Radius: Journal of Science and Technology
Vol. 3, Issue 1, (2026) 261005 (14pp)

A conceptual framework for deep learning-based multimodal emotion detection using facial expressions and physiological signals

Akshata G. Shinde[1,*] | Shivaji G. Shinde[1]

[1]TPCT's College of Engineering, Dharashiv - 413501, Maharashtra (India)

*Corresponding author: akshatashinde620@gmail.com

Abstract: There is increased interest in one aspect of this, namely the identification of emotion as an important field in affective computing, in which the ability of intelligent systems to analyze human emotions has been introduced and the efficiency of human–machine interactions has been improved. Facial expression recognition (FER) is the only modality used in traditional approaches to emotion recognition and consequently these approaches are less robust because of inter-personal and inter-environmental variability. In this paper, the recent advances in multimodal emotion recognition (MER) systems that involve combining the facial expression recognition through deep learning approaches with physiological signal-based emotion recognition (PSER) are reviewed. An overview of the reviewed works shows that a conceptual multimodal framework can be presented that consists of data acquisition, data preprocessing, feature extraction, multimodal fusion, and classification stages. Previous studies have extensively used Convolutional Neural Networks (CNNs) for facial feature extraction, while physiological signals like electrodermal activity (EDA) and heart rate variability (HRV) have been traditionally analyzed using statistical and frequency domain approaches. Emotional pattern learning across the various domains and fusion of heterogeneous multimodal representations can be achieved by feature-level fusion strategies, as reported in the literature. Recent studies in the literature have shown that temporal modeling approaches have proven to be more stable when facing dynamic environmental conditions. Architectures reviewed are typically modular and designed to be scalable and adaptable for future real-life implementation scenarios. The aim of the present work is primarily to review the latest emotion recognition systems based on multimodal approach and to present a conceptual framework based on the literature. Future research directions are suggested to be experimental implementation and quantitative validation.

Keywords: Multimodal Emotion Recognition (MER), Facial Expression Recognition (FER), Physiological Signal-Based Emotion Recognition (PSER), Deep Learning, Convolutional Neural Network (CNN)

1. Introduction

One of the recent emerging topics in affective computing is emotion recognition. Emotionally-aware technologies have been extensively utilized in applications such as adaptive learning systems, driver assistance systems, health care surveillance systems and human-computer interaction systems. When machine is able to understand the emotional state with accuracy, it can respond to it in an adaptive and contextual way, increasing effectiveness of the system and the user experience. Traditionally, emotion recognition has relied heavily

https://doi.org/10.5281/zenodo.20514398
Received: 30 April 2026 | **Revised:** 18 May 2026
Accepted: 01 June 2026 | **Published Online:** 05 June 2026

on evaluating facial expressions, providing one of the most visible and easily accessible forms of human emotion display. Progress towards improved performance in facial emotion recognition has been remarkable with the advent of deep learning methods, especially convolutional neural network (CNN) models [1]. However, the performance of these models is adversely affected by real world conditions, including illumination differences, occlusions, pose changes, and inter-individual differences in expressing emotions. As subsequent studies examined the limitations of using vision-based systems as an input, many have also turned to the use of physiological signals such as heart rates, galvanic skin response (GSR), and electroencephalograms (EEG) to represent internal and involuntary emotional responses [2]. Physiological signals are less likely to be affected by external factors than vision-based systems, and therefore provide a more legitimate measure of emotional condition. However, systems that rely solely on the use of physiological signals lack contextual richness, and may also be affected by sensor noise or calibration error. Research in the recent past has shown that the combination of multiple modalities enhances the robustness and accuracy of emotion detection systems. Combining facial expressions with physiological cues, multimodal emotion recognition has been considered an efficient way to address the shortcomings of single-modality methods [3]. The heterogeneous data sources can be combined such that complementary information is used leading to better classification performance and reliability of the system.

The deep learning structures are significant in promoting effective multimodal information processing. Hybrid architectures combining CNN-based feature extraction with recurrent models have proven highly effective in learning both spatial and temporal emotional patterns, which combine CNN-based feature extractions with recurrent models, have proven highly effective in learning both spatial and temporal patterns in emotional data [4]. Additionally, multimodal fusion methods (like feature-level and decision-level fusion) have become common to combine various data representations into a single structure [5]. Simultaneous progress in wireless communication and sensing technologies, especially in sub-6 GHz 5G systems have enabled the creation of high-speed, low-latency data acquisition platforms. To use physiological signal data in generating real-time multimodal emotion data to support emotion-aware applications, the technologies are required to have a reliable means by which to transmit these multimodal data sets. Previous research contributions in providing antenna design and multiple-input/multiple-output (MIMO) communication systems to facilitate 5G communications reveal the necessity of reliable data transmission infrastructure to be used in the development of intelligent systems and applications [6-8].

