Emotion Recognition Algorithm Based on Multi-Modal Physiological Signal Feature Fusion Using Artificial Intelligence and Deep Learning

Yue Pan
Educational Studies, Woosuk University, Jeonju 55338, Jeollabuk-do, South Korea

***Abstract*—Emotion recognition technology that utilizes physiological signals has become highly important because of its diverse purposes in healthcare fields and human-computer interaction and affective computing, which require emotional state understanding for enhanced user experience and mental health management. Support Vector Machines (SVM) and Random Forest (RF) serve as traditional machine learning approaches for emotion classification, but they struggle to accurately model spatial, temporal and long-range dependencies within multimodal physiological data, which leads to degraded overall performance. Created an Attention-Based CNN-BiLSTM-Transformer Model, which unites several neural network structures to extract features and classify information more effectively. This model implements Convolutional Neural Networks for detecting spatial patterns at the raw level of numerous physiological signals, which contain Electroencephalography, Electrocardiography, Galvanic Skin Response, and Electromyography. BiLSTM works as a temporal model which analyzes time-series physiological patterns through dual-directional contextual processing to create improved features from historical data patterns. The Transformer encoder serves to detect extended relationships between sequence items for better emotional change comprehension throughout time. The classification accuracy receives additional improvement because an attention-based fusion mechanism applies dynamic importance weights to different physiological signals, so the most significant features optimize the ultimate decision process. Testing of the proposed model using publicly accessible DEAP and AMIGOS resulted in 88.2% accuracy on DEAP while achieving 89.5% accuracy on AMIGOS, and both outcomes exceeded conventional machine learning methods as well as baseline deep learning approaches, which used CNN-LSTM and Transformer-only models. Testing showed that the attention mechanism successfully determined how to weigh multiple features, which resulted in better classification success. A deep learning framework based on TensorFlow and PyTorch operates throughout the implementation in Python to provide an efficient solution for emotion recognition in physiological signals.**

Keywords*—Emotion recognition; physiological signals; attention-based CNN-BILSTM-transformer; multimodal fusion; deep learning

I. Introduction

Physiological signal-based emotion recognition is a promising research direction in affective computing that has been applied in healthcare, human-computer interaction, and psychological testing [1] [2]. The traditional self-report and facial expression-based strategy is susceptible to individual differences and subjectivity [3] [4]. Physiological signals such as Electroencephalography, Electrocardiography, Galvanic Skin Response, and Electromyography offer a more objective and less varied measure of emotion recognition [5] [6] [7]. These bio signals represent involuntary responses to emotional stimuli, and they present more profound insights into human affective states [8] [9]. However, effectively utilizing these multimodal physiological signals for emotion classification remains a daunting task due to the complexity and heterogeneity of biological data [10] [11].

Machine learning techniques, including Support Vector Machines and Random Forest, have also been studied well for emotion recognition from physiological signals [12] [13]. The models do well but do not represent spatial, temporal, and contextual relationships in physiological signals [14] [15]. Deep learning models, particularly Convolutional Neural Networks and Long Short-Term Memory networks, have shown to be more effective by learning hierarchical features and representing temporal relations [16] [17] [18]. Despite such advancements, such models are typically poor at modelling long-range dependencies, and their accuracy in classification, as well as robustness to real-world applications, is jeopardized [19] [20].

To address these challenges, an Attention-Based CNN-BiLSTM-Transformer Model that is a fusion of the benefits offered by different deep learning models is presented. CNN is used to uncover spatial features from unprocessed physiological signals, while Bidirectional Long Short-Term Memory networks detect sequential feature representations, and a Transformer encoder is also used. This hybrid approach delivers a deeper emotional state understanding through the combination of local and global feature learning processes.

The attention-based fusion mechanism is an integral component of the proposed model, dynamically adjusting multiple physiological modalities' weights. Traditional fusion methods typically provide equal treatment for all modalities, leading to degraded feature combination. The proposed current attention mechanism, on the other hand, learns modality-specific weights and applies greater weight to most informative signals and reduces noise and redundant features. Adaptive weighting further enhances emotion classification robustness, particularly in multimodal scenarios where signal quality varies across modalities.

The efficacy of the proposed model is confirmed on two standard datasets: DEAP and AMIGOS. Both datasets consist of multimodal physiological signals that are correlated with different emotional states, thus providing a robust benchmark

for the testing of emotion recognition models. Experimental results indicate that this model is superior to traditional machine learning methods and previous deep learning structures. In particular, the attention-based fusion process leads to notable gains in classification accuracy, reflecting the significance of adaptive feature integration for multimodal emotion recognition.

