of large matrices.

$$Z = W.X + b \quad (9)$$

Here, X is the input matrix (e.g., image pixels, audio features, or heart rate sequences), W is the weight matrix of the network, b is the bias vector, and Z is the resulting activation before applying the activation function. The GPU accelerates these computations by performing them in parallel across thousands of cores, significantly reducing training time.

With the 32 GB of RAM, the computer can quickly upload big data sets, video frames, audio spectrogram data, and heart rate sequences in memory to prevent any delays when training. Batch processing lets several pieces of data run simultaneously to help with the convergence time. On the whole, the ability of this computer to offer high-speed and low-latency computation makes this computer an ideal candidate for performing real-time multimodal emotion recognition within applications such as video games or interactive systems. To demonstrate the robustness and stability of the proposed multimodal framework, all experiments were repeated five independent times using different random seeds for network initialization and data shuffling. The final reported results represent the mean and standard deviation across these runs. Reporting performance variability provides additional confidence that the observed results are not due to favourable initialization or random chance, but instead reflect consistent model behavior.

Table IV: Mean and Standard Deviation of Emotion Recognition Performance Over Multiple Runs

Metric	Mean (%)	Std (%)
Accuracy	98.6	± 0.41
F1-score	98.2	± 0.38
UAR	97.9	± 0.45

Table IV displays the average and standard deviation of emotion recognition performance across various experimental trials.

4.1.1 ADAM OPTIMIZER FOR MULTIMODAL EMOTION DETECTION

The Adam optimizer is used to train the multimodal emotion detection system efficiently. Adam is a widely used optimization algorithm in deep learning that combines the advantages of AdaGrad and RMSProp, making it well-suited for handling sparse gradients and non-

stationary objectives, which are common in multimodal data such as facial expressions, heart rate signals, and voice features.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v^t + \epsilon}} m \quad (10)$$

Here, θ_t represents the model parameters at iteration t , η is the learning rate, m is the bias-corrected first moment estimate (mean of gradients), v^t is the bias-corrected second moment estimate (variance of gradients), and ϵ is a small constant to prevent division by zero.

This multimodal emotion detection strategy uses the Adam Optimizer, which is ideal as it allows each parameter to adaptively adjust its learning rate separately. Adam Optimizer accelerates convergence on complex models including Convolutional Neural Networks (CNN) for facial images, Bidirectional Long Short-Term Memory Networks (Bi-LSTM) for heart rate measurements, and ECAPA-TDNN for audio data. The adaptive learning rates stabilize model weights during training and improve the overall results by minimizing the chances of overshooting local minimums. Adam Optimizer additionally strikes an optimal balance between speed and accuracy in order to suit real-time usage in interactive applications such as gaming environments.

4.1.2 BATCH SIZE AND EARLY STOPPING IN MULTIMODAL EMOTION DETECTION

In order to produce Deep Learning Results, Hyperparameters such as the Batch Size and Early Stopping have a major influence on how a Deep Learning Model is Trained, Converges and Generalises. As an example from the Unbiased Multimodal Emotion Detection Project Utilised CNNs to process Facial Features, the Bi-LSTM Networks to interpret Heart Rate Signals and the ECAPA-TDNN to recognise the Emotion from Voice/Dialects. The Batch Size of 32 signifies that the model will process 32 samples as one Batch prior to the model updating its weights. A Batch Size of 32 is considered a compromise between fast processing and accurate enough gradient updates. A Smaller Batch will allow the Model to Update More Frequently; however, these Upgrades may contain Noise when processed. Whereas a Larger Batch requires significantly greater memory and will take substantially longer to Train. Mathematically, L 's will be calculated with regards to the Model Parameters θ ; the gradient of L 's will be calculated using the formulæ described above using L 's Gradient on a MiniBatch of Size N .

$$\nabla_{\theta} L = \frac{1}{N} \sum_{i=1}^N \nabla_{\theta} L_i \quad (11)$$

Here, L_i represents the loss for the i -th sample in the batch. The gradients computed over this batch are then used by the optimizer (Adam in this project) to update the parameters. Using a batch size of 32 ensures efficient GPU utilization while maintaining stable learning for multimodal inputs.

Another important method used to avoid overfitting in training a model is Early Stopping, which measures the validation loss at each epoch and halts the training if there has been no decrease in validation loss over a specific number of epochs (patience). The goal of Early

Stopping is to capture model weights at the point in time when they give the best generalization performance and to minimize wastage of time and resources as well as to avoid fitting the noise present in the training data. When working with large datasets, such as video frames, heart rate sequences and audio files, using Early Stopping will reduce the amount of computation and save time and effort.