Unfortunately, while there are many promising developments in the area of physiological signal processing, there continues to be obstacles in the creation of an efficient and scalable multimodal emotion detection system that successfully integrates disparate data sources while maintaining computing efficiency. Based on the reviewed literature, this paper discusses a conceptual deep-learning-based multimodal framework that combines facial expression analysis with physiological signal processing for emotion recognition applications.

2. Literature Review

Emotion recognition has made great strides recently, especially due to deep learning methods combined with multiple types of data input. Previous research studies on facial expression

recognition using CNN models were very effective when tested in a controlled environment, yet they have low generalization ability to real-world variations (e.g., light variation, occlusion, and pose variation). From literature survey, it is known that CNN based FER models are found to be very accurate on the benchmark dataset but there are still some concerns regarding their robustness in non-constrained settings [9].

It is in this regard that researchers are starting to explore the use of PSER as a means of enhancing reliability and robustness. PSER uses physiological parameters including EDA, heart rate variability (HRV) and EEG to assist in doing this. Such signs give inherent information about the moods and are not so affected by external disruptions. In research done on multimodal physiological signals, it was observed that deep learning networks are capable of emotion classification by identifying time- and frequency-domain features of these signals [10]. Although PSER systems are reliable, they also frequently need a complicated setup of sensors as well as preprocessing.

The shortcomings of unimodal systems have given way to multimodal emotion recognition (MER) systems. Such systems combine complementary information among various sources which enhances accuracy and strength. Recent deep learning-based MER model that integrates both facial and physiological features showed that multimodal fusion models can achieve significantly better classification than single models [11]. Likewise, a study has highlighted that feature-level fusion (FLF) represents inter-modal relations better than decision-level fusion (DLF) in case data synchronization is ensured [12].

Besides extracting spatial features, modeling the temporal dependencies has emerged as an enlightened research direction. CNNs used with long short-term memory (LSTM) networks to form a hybrid network have proven to be effective in both temporal and spatial analysis of emotional data. Models such as these can be used to analyze types of sequential data such as videos and time-varying physiological signals so that more stable and consistent predictions are achieved [13].

Another important consideration in MER systems is the proper way to merge the input from different modalities. Fusion techniques are typically divided into three types – FLF, DLF and hybrid fusion. Recent review studies indicated that FLF delivers the most effective outcomes if their input modalities are properly matched with each other and DLF can process asynchronously input and heterogeneous data streams [14]. Both hybrid fusion techniques are being tried, which try to combine some of the best of both technologies with additional computational complexity.

Transformer based multimodal architectures have been recently shown to have promising abilities in capturing long-range dependencies and contextual interactions between heterogeneous emotional modalities. Furthermore, self-supervised learning methods have attracted interest for their ability to alleviate the reliance on manually annotated emotional databases and facilitate the efficient representation learning within multimodal emotion recognition systems.

Much work is done in the design of efficient MER systems. However, the following challenges continue to exist: High computational cost, availability of sufficiently large collections of synchronized multimodal datasets and challenges associated with implementing MER systems in real time. Further, the scalability and adaptability of MER systems to various applications remain an open research problem.

Table 1: Summary of Recent Emotion Detection Approaches

Ref.	Approach	Modalities	Technique	Key Findings
[9]	FER	Facial	CNN	High accuracy on benchmark datasets; limited real-world robustness
[10]	PSER	Physiological (EDA, HRV, EEG)	DL Models	Reliable emotion inference; requires sensor setup
[11]	MER	Facial + Physiological	CNN-based Fusion	Improved accuracy over unimodal systems
[12]	MER	Visual + Physiological	FLF/DLF	FLF shows better performance with synchronized data
[13]	MER	Video + Signals	CNN-LSTM	Captures spatial-temporal features effectively
[14]	MER	Multimodal	Hybrid Fusion	Balances flexibility and performance

2.1 Comparative Analysis of Existing Emotion Recognition Approaches

The reviewed studies have shown that the multimodal emotion recognition systems usually perform better than unimodal systems, as they exploit complementary emotional information from different sources. Despite the good visual cues provided by facial expression recognition methods, they can be affected by change in facial pose, illumination and occlusions. Physiological signal-based approaches are data-driven, and they convey some internal emotional information, but they typically need special sensors and a large amount of processing. Multimodal systems try to address these shortcomings by fusing both outward and internal emotional signals.

As shown in Table 1, CNN-based FER systems can attain high accuracy in controlled environments, while multimodal fusion systems are more robust and reliable in various scenarios. Moreover, the hybrid architectures built on temporal modeling method (CNN-LSTM networks) can capture both spatial and temporal emotional features. The observations here suggest that the multimodal fusion with temporal modelling is a promising direction for emotion recognition systems.