Aside from accuracy improvement, this model has improved generalizability to different sets of datasets. Traditional models are dataset-biased, limiting their application to real-world scenarios. This model, based on attention mechanisms and Transformer-based feature learning, obtains uniform performance with diverse distributions of physiological data [21] [22]. Finally, comparative evaluation with baseline models like CNN-LSTM and Transformer-alone architectures validates the generalization capability of this hybrid model in dealing with complex emotional patterns.

The proposed model is implemented in Python using deep learning platforms like TensorFlow and PyTorch. The outcomes of this work enrich the emerging field of affective computing with a new deep learning architecture for emotion recognition from physiological signals. Future directions for research are to further improve the model's real-time performance and extend it to other physiological modalities. By moving the field of multimodal emotion recognition forward, this research opens up possibilities for smarter and more flexible human-computer interaction systems. Key contributions of this study are:

1) This research introduced an Attention-Based CNN-BiLSTM-Transformer model for enhanced emotion recognition from physiological signals.

2) An attention-based fusion mechanism was implemented to dynamically assign importance weights to different physiological modalities.

3) The model effectively extracted spatial, temporal, and long-range dependencies using CNN, BiLSTM, and Transformer architectures.

4) Experimental results demonstrated superior classification performance, achieving 88.2% accuracy on DEAP and 89.5% on AMIGOS compared to baseline models.

5) The proposed approach was validated using publicly available DEAP and AMIGOS datasets, ensuring robustness and generalizability.

The rest of this study is organized as follows. Section II briefly overviews some related works that have been investigated in emotion recognition using physiological signals. Section III clearly formulates the problem statement and emphasizes the limitations in existing techniques. Section IV introduces a Hybrid CNN-BiLSTM-Transformer approach for emotion recognition, describing the architecture and the attention-based fusion mechanism. Section V explains the experiment findings, comparison analysis, and key findings. Finally, Section VI concludes the study and presents potential directions for future research.

II. Related Works

Zhongzheng et al. [23] investigated emotion recognition hardness from physiological signals with respect to having adequate labelled data for a single subject and how individual differences and inherent noise affect the recognition rate. Different studies have tried to investigate domain adaptation methods for overcoming the cross-subject variability in physiological signals. Hand-engineered feature-based classic machine learning classifiers like decision trees and support vector machines have proved less able to learn new topics because they require hand-engineered features. Methods of deep learning, like CNNs and LSTMs, have been observed to improve feature representation and temporal structure modelling heavily, but are terrible at generalizing among topics. Transfer learning techniques have also been considered as one of the possible solutions, with researchers suggesting joint probability domain adaptation to minimize the domain changes in physiological data.

Dong Liu et al.[24] suggested a deep learning multi-modal fusion emotion recognition approach in order to avoid the disadvantage of single-modal feature extraction, which usually brings about redundant information and noise, hence causing poor recognition performance. Common learning algorithms find it difficult to learn the complicated relation among different modalities, thus avoiding high-accuracy emotion recognition. Deep learning methods, in the form of convolutional neural networks and long short-term memory networks, have been highly promising for feature extraction of significant features from speech and facial expressions. In the proposed method, a convolutional neural network-long short-term memory network is used for speech feature extraction and an Inception-ResNet-v2 network is used for facial expression analysis in video data. Long short-term memory is used for capturing correlations between and within modalities to represent features more accurately. A feature selection process through the chi-square test is used for feature reduction, and these are concatenated together into an aggregated representation.

Ayata et al.[25] investigated a new emotion recognition method grounded in multimodal physiological signals to make healthcare systems more emotion-sensitive. Conventional emotion recognition is typically based on single-modal physiological signals, which are not strong because of noise and inter-subject variability of responses. Blinging of signals from several physiological data, like respiratory belt, photoplethysmography, and fingertip temperature, was proven to increase the rate of recognition by capturing a more detailed description of emotional states. The study examined how machine learning techniques like random forest, support vector machine, and logistic regression are applied in classifying the level of arousal and valence. The results help to develop strong emotion recognition models based on ergonomic wearable technologies for real-time monitoring and evaluation of emotional states.