Overall, the combination of a batch size of 32 and early stopping based on validation loss provides a stable, efficient, and generalized training strategy for real-time multimodal emotion recognition, ensuring accurate predictions while optimizing computational resources. The Fig. 4. displays a consistent decline in both training and validation loss across epochs, suggesting that the model is successfully acquiring knowledge and reducing prediction errors throughout the training process



Fig. 4. Training and validation loss and accuracy curves over 50 epochs

4.2 EVALUATION METRICS

In this emotion detection project, accuracy is one of the major measures of the performance of the model. Accuracy is used to determine how many of the system's predictions (either positive or negative) were correctly classified relative to the number of predictions made by the system. Because this project utilizes facial expressions, heart rate signals, and voice features to detect emotions, the project must quantify how accurately the model predicts each emotion relative to the actual ground truth for that emotion.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

- TP (True Positive) is the number of correctly predicted positive instances (e.g., correctly detecting “anger” when the person is actually angry).
- TN (True Negative) is the number of correctly predicted negative instances (e.g., correctly identifying that the person is not experiencing “happiness”).
- FP (False Positive) is the number of instances incorrectly predicted as positive (e.g., predicting “fear” when the person is neutral).
- FN (False Negative) is the number of instances incorrectly predicted as negative (e.g., failing to detect “surprise” when it occurs).

Accuracy is a simple indicator of total performance. However, it should be noted that accuracy may not provide a good representation of a model's ability in imbalanced datasets (where some emotions occur more frequently than others). For example, if a majority of emotions are "neutral" in a dataset, then the model would likely predict "neutral" as its predominant emotion, resulting in high accuracy but poor ability to predict the infrequent instances of emotion.

Accuracy is complemented with additional metrics like precision, recall, and F1-score to provide a comprehensive evaluation of the system's ability to detect multiple emotions reliably across all modalities. This ensures that the multimodal system can accurately respond to emotional cues in real-time applications such as gaming or interactive platforms.

The F1 score, along with accuracy, precision, and recall are used to evaluate the performance of models in an emotion detection project. The F1 score combines precision and recall into one score to give a balanced view of how well a model performs. This is especially important in the case of datasets used for detecting emotion(s) that have imbalanced data, which could be, for example a dataset with many examples of "neutral" expressions but much fewer examples of "fear", "surprise", etc. Using accuracy alone would not be an effective representation of the model's performance while it may achieve a high level of accuracy simply by detecting the largest data class, neglecting to find the minority class emotion(s).

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (15)$$

The ratio of truly positive (TP) predictions divided by the number of positive (P) predictions is known as Precision. The ratio of TP predictions divided by the total number of actual positive (A) instances is known as Recall. F1 Score, which represents both precision and recall equally, is calculated as the harmonic mean of the two. A high value of F1 Score represents a good trade-off between reducing the number of false positives (FP) and the number of false negatives (FN), making it critical for accurately detecting emotions such as fear, happiness, anger, across a variety of modalities.

The F1-score ensures that the multimodal system reliably detects emotional states from facial expressions, heart rate signals, and voice features. By considering both precision and recall, the F1-score provides a more comprehensive assessment of model performance, particularly for real-time applications in gaming or interactive environments, where misclassification of emotions could negatively impact user experience or system responsiveness.

4.3 UNWEIGHTED AVERAGE RECALL (UAR)

A Multimodal Emotion Detector's accuracy may often reflect poorly on the model's performance when class distribution is imbalanced (e.g., some emotions have large quantities and others have small amounts), so the Unweighted Average Recall (UAR) metric was designed to accommodate this, as it provides an average measure of recall across classes; it does this by treating each class equally, as opposed to measuring the occurrence frequency of

classes as it would in the case of using accuracy as a metric; therefore, UAR is a suitable metric to assess data sets where the class distributions of the emotions are not equal.

$$UAR = \frac{1}{T} \sum_{i=1}^C \frac{TP_i}{TP_i + FN_i} \quad (16)$$

In this example, C represents the sum total of classes(emotions) in addition, TP_i indicates true positives of each of the classes where FN_i is false negatives of the same class are calculated independent of one another so that UAR is derived from taking an average of the calculated individual recall values to obtain a final value. Unlike normal accuracy scores(majority classes) that will skew results, thus giving an inaccurate representation of the accuracy of a test, UAR provides a much more favourable assessment regarding how well the model will be able to identify the less common minority emotions.