MER systems are better than unimodal emotion detection methods as identified above, but there remains a number of challenges related to the efficient computation, sync and real time application of MER systems. This project will develop a deep learning MER framework that involves the two sources of emotion information (face and physiology) and will implement an optimized data fusion process that will increase the performance of the whole system.

The reviewed studies show that multimodal emotion recognition systems tend to achieve better accuracy than unimodal emotion recognition systems because they are able to combine the complementarity of emotional cues. But there are problems of computational complexity, multimodal synchronization, data availability, and the real-time implementation that still remain to be explored as research issues. Hence, in the present study, the recent developments have been summarized and a conceptual frame work has been proposed based upon the existing research findings.

2.2 Research Contribution and Conceptual Novelty

The recent progress in the field of multimodal emotion recognition is summarized and a conceptual framework that unites the FER and PSER in the same DL architecture is discussed. The purpose of the framework is to incorporate both external face expression and internal physiological information cues, which are complementary [15].

The main conceptual advances of the present work are summarized below:

- Modular multimodal framework discussion of FER and PSER approaches.
- Incorporation of the concepts of feature-level fusion and decision-level fusion for multimodal integration [16].
- Inclusion of temporal stabilisation ideas aimed at achieving consistency in prediction.
- Presentation of a scalable conceptual architecture suitable for future real-time implementation and extension toward additional modalities such as speech and textual emotion analysis [17].
- Consolidation of preprocessing, feature extraction, fusion, and classification stages into a unified methodological framework derived from existing literature.

3. Methodology

This section presents the design and operational workflow of the MER system. The framework contains both Face Emotion Recognition (FER) and Physiology Emotion Recognition (PSER) through the use of a deep learning framework. The goal is to develop a robust MER system that is able to recognize external (faces) and internal (physiology) emotional inputs thus, improving overall reliability and consistency [18].

3.1 System Architecture Description

The conceptual MER framework consists of several processing stages in which input data may be transformed into progressively higher-level representations. The proposed conceptual architecture consists of four major functional modules, as shown in Figure 1 below.

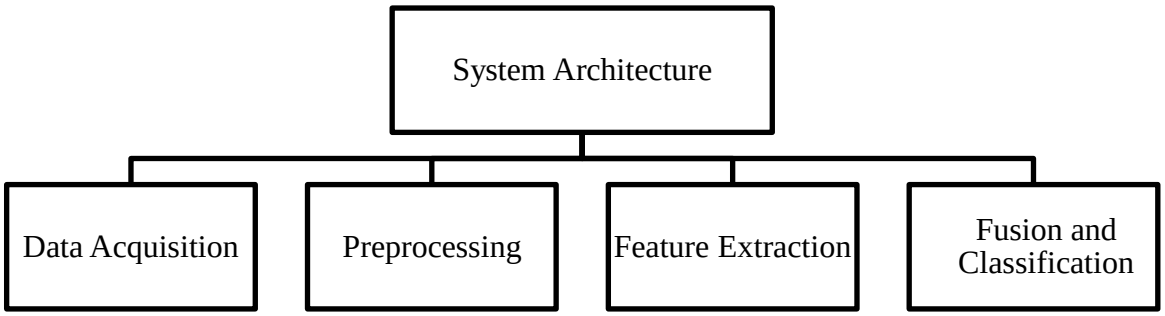

Figure 1: System Architectures module

The proposed multimodal emotion recognition architecture is shown in Figure 2, illustrating that both the facial and physiological signal inputs are processed by parallel processing streams with unique preprocessing and feature extraction processes. The resulting feature sets undergo a feature-level fusion process and are submitted to a deep learning classifier to predict emotion. A component for temporal modelling is added to enhance temporal consistency of predictions, and the design is optimised to provide the maximum possible balance of computation efficiency and performance for real-time applications.