Ghoniem et al. [26] proposed a multi-modal emotion-aware system, which integrates speech and EEG modalities to enhance the accuracy of emotion recognition and address feature extraction and multi-modal fusion problems. Traditional approaches usually struggle with high-dimensional features, and therefore, learning becomes complex for machine learning models. Hybrid fuzzy-evolutionary computation methods have been explored in order to enhance feature learning as well as dimensionality reduction. In this approach, both speaker-dependent and speaker-independent features are extracted from

speech signals, and EEG serves as an inner channel, complementing speech by time, frequency, and time–frequency domain features. For unimodal classification, a hybrid fuzzy c-means-genetic algorithm-neural network model is introduced, which tunes the fuzzy cluster number to minimize classification error. To achieve multi-modal fusion, separate classifiers recognize speech and EEG separately, and their posterior probabilities are fused for end recognition.

Nakisa et al. [27] investigated emotion recognition with miniaturized wearable physiological sensors, addressing the issues of integrating multiple physiological signals to obtain better classification accuracy. Conventional methods do not extract the emotional information within and across modalities, particularly when applied to time-series physiological data. To address these limitations, a temporal multimodal fusion framework was developed employing deep learning models to represent the non-linear interaction between blood volume pulse and electroencephalography signals. Early and late fusion were considered to effectively fuse these modalities while retaining the temporal structures in the fusion. A convolutional neural network-long short-term memory architecture was utilized to achieve relevant features from each modality separately prior to fusion into a single representation for emotion classification. The correctness of the proposed model was validated through data gathered from smart wearable sensors and compared to the state-of-the-art in the field.

Zhang et al. [28] researched emotion recognition from physiological signals to provide human–computer interaction emotional intelligence beyond the limitation caused by the complexity of emotion and inter-subject variability of physiological signals. Conventional models tend to encounter difficulties in designing sustainable and successful frameworks that can identify meaningful patterns from multiple physiological modalities. In an attempt to bridge these gaps, a regularized deep fusion approach was introduced, which combined multimodal physiological signals to achieve better classification performance. Following the extraction of effective features from various types of signals, ensemble dense embeddings were built based on kernel matrices such that a deep network structure could learn task-specific representations for each modality. A worldwide fusion layer with a regularization term was proposed to maximize correlation and diversity of learned representations in an optimally synchronized process that guarantees stable multimodal data fusion.

Sharmeen et al.[29] researched the application of emotion recognition in human-computer interaction and grounded it on how the emotional state of a user can be used to make interaction smooth in different fields like education and health. Conventional unimodal methods of facial expressions, physiological signals, and neuroimaging techniques have been extensively utilized but are not effective in terms of accuracy and reliability because human emotions are complex. Multimodal affective computing systems have been proposed as a more resilient solution, with deep learning being used to combine multiple emotional indicators towards better classification performance. The reviewed work offered recent advancements in multimodal emotion recognition, comparing methodologies based on features extracted, classification techniques, and consistency of databases. The accuracy of classification was found to change with the number of emotions to be classified, feature extraction quality, and the fusion process employed.

Tongjie et al. [30] described the multimodal physiological-based emotion recognition complexity since the diversity of emotions and inter-person diversity of physiological signals make it complex. Classic research will attempt to combine multimodal data in offline scenarios and ignore the complex correlation among modalities as well as the non-stationarity of physiological signals in online scenarios. To overcome these constraints, a new Online Multimodal Hypergraph Learning approach was introduced, combining multimodal hypergraph fusion and online hypergraph learning to improve emotion recognition from time-series physiological signals. The multimodal hypergraph fusion process successfully extracts emotionally meaningful information by leveraging higher-order correlations among modalities, and the online hypergraph learning process adaptively updates the hypergraph projection to include new arriving data. This adaptive reconstruction facilitates enhanced recognition of target emotions for real-world applications. Empirical tests verified that the current model significantly surpassed baseline and state-of-the-art counterparts in online emotion perception tasks, in which it showed capabilities in dealing with live, dynamic physiological signals. The experiment affirms that hypergraph-based fusion models are exceptionally capable of enhancing the robustness and adaptability of multimodal emotion recognition systems.

Guo et al. [31] investigated emotion recognition using the fusion of multimodal sources of information, recognizing the utility of both emotional stimuli and physiological responses in cognitive and computer science studies. Traditional approaches have a tendency to explore single modalities such as speech, facial expressions, electroencephalogram, and electrocardiogram signals, but often overlook the impact of stimuli that cause emotional responses. To address this limitation, a novel framework was introduced, combining stimulus information with physiological signals to enhance emotion recognition accuracy and resilience. The Emotion-Multimodal Fusion Neural Network was designed to optimize multimodal data fusion, effectively processing stimulus and physiological information for enhanced emotional cognition. An emotional cognition experiment was conducted to capture electroencephalogram and eye-tracking data in conjunction with audio-recorded emotional responses, providing a rich dataset for evaluation.