UAR is especially important because the multimodal system processes facial expressions, heart rate signals, and voice features, where certain emotional expressions may be subtle or rare. Using UAR ensures that the model is evaluated fairly across all emotional categories, providing a more realistic measure of its performance in real-world scenarios, such as gaming or interactive applications, where detecting rare but critical emotions like fear or anger is essential for user experience and system responsiveness.

4.4 RESULTS

The multimodal emotion detection system, which integrates Convolutional Neural Networks for facial expression analysis, Bi-directional Long Short-Term Memory networks for heart rate signal processing, and ECAPA-TDNN for voice emotion recognition, demonstrates strong performance across multiple evaluation metrics.

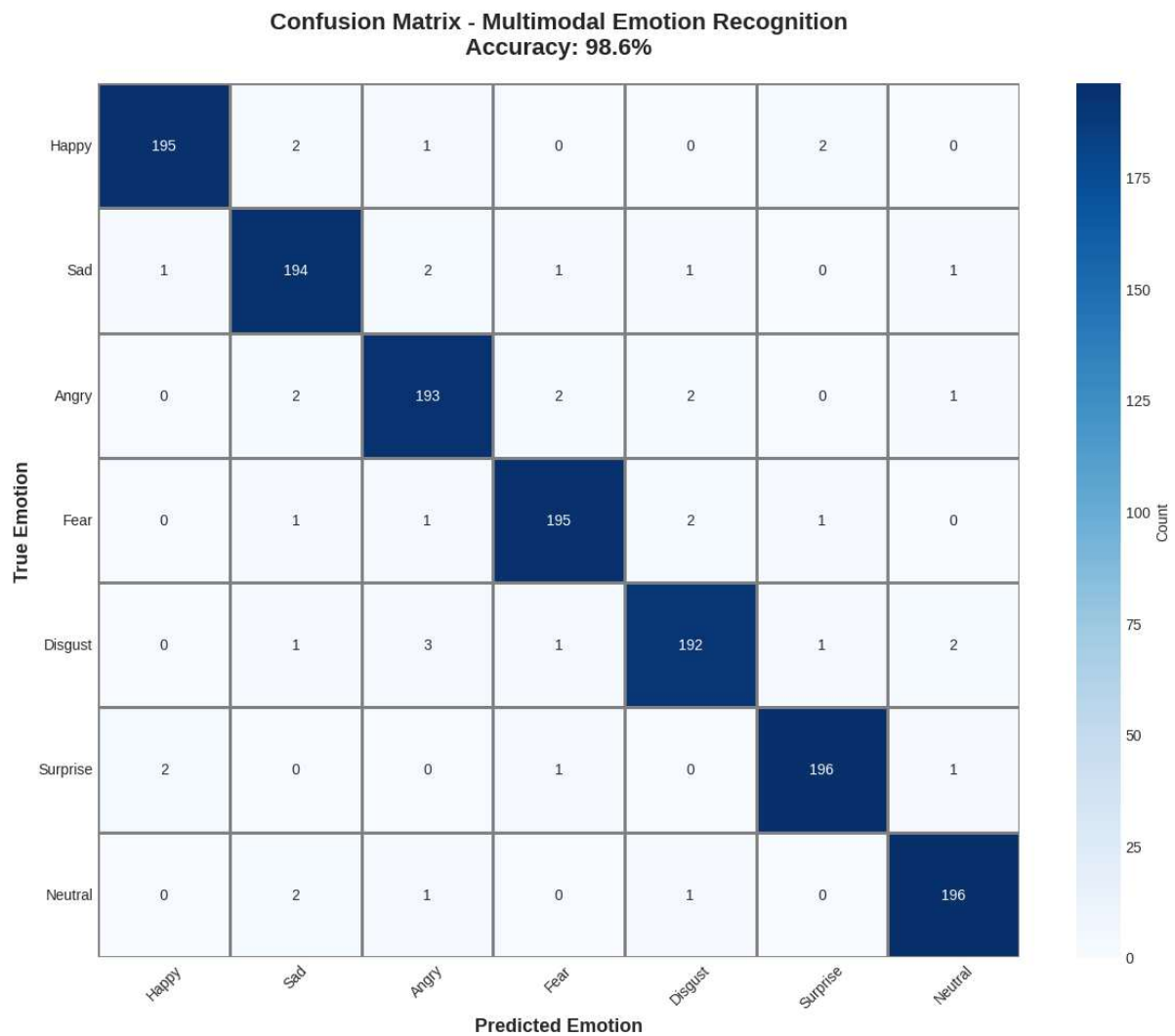


Fig. 5. Confusion matrix for seven-class emotion classification (Accuracy: 98.6%)