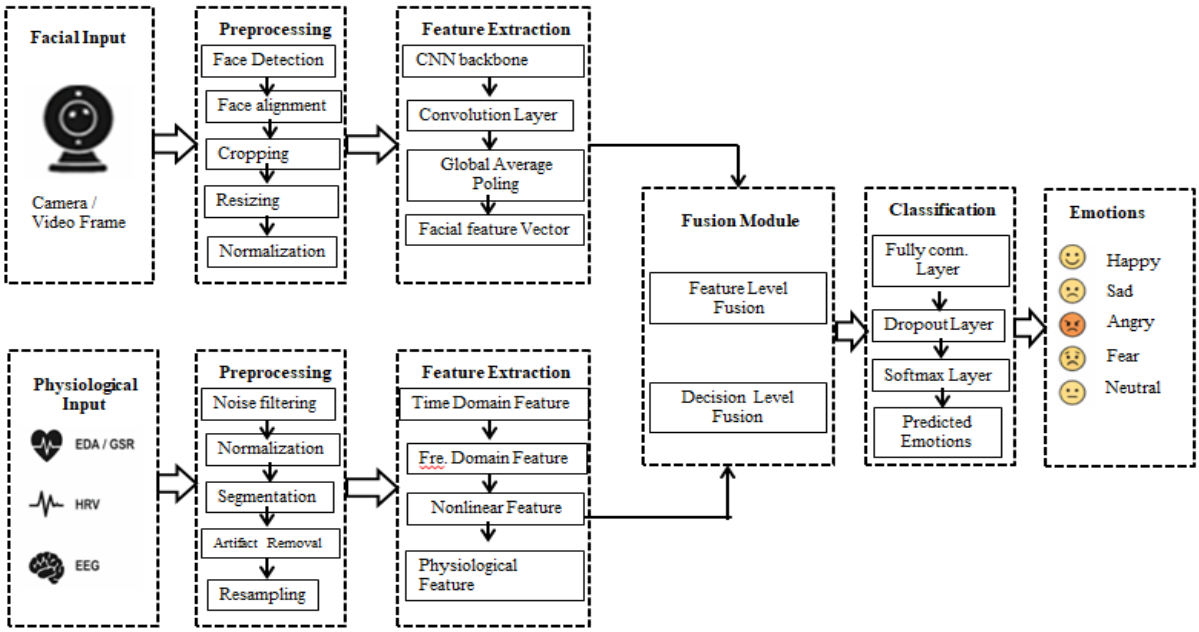

Figure 2: Detailed Architecture of Proposed MER System

3.2 Data Acquisition Strategy

Quality and diversity of the input data that is obtained are significant determinants of the performance of an MER system. In this work two complementary modalities are taken into account:

3.2.1 Facial Data (FER Input)

In multimodal emotion recognition studies, facial image acquisition is commonly performed using real-time webcams or benchmark datasets. Every frame is a record of facial expressions: inherently, face expressions encode emotional conditions. Temporal tracking of the expressions is made possible by continuous frame capture [19].

3.2.2 Physiological Data (PSER input)

Among the various studies that have been previously carried out, wearable sensors are used to capture physiological data such as EDA, HRV, and EEG. These cues are the reactions of the autonomic nervous system and give subconscious emotional clues. These signals are time-series and thus they need to be synchronized with the facial data in order to be effectively fused.

One important design factor is that of temporal alignment, so that both modalities represent the same emotional instance. The temporal synchronization of facial video frames and physiological signals can be done with the help of timestamped alignment and fixed-length sliding windows. Physiological signals can have different sampling rates and can be resampled or interpolated to maintain consistency of synchronization between modalities [20].

3.2.3 Proposed Benchmark Datasets for Future Validation

Table 2 summarizes some common benchmark datasets used in multimodal emotion recognition studies. These datasets could be used in the future to implement and validate the proposed conceptual framework. Training, validation and testing methods will vary based on the dataset and application needs, for further implementation studies, such as methods as 70:15:15 or cross-validation methods can be used.

Table 2: Benchmark Datasets for Future Validation

Ref.	Dataset	Modality	Samples	Emotion Classes
[38]	CK+	Facial Expressions	593 sequences	Happiness, Sadness, Anger, Fear, Surprise
[39]	FER2013	Facial Images	35,887 images	Seven basic emotions
[40]	DEAP	Physiological Signals	32 participants	Valence/Arousal emotions
[41]	MAHNOB-HCI	Multimodal	527 sessions	Multimodal emotional responses

3.3 Data Preprocessing

Raw data of the two modalities can be noisy and inconsistent. Hence, the data should be preprocessed for improving the quality of data and the performance of the models.

3.3.1 Facial Data Preprocessing

The facial preprocessing step involves a number of steps:

- **Face Detection:** Use deep learning detector to detect the location of the face area in each frame.
- **Face Alignment:** The facial landmarks are used for normalization of face orientation and scale to reduce the effects of the head pose on the variability in the facial landmarks.
- **Normalization:** Normalize pixel values to a fixed range for stable gradient behavior when training the model.
- **Resizing:** The images are resized according to the requirement of the CNN input.

These pre-processing methods are usually employed for ensuring good quality and stability of the input data of CNN-based emotion recognition systems [21].

3.3.2 Physiological Signal Preprocessing

Physiological data can be disturbed by motion of the subject, and external noise in the environment. There is a pre-processing pipeline of:

- Band-pass filters: Filter noise are those which remove unwanted frequency components.
- Segmentation: The signals are broken in fixed length windows to be analyzed.
- Normalization: By standardizing, similar scales of features are provided.
- Artifact Removal: Artifacts are removed (spikes, irregularities).

Such steps are important in deriving significant features of physiological data [22].