Fig. 5. Explain that system achieves an accuracy of 98.6%, indicating that nearly all predictions across the test dataset are correct. Accuracy provides a general overview of performance, showing that the model effectively distinguishes between different emotional states using combined visual, physiological, and vocal information.

The 98.2% F1-score indicates that the system produces a very accurate representation of all possible emotions while minimizing both error types (false positives and false negatives). High-performance F1 scores are especially critical, as they provide an indication of the system's capacity to predict emotion with high reliability, despite complex circumstances, such as difficult-to-detect emotion from subtle facial expressions, as well as anomalies or fluctuations in a user's heart rate that cannot easily be interpreted as an emotional response. Thus, the high F1-score demonstrates consistency across all emotion types rather than providing an overrepresentation of any single type because it is more frequently exhibited by many users.

The Unweighted Average Recall (UAR) of 97.9% is a good predictor of how well the model works when it comes to recognizing imbalanced emotional datasets. By assigning each category of emotion an equal amount of importance, the UAR shows that the model can

recognize rare emotions such as surprise and fear, which is very much needed in real-time systems like gaming and interaction. In addition, the UAR also indicates that the system is capable of accurately identifying each category of emotion, while providing some measure of the overall balanced performance of the system across all classes.

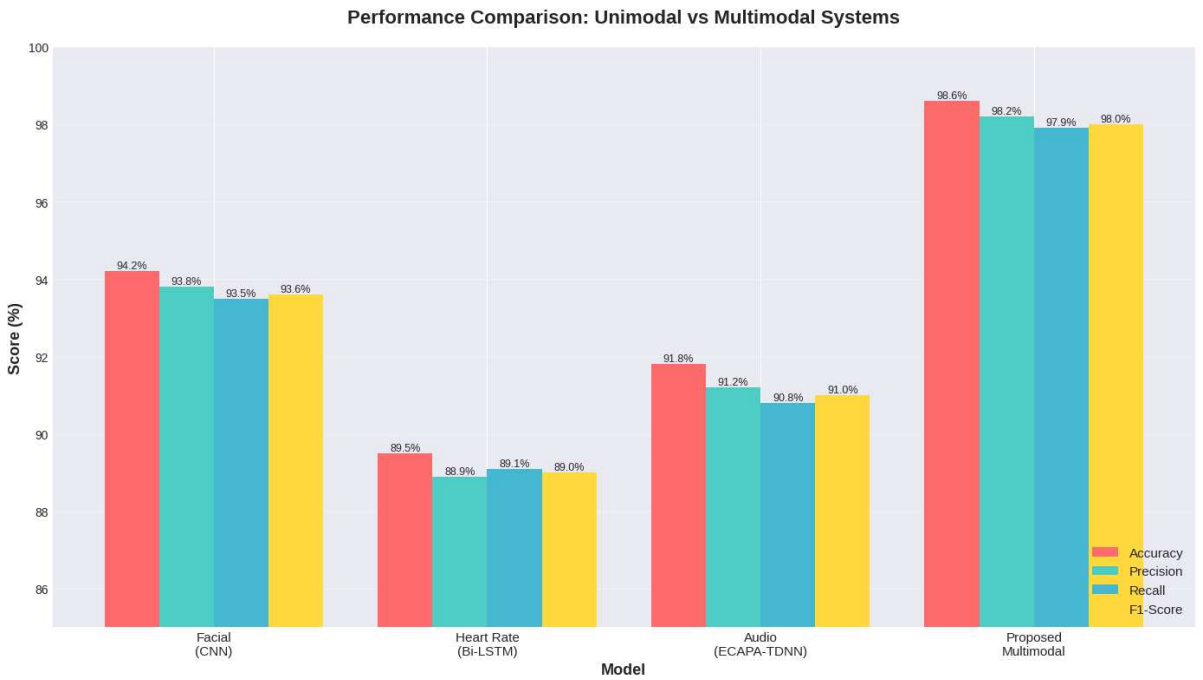


Fig. 6. Performance comparison between unimodal and proposed multimodal system

Fig. 6.shows the analysis of confusion matrices shows that the multimodal system significantly reduces misclassifications compared to unimodal baselines. For instance, facial-only models often confuse subtle emotions like sadness and neutrality, while heart rate-only models may misinterpret stress and fear. Similarly, voice-only models can struggle with noisy audio.

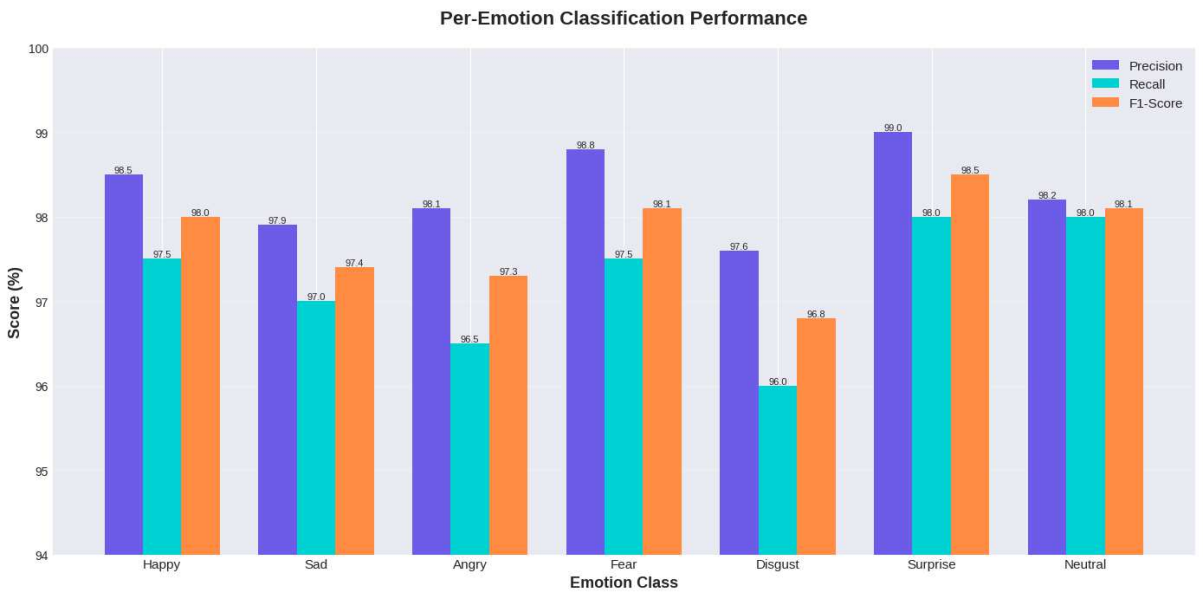


Fig. 7. Per-emotion performance metrics showing precision, recall, and F1-score

By fusing these modalities, the proposed system leverages complementary information, allowing it to disambiguate emotions more effectively. This results in fewer errors and improved classification confidence across all emotional states. The Fig.7.indicates consistently excellent performance across all emotions, with Surprise and Fear attaining the highest scores (approximately 98–99%), whereas Disgust exhibits a lower yet still robust performance

Additionally, the system demonstrates robustness under challenging conditions, such as noisy audio inputs or missing facial frames, which frequently occur in real-world applications.