3.4 Feature Extraction Mechanism

The feature extraction process brings the raw data into informative features that can be used in classification.

3.4.1 FER using CNN

In emotion recognition, CNN architectures like ResNet are frequently used for extracting spatial facial features. The CNN is a hierarchical learner:

- Low-level features (edges, textures) are captured at the first few layers.
- Intermediary layers detect face parts (eyes, lips).
- Higher layers encode emotional patterns at high levels.

Such models can generate high-dimensional feature embeddings representing emotional facial characteristics [23].

3.4.2 PSER Feature Extraction

Both statistical and spectral methods are used to analyze physiological signals:

- Time-domain features: mean, variance, peak amplitude.
- Frequency-domain properties: power spectral density.
- Nonlinear quantities: entropy and complexity measures.

These characteristics record differences in physiological reactions that relate to various emotional conditions [24].

3.5 Temporal Modeling and Stability Enhancement

The naturally occurring nature of emotion is not fixed. Emotion constantly changes as time goes on, and can never truly be isolated to one point of time. Therefore, instantaneous prediction using a variety of individual frames or very short segments of the signals may not consistently or accurately predict the same value [25]. To address this issue, temporal stability techniques have been discussed in the literature (time-stable adjustment) to ensure a smooth transition between the current value and previous values (frames). Instead of using only the output from one frame as an indicator of emotion, the same frame may also be considered as part of a temporal sequence (group of frames) and measure an adjusted prediction probability for the group using a sliding window approach. By using the average or weighted (decentred) averages across the values from the previous group, abrupt or sudden transitions between emotions are minimised. This technique is also suitable for real-time systems because rapid changes in facial expressions or physiological signals can hinder the classifying process. Furthermore, extensions of the temporal modeling part like Long Short-Term Memory (LSTM) networks can be applied to the model to resolve the underlying temporal pattern. The overall forecasting is, as a result, more uniform, specific, and true to lasting emotional disposition.

3.6 Multimodal Fusion Strategy

Combining the heterogeneous sources of data is a vital aspect of the suggested MER system. As facial expressions and physiological signals offer two different perspectives on emotion recognition, the combination of these two gives significant improvement in the performance of emotion recognition as facial expression and physiological responses are complementary. Structured fusion strategies can be adopted into multimodal emotion recognition systems in a way that features are seamlessly integrated.

The first stage is a feature-level fusion (FLF) which involves fusing the features from the FER and PSER elements into one feature set. This lets interactions between them be directly learnt from the model, which in turn leads to an enhanced expressiveness of the features. Though, FLF presupposes strict synchronization of modalities in order to make sure that the similar features reflect identical time occurrence [26].

In addition to FLF, there is also decision-level fusion (DLF), which is regarded as complementary mechanism. The method is based on the idea that independent predictions of the same variable from different modalities are combined, weighted by averaging, and the weights can be weighted based on the reliability and/or confidence score of the modality. DLF method can handle asynchronous inputs, but cannot make use of the maximum possible correlation of each feature as well as their combination. In practice, the loss or corruption of a modality might be temporary due to sensor noise, occlusion, or communication failure. In

such cases, the system can be made robust using a confidence-based weighting system or unimodal fallback system [27].

So, hybrid fusion models can be capable of performing FLF when a complete integration is needed and can provide the option of DLF for better robustness, thus enabling fine emotion detection and more flexibility in various operational conditions.

3.7 Emotion Classification Model

Once features have been fused, the resulting fused feature space can be processed using a DNN classifier to determine the predicted emotion (E-Motion) of that input sample. In this type of classifier, nonlinear mappings between high-dimensional fused feature spaces and discrete (i.e., finalized) emotional responses are used. It is a series of interconnected layers that decrease the dimensionality in a sequential manner that does not remove discriminative information [28].

To reduce overfitting, dropout regularization is added between layers, and will randomly drop neurons during training, enhancing the generalization ability. The last stage is an output stage that uses a Softmax activation function to transform the model output into the probability scores in the form of the classes. The predicted emotion class is generally determined using maximum probability estimation, which is selected as the most likely [29].

This classification scheme is not rigid and can be modified to other emotional labeling schemes. In the current implementation, the typical types of feelings like happiness, sadness, anger, fear and neutral are used. This will simplify the model to be interpreted and will also enable scaling of the model [30].

3.8 System Design Summary

The general approach is planned in such a way that the raw data collection can flow in to the emotion forecast without obstacle. Each step carries out a certain transformation which improves the quality of data and interpretability. The multimodal inputs are first collected and noise and inconsistencies are removed. Then strong feature extraction techniques are applied to the face and physiology data to extract important features out of the data. These characteristics are then fused with an optimized fusion strategy that enables the system to utilize the complementary information [31]. Lastly, deep learning classifiers can be used to handle the fused representations for prediction of emotion.

Designed in a modular format so that each module can be modified, upgraded or replaced without impacting the whole system. For instance, a different CNN architecture can be applied or advanced signal processing techniques to enhance performance. The proposed framework will be generalizable across a range of applications including areas such as healthcare monitoring, intelligent user interface and adaptive systems. The modular architecture is also intended to support future optimization for low-latency and real-time emotion recognition applications, particularly in edge-computing and wearable-device environments.

Algorithm 1: Conceptual Workflow of the Proposed MER Framework

Step 1: Acquire facial and physiological signals

Step 2: Perform preprocessing and normalization

Step 3: Extract facial features using CNN models

Step 4: Extract physiological features

Step 5: Synchronize multimodal data streams
Step 6: Apply feature-level fusion
Step 7: Perform temporal stabilization
Step 8: Classify emotions using DNN classifier
Step 9: Generate final emotion prediction

3.9 Expected Advantages of the Proposed Conceptual Framework

The proposed conceptual framework may provide several advantages for multimodal emotion recognition systems in several ways. First, it introduces an integrative framework capable of integrating FER and PSER techniques, thus making it possible to analyze both observable expressions and underlying physiological responses. Second, the application of deep learning algorithms helps to automate the process of feature extraction without handcrafting them, which improves the flexibility of the technique [32]. Third, the use of a structured method of information fusion increases the performance of the model through more effective data combination. Fourth, the inclusion of temporal stability techniques is expected to improve prediction consistency even in dynamic conditions. Finally, the modularity and flexibility of the architecture facilitate the development of a real-time system capable of integrating other types of data, such as speech [33].

In general, the discussed ideas and suggested framework try to overcome some of the drawbacks of the traditional unimodal methods used for emotion recognition and find a balance between accuracy and feasibility. According to the reviewed literature, multimodal emotion recognition frameworks have been observed to achieve better performance than unimodal emotion recognition systems because of the combination of complementary emotional information. Though this work doesn't include experimental validation, the proposed framework embeds architectural elements previously reported in literature and demonstrated to enhance the robustness, temporal stability, and classification effectiveness. Future implementation and benchmarking studies will need to be done to quantify these anticipated benefits.

3.10 Ethical Considerations

The use of emotion recognition systems requires data from the face and physiological responses which can lead to privacy and ethical issues. In the future, informed consent, personal information anonymization, secure storage of personal data, and fairness among various demographic groups should be taken into account when implementing the proposed mechanism. Additionally, bias in emotion datasets and automated decision-making systems needs to be thoroughly assessed before implementing these in the real world [34-35].

3.11 Comparative Positioning of the Proposed Framework

Proposed conceptual framework combines facial expression analysis and physiological signal processing, multimodal fusion, temporal stabilization, and scalable deep learning classification in a single framework and differs from many previous studies. Some of the earlier works are dedicated to the individual components of emotion recognition like FER, PSER, and fusion mechanisms individually whereas the present framework is oriented to integrate the various facets of emotion recognition within a single multimodal emotion recognition pipeline. The modular design is envisioned to ease the implementation, optimization, and expansion to other modalities, such as speech and text information.

3.12 Limitations of the Current Study

The present work mainly focuses on reviewing multimodal emotion recognition techniques and presenting a conceptual framework derived from existing literature. Experimental implementation, quantitative evaluation, real-time deployment analysis, and comparative benchmarking are beyond the scope of the current study. Future research will focus on practical validation using benchmark multimodal datasets. In addition, quantitative experimental validation and comparative performance analysis with state-of-the-art methods remain important future research directions [36-37].

4. Conclusion

This paper reviews recent multimodal emotion recognition approaches and discusses a conceptual framework for multi-modal emotion detection through the fusion of FER and PSER using deep learning techniques. Through the use of visible facial activities along with the physiological reactions, the proposed conceptual framework attempts to address limitations posed by the unimodal systems and ensures better results when it comes to emotion interpretation. The proposed pipeline, comprising three main processes of preprocessing, feature extraction, and feature-level fusion, is intended to support effective emotional representation learning. Furthermore, the inclusion of temporal stabilization enables the algorithm to produce consistent outputs in non-stationary environments. Despite being scalable and suitable for practical applications, the current work is limited to conceptual analysis and framework design without experimental implementation of the proposed system.

For future research, the implementation and evaluation of the real-time system using popular benchmarking datasets will be considered. The extension of the scope of the architecture can be adopted to include other modalities such as speech and textual information that would improve the understanding of context and the accuracy of the classification. Attention-based models and transformer networks could be helpful with multi-modal feature learning. In resource constrained applications such as edge and wearable devices, the efficiency of the computation needs to be taken into account. Finally, challenges with the synchronization and sensing accuracy and data privacy should not be neglected.