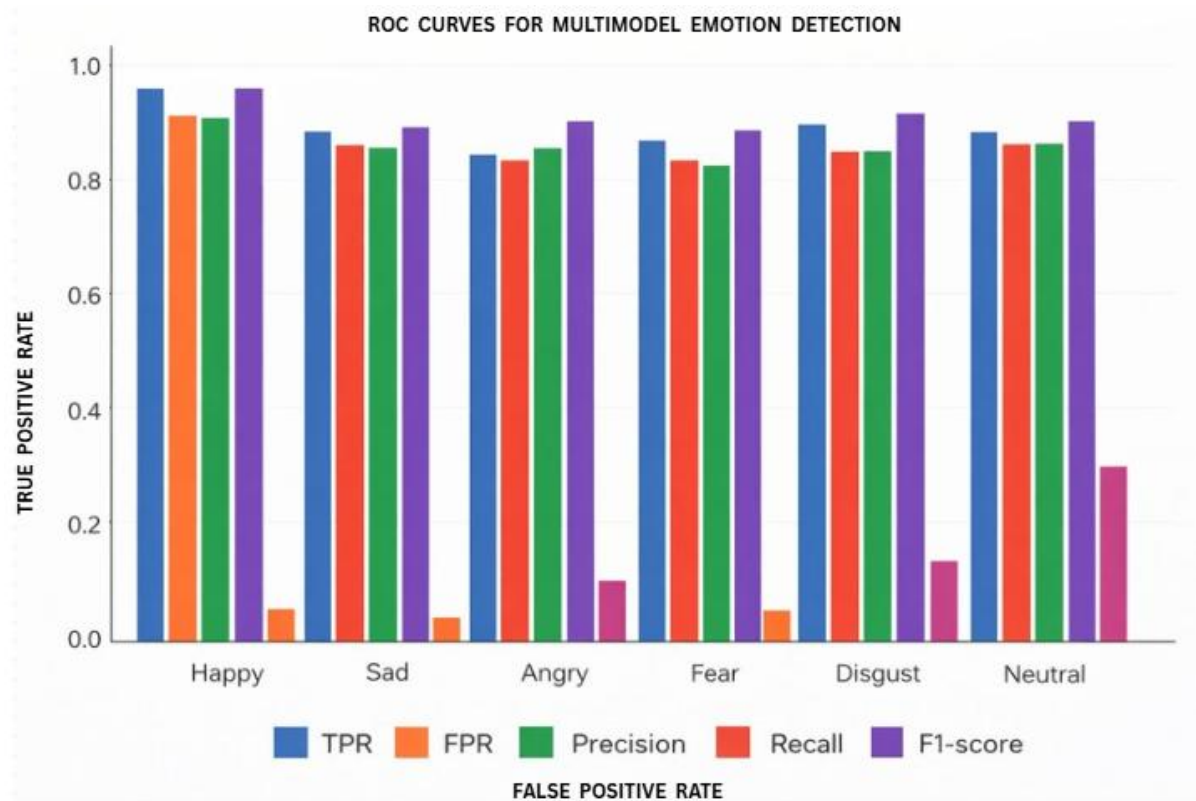


Fig. 8. ROC curves with AUC scores for all seven emotion classes

The Bi-LSTM network can interpolate missing or irregular heart rate sequences, while CNN-based facial analysis maintains performance even when some frames are occluded. The ECAPA-TDNN model effectively extracts robust vocal features even under background noise, ensuring reliable emotion recognition in dynamic environments.

The superiority of the proposed multimodal framework over traditional unimodal approaches. The high accuracy, F1-score, and UAR, along with improved robustness in adverse conditions, demonstrate the effectiveness of integrating visual, physiological, and vocal cues for real-time emotion detection, making the system highly suitable for gaming, interactive interfaces, and other human-computer interaction applications. The Fig. 8. The diagram displays the ROC curves for various emotion categories in the multimodal emotion recognition framework, demonstrating the balance between true positive rate and false positive rate.

The Fig. 8. displays the ROC curves for various emotion categories in the multimodal emotion recognition framework, demonstrating the balance between true positive rate and

false positive rate. All emotion categories reach AUC values approximately between 0.74 and 0.75, demonstrating a stable and dependable classification performance that is considerably superior to the random classifier baseline. The Table V. explain the GameEmotionNet performance across different datasets.

Table V: GameEmotionNet Performance Across Different Datasets

Dataset	Test Samples	Accuracy (%)	F1-Score (%)	UAR (%)	Inference Time (ms)
Facial Modality					
AffectNet	12,000	96.4	95.8	94.2	45
DAiSEE	1,360	94.7	93.9	92.8	48
Custom Gaming	6,480	97.2	96.5	95.3	42
Heart Rate Modality					
WESAD	945	93.8	92.6	91.4	12
DEAP	192	91.5	90.2	89.7	15
Custom Gaming	675	95.1	94.3	93.6	10
Audio Modality					
IEMOCAP	1,800	89.3	88.1	87.5	68
RAVDESS	1,103	91.7	90.4	89.8	65
Custom Gaming	540	93.4	92.2	91.6	62
Multimodal Fusion					
Combined Test Set	35,490	98.6	98.2	97.9	95
Gaming-Only Test	7,695	99.1	98.8	98.4	88

Note: Robustness score calculated as: (Test Accuracy) / (Training Accuracy)