References

1. Kopalidis, T., Solachidis, V., Vretos, N., & Daras, P. (2024). Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets. *Information*, *15*, 135. https://doi.org/10.3390/info15030135
2. Kim, J., & André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*, 2067–2083. https://doi.org/10.1109/TPAMI.2008.26
3. Barua, A., Ahmed, M. U., & Begum, S. (2023). A systematic literature review on multimodal machine learning: Applications, challenges, gaps and future directions. *IEEE Access*, *11*, 14804–14831. https://doi.org/10.1109/ACCESS.2023.3243854
4. Zhang, Z., Cui, P., & Zhu, W. (2022). Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *34*, 249–270. https://doi.org/10.1109/TKDE.2020.2981333
5. Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, *3*, 42–55. https://doi.org/10.1109/T-AFFC.2011.25

6. Sandeep, K., Sharma, N., & Narwade, N. (2024). Design of slotted patch MIMO antenna and investigation of antenna parameters for sub-6 5G network. *International Research Journal of Multidisciplinary Scope, 5*, 514–523. https://doi.org/10.47857/irjms.2024.v05i03.0994
7. Sandeep, K., Sharma, N., & Narawade, N. (2025). High-isolation dual-band slotted patch MIMO antenna for sub-6 GHz 5G applications. *International Journal of Advanced Technology and Engineering Exploration, 12*, 301. https://doi.org/10.19101/IJATEE.2024.111100612
8. Sandeep, K., Asmita, D., Shrishail, P., Shivale, N., & Sonawane, V. (2025). A novel four-element button mushroom MIMO antenna for enhanced sub-6 GHz 5G communication. *International Research Journal of Multidisciplinary Scope, 6*, 397–409. https://doi.org/10.47857/irjms.2025.v06i02.03051
9. Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13*, 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446
10. Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., & Yang, X. (2018). A review of emotion recognition using physiological signals. *Sensors, 18*, 2074. https://doi.org/10.3390/s18072074
11. Jiao, T., Guo, C., Feng, X., Chen, Y., & Song, J. (2024). A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. *Computers, Materials & Continua, 80*, 1-35. https://doi.org/10.32604/cmc.2024.053204
12. Ma, X., & Sun, Y. (2022). Special issue on multi-modal information learning and analytics on big data. *Neural Computing and Applications, 34*, 3299–3300. https://doi.org/10.1007/s00521-021-06363-2
13. Cheng, W. X., Gao, R., Suganthan, P. N., & Yuen, K. F. (2022). EEG-based emotion recognition using random convolutional neural networks. *Engineering Applications of Artificial Intelligence, 116*, 105349. https://doi.org/10.1016/j.engappai.2022.105349
14. Al-Zoghby, A. M., Al-Awadly, E. M., Ebada, A. I., & Awad, W. A. (2025). Overview of multimodal machine learning. *ACM Transactions on Asian and Low-Resource Language Information Processing, 24*, 1–20. https://doi.org/10.1145/3701031
15. Kumar, P. S., Govarthan, P. K., Gadda, A. A. S., Ganapathy, N., & Ronickom, J. F. A. (2024). Deep learning-based automated emotion recognition using multimodal physiological signals and time-frequency methods. *IEEE Transactions on Instrumentation and Measurement, 73*, 1-12. https://doi.org/10.1109/TIM.2024.3420349
16. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion, 37*, 98-125. https://doi.org/10.1016/j.inffus.2017.02.003
17. Dzedzickis, A., Kaklauskas, A., & Bucinskas, V. (2020). Human emotion recognition: Review of sensors and methods. *Sensors, 20*, 592. https://doi.org/10.3390/s20030592
18. Malik, S. S., Ilyas, M., Haq, Y. U., Sana, R., Razzaq, M. S., Maqbool, F., & Pathan, M. S. (2025). Multi-modal emotion detection and sentiment analysis. *IEEE Access, 13*, 59790-59810. https://doi.org/10.1109/ACCESS.2025.3552475
19. Dewi, C., Gunawan, L. S., Hastoko, S. G., & Christanto, H. J. (2024). Real-time facial expression recognition: advances, challenges, and future directions. *Vietnam Journal of Computer Science, 11*, 167-193. https://doi.org/10.1142/S219688882330003X
20. Priyadarshini, N., & Aravinth, J. (2023, May). Emotion Recognition based on fusion of multimodal physiological signals using LSTM and GRU. In *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICSCCC58608.2023.10176510
21. Dalal, D., Talreja, D., Vaidya, D., Narvekar, M., & Ghag, K. (2023, October). Comparing the Effects of Various Preprocessing Techniques on the Performance of CNN for Facial Emotion

Recognition. In *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICACTA58201.2023.10392689

22. Chen, P., Li, J., Peng, B., Liu, Z., & Zhou, L. (2025). A 1-Dimensional Physiological Signal Prediction Method Based on Composite Feature Preprocessing and Multi-Scale Modeling. *Sensors*, *25*, 6726. https://doi.org/10.3390/s25216726
23. Wu, Z., Gan, J., Liu, J., & Wang, J. (2023, November). A multimodal emotion recognition method based on multiple fusion of audio-visual modalities. In *Proceedings of the 2023 5th International Conference on Video, Signal and Image Processing* (pp. 108-114). https://doi.org/10.1145/3638682.3638698
24. Patel, P., R, R., & Annavarapu, R. N. (2021). EEG-based human emotion recognition using entropy as a feature extraction measure. *Brain informatics*, *8*, 20. https://doi.org/10.1186/s40708-021-00141-5
25. Das, R., & Singh, T. D. (2023). Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, *55*, 1-38. https://doi.org/10.1145/3586075
26. Zhao, K., Zheng, M., Li, Q., & Liu, J. (2025). Multimodal sentiment analysis-a comprehensive survey from a fusion methods perspective. *IEEE Access*, *13*, 64556-64583. https://doi.org/10.1109/ACCESS.2025.3554665
27. Molino-Minero-Re, E., Aguileta, A. A., Brena, R. F., & Garcia-Ceja, E. (2021). Improved accuracy in predicting the best sensor fusion architecture for multiple domains. *Sensors*, *21*, 7007. https://doi.org/10.3390/s21217007
28. Abdulrahman, R., Jamil, A., Amjad, A., Hussain, S., Azhar, M., Aslam, Z., Shabbir, I., Ahmad, W., Mansab, A. A., Akbar, M. H., & Waqas, M. (2025). Automated Deep Learning Approaches for Multimodal Emotion Recognition: A Review of Fusion Strategies, Modalities and Architectures. *Machines and Algorithms*, *4*, 198-214. https://doi.org/10.66108/mna.v4i3.103
29. Peña, D., Aguilera, A., Dongo, I., Heredia, J., & Cardinale, Y. (2023). A framework to evaluate fusion methods for multimodal emotion recognition. *IEEE Access*, *11*, 10218-10237. https://doi.org/10.1109/ACCESS.2023.3240420
30. Sanku, R., Singireddy, S., Nandini, M. R., Dhanamalar, M., & Soni, M. (2025). Comprehensive Insights Into Multimodal Emotion Recognition Using Machine Learning and Deep Learning. In *2025 International Conference on Communication, Computer, and Information Technology (IC3IT)* (pp. 01-08). IEEE. https://doi.org/10.1109/IC3IT66137.2025.11340782
31. Kalateh, S., Estrada-Jimenez, L. A., Nikghadam-Hojjati, S., & Barata, J. (2024). A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges. *IEEE Access*, *12*, 103976-104019. https://doi.org/10.1109/ACCESS.2024.3430850
32. Zhu, X., Liu, Z., Cambria, E., Yu, X., Fan, X., Chen, H., & Wang, R. (2025). A client–server based recognition system: Non-contact single/multiple emotional and behavioral state assessment methods. *Computer Methods and Programs in Biomedicine*, *260*, 108564. https://doi.org/10.1016/j.cmpb.2024.108564
33. Zhu, X., Guo, C., Feng, H., Huang, Y., Feng, Y., Wang, X., & Wang, R. (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
34. Wehrli, S., Hertweck, C., Amirian, M., Glüge, S., & Stadelmann, T. (2022). Bias, awareness, and ignorance in deep-learning-based face recognition. *AI and Ethics*, *2*, 509-522. https://doi.org/10.1007/s43681-021-00108-6
35. Serna, I., Morales, A., Fierrez, J., & Obradovich, N. (2022). Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence*, *305*, 103682. https://doi.org/10.1016/j.artint.2022.103682

36. Ramaswamy, M. P. A., & Palaniswamy, S. (2024). Multimodal emotion recognition: A comprehensive review, trends, and challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*, e1563. https://doi.org/10.1002/widm.1563
37. Zhao, S., Jia, G., Yang, J., Ding, G., & Keutzer, K. (2021). Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, *38*, 59-73. https://doi.org/10.1109/MSP.2021.3106895
38. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010, June). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshop*s (pp. 94-101). IEEE. https://doi.org/10.1109/CVPRW.2010.5543262
39. Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2015). Challenges in representation learning: A report on three machine learning contests. *Neural networks*, *64*, 59-63. https://doi.org/10.1016/j.neunet.2014.09.005
40. Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2011). Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, *3*, 18-31. https://doi.org/10.1109/T-AFFC.2011.15
41. Sharma, G., & Dhall, A. (2020). A survey on automatic multimodal emotion recognition in the wild. In *Advances in data science: Methodologies and applications* (pp. 35-64). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-51870-7_3