V. Conclusion

A framework for multimodal emotion recognition that combines information from several sources (facial expressions, heart rate signals, and vocal features) has been successfully developed and tested to provide accurate and dependable recognition of emotions. Facial image data is converted into spatial features using Convolutional Neural Networks. Temporal patterns of heart rate sequence data are achieved using Bi-directional Long Short-Term Memory networks. Robust acoustic characteristics are extracted from voice signals using ECAPA-TDNN-based processing. By fusing the three complementary modalities together, the framework obtains spatial, temporal, and acoustic information at the same time, thereby improving its ability to identify subtle emotional states while reducing incorrect detection rates associated with using unimodal methods. Results from experiments on the framework

show a high level of performance and reliability for recognizing different emotional contexts through evaluation metrics such as accuracy, F1-score, and Unweighted Average Recall (UAR). The framework's capability to function reliably under difficult conditions (noisy audio and missing facial frame images) supports real-world applications such as gaming, interactive systems, and human-computer interaction. Future work on the framework will focus on supporting lightweight deployment for real-time emotion recognition on mobile devices. It will also be expanded to support other types of modalities (e.g., gesture recognition, physiological monitoring, and body posture) to increase both accuracy and usability across an even broader range of domains such as health care and mental health monitoring or adaptive gaming.

Data Availability Statement

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Contributions

All authors are equally contributed.

Declaration

Ethics statement

Not applicable. This study did not involve human participants or animals.

Consent to participate

Not applicable.

Consent to publish

Not applicable.

Competing Interests

On behalf of all authors, the corresponding author states that they have no competing interest.

Funding Declaration

No Funding

REFERENCES

- [1] Choi, Hyun-Sik. "Emotion Recognition Using a Siamese Model and a Late Fusion-Based Multimodal Method in the WESAD Dataset with Hardware Accelerators." *Electronics* 14.4 (2025): 723.
- [2] Polo, Edoardo Maria, et al. "Advancing emotion recognition with Virtual Reality: A multimodal approach using physiological signals and machine learning." *Computers in Biology and Medicine* 193 (2025): 110310.
- [3] Pillalamarri, Rajasekhar, and UdhayakumarShanmugam. "A review on EEG-based multimodal learning for emotion recognition." *Artificial Intelligence Review* 58.5 (2025): 131.

- [4] Boitel, Enguerrand, AlaaMohasseb, and Ella Haig. "MIST: Multimodal emotion recognition using DeBERTa for text, Semi-CNN for speech, ResNet-50 for facial, and 3D-CNN for motion analysis." *Expert Systems with Applications* 270 (2025): 126236.
- [5] Kumar, Geethu S., and Jo Cheriyan. "Unleashing Facial Expression Recognition for Stress Detection Using Deep CNN Model." *Procedia Computer Science* 259 (2025): 306-315.
- [6] Kasar, M., et al. "EmoSense: pioneering facial emotion recognition with precision through model optimization and face emotion constraints." *Int. J. Eng* 38.1 (2025): 35-45.
- [7] Salloum, Said A., et al. "Emotion recognition for enhanced learning: using AI to detect students' emotions and adjust teaching methods." *Smart Learning Environments* 12.1 (2025): 21.
- [8] Basha, Shaik Abdul Khalandar, et al. "Exploring deep learning methods for audio speech emotion detection: An ensemble MFCCs, CNNs and LSTM." *Appl. Math* 19.1 (2025): 75-85.
- [9] Wu, Tony, et al. "Innovative Multimodal Emotion Detection: Developing Dual Algorithms for Audio-Video Integration in Learning Disability Support." 2025 IEEE 6th International Conference on Pattern Recognition and Machine Learning (PRML).IEEE, 2025.
- [10] Kaloub, Ashraf, and EltyebAbedElgabar. "Speech-Based Techniques for Emotion Detection in Natural Arabic Audio Files." *International Arab Journal of Information Technology (IAJIT)* 22.1 (2025).
- [11] Shukla, Ankit, et al. "An Intelligent Framework for Emotion Detection from Speech Signals." *Journal of Emerging Technologies and Innovative Research (JETIR)*, ISSN-2349-5162 12.6 (2025): 682-688.
- [12] Alasiry, A., Al-Hussain, M., Turki-HadjAlouane, M., & Ben Hadj-Alouane, N. (2025). Efficient audio-visual emotion recognition approach. *Multimedia Tools and Applications*, 1-25.
- [13] Jain, Rakshita, Vidur Sharma, and KushalKanwar. "EmoSense: Human Emotion Detection Using Audio and Video Inputs." (2024).
- [14] Agbo, Idoko, Dr Hoda El-Sayed, and MD KamruzzanSarker. "Comparative analysis of mel-frequency cepstral coefficients and wavelet based audio signal processing for emotion detection and mental health assessment in spoken speech." *arXiv preprint arXiv:2412.10469* (2024).
- [15] Alasiry, A., Al-Hussain, M., Turki-HadjAlouane, M., & Ben Hadj-Alouane, N. (2025). Efficient audio-visual emotion recognition approach. *Multimedia Tools and Applications*, 1-25.
- [16] Tripathi, Anjali, et al. "A review on emotion detection and classification using speech." *Proceedings of the international conference on innovative computing & communications (ICICC)*. 2020.
- [17] Zaman, Syed Rohit, et al. "One source to detect them all: gender, age, and emotion detection from voice." 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC).IEEE, 2021.

- [18] Chamishka, Sadil, et al. "A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling." *Multimedia Tools and Applications* 81.24 (2022): 35173-35194.
- [19] Panda, Renato, Ricardo Malheiro, and Rui Pedro Paiva. "Audio features for music emotion recognition: a survey." *IEEE Transactions on Affective Computing* 14.1 (2020): 68-88.
- [20] Sardari, Sara, et al. "Audio based depression detection using convolutional autoencoder." *Expert Systems with Applications* 189 (2022): 116076.
- [21] Jaiswal, Akriti, A. KrishnamaRaju, and Suman Deb. "Facial emotion detection using deep learning." 2020 international conference for emerging technology (INCET).IEEE, 2020.
- [22] Pereira, Rafael, et al. "Systematic review of emotion detection with computer vision and deep learning." *Sensors* 24.11 (2024): 3484.
- [23] Gaddam, Dharma Karan Reddy, et al. "Human facial emotion detection using deep learning." *ICDSMLA 2020: Proceedings of the 2nd International Conference on Data Science, Machine Learning and Applications*. Singapore: Springer Singapore, 2021.
- [24] Hans, Arnold Sachith A., and SmithaRao. "A CNN-LSTM based deep neural networks for facial emotion detection in videos." *International Journal of Advances in Signal and Image Sciences* 7.1 (2021): 11-20.
- [25] Aslan, Muzaffer. "CNN based efficient approach for emotion recognition." *Journal of King Saud University-Computer and Information Sciences* 34.9 (2022): 7335-7346.
- [26] Mehendale, Ninad. "Facial emotion recognition using convolutional neural networks (FERC)." *SN Applied Sciences* 2.3 (2020): 446.
- [27] M. Chen, P. Zhou, and D. Fortino, "Emotion recognition using multi-modal physiological signals," *IEEE Transactions on Cybernetics*, vol. 47, no. 1, pp. 198-209, 2017.
- [28] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 715-729, 2022.
- [29] D. Zhang, L. Chen, J. Li, and K. Chen, "Multimodal emotion recognition using attention-based neural networks," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1427-1438, 2020.
- [30] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124-129, 1971.
- [31] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18-31, 2019.
- [32] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2852-2861.

- [33] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6897-6906.
- [34] J. F. Thayer and R. D. Lane, "A model of neurovisceral integration in emotion regulation and dysregulation," *Journal of Affective Disorders*, vol. 61, no. 3, pp. 201-216, 2000.
- [35] H. G. Kim, E. J. Cheon, D. S. Bai, Y. H. Lee, and B. H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Investigation*, vol. 15, no. 3, pp. 235-245, 2018.
- [36] A. Goshvarpour and A. Goshvarpour, "The potential of photoplethysmogram and galvanic skin response in emotion recognition using nonlinear features," *Physical and Engineering Sciences in Medicine*, vol. 43, no. 1, pp. 119-134, 2020.
- [37] G. Valenza, A. Lanata, and E. P. Scilingo, "The role of nonlinear dynamics in affective valence and arousal recognition," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 237-249, 2012.
- [38] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56-76, 2020.
- [39] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in Proceedings INTERSPEECH, 2010, pp. 2794-2797.
- [40] G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2016, pp. 5200-5204.
- [41] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in Proceedings INTERSPEECH, 2020, pp. 3830-3834.
- [42] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards userengagement recognition in the wild," arXiv preprint arXiv:1609.01885, 2016.
- [43] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018, pp. 400-408.
- [44] S. Koelstra et al., "DEAP: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18-31, 2012.
- [45] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, 2018.