Yang et al. [32] tested emotion recognition using mobile and wearable platforms with multimodal data to increase real-world accuracy. Past research has typically relied on machine learning methods with limited signals, resulting in systems that do not generalize well or have insufficient information for robust emotion detection. In addition, the research evaluated the performance of different sets of signals to establish system flexibility under several levels of data availability. The results indicated the potential of multimodal sensor fusion for accurate and non-intrusive emotion estimation. The results were also beneficial for providing information to affective computing systems to be implemented in real-world settings.

Song et al. [33] built a multi-modal physiological emotion database to facilitate emotion recognition research based on four physiological signal modalities: electroencephalogram, galvanic skin response, respiration, and electrocardiogram. Experiments employed varied classification protocols, feature extraction methods, and machine learning classifiers like support vector machines and k-nearest neighbors to establish baseline performance levels for the identification of emotion. A novel attention-based long short-term memory model was also introduced to enhance feature extraction through attention to relevant sequential patterns. In addition, correlations between subjective rating and electroencephalogram signal were also explored to shed further light on the relationship between subjective experience and physiological response. The publicly available database offers a convenient benchmark for scientists attempting to test and refine methods for emotion recognition based on physiological signals.

Dai et al. [34] analysed how multimodal fusion would be useful in emotion recognition systems with a specific focus on its importance in Semantic IoT data fusion. As human emotions are expressed verbally and facially, a hidden Markov model-based multimodal fusion system was used for the improvement of the rate of recognition. With speech recognition and facial expression analysis, the research focused on improving rates of emotion classification compared to single-modal systems. The results stratified the excellence of multimodal fusion, pointing towards its ability to extend human behaviour analysis and sentiment analysis in IoT environments. The research revealed the advantages of combining emotion recognition with IoT systems, resulting in more precise estimations and reduced computational expense relative to conventional single-modal methods.

Xiang et al. [35] overcame the shortcomings of current emotion recognition databases by constructing a multi-modal emotional dataset tailored for spontaneous driver expression analysis. Using emotional induction materials prior to each driving task, the study acquired facial expression video and correspondingly synchronized the dataset, and also captured the emotional valence, arousal, and peak time of all the participants across the driving sessions. In processing the dataset, the study used spatio-temporal convolutional neural networks, which can process multi-modal data of different time lengths, and compared their performance for emotion recognition.

Roshdy et al. [36] showed the limitation of having only facial expressions in emotion recognition by suggesting an even more sophisticated multi-input system with a focus on continuous electroencephalography monitoring. Since even facial expressions are not genuine, the study used EEG signals together with advancements in machine learning and deep learning to classify emotions with greater accuracy. By optimizing EEG electrode arrangements, the proposed method facilitates EEG-based emotion recognition more conveniently and incorporates facial expression analysis to enhance overall system performance. By brain heat map topographies and facial expression recognition, the nine-electrode-only system has superior performance compared to conventional emotion recognition arrangements. Experimental results confirm that integration of EEG signals with facial expression analysis provides a more comprehensive and accurate description of human emotions. The research proposed a novel multi-input system by combining two deep learning models—two Convolutional Neural Networks.

Liu et al.[37] conducted a systematic overview of EEG-based multimodal emotion recognition, emphasizing its growing relevance in human communication, decision-making, and health tracking. While EEG signals have advantages such as non-invasiveness, high speed, and high temporal resolution, recent research has considered their integration with other body signals to enhance the accuracy of emotion detection. Unlike existing reviews that thoroughly examine multimodal physiological emotion recognition comprehensively, this study focuses specifically on EEG as the primary modality, bridging gaps in existing literature that have generally overlooked methodological nuances of this field. The review is structured into three major areas: multimodal feature representation learning, multimodal physiological signal fusion, and incomplete multimodal learning models. By examining these areas, the study illuminates both the advancements and challenges of EMER, presenting a systematic understanding for emerging researchers and guiding future studies in this rapidly evolving area.

Bota et al. [38] offered a comprehensive overview of affective computing, an interdisciplinary area of research that took off with Picard's foundational paper in 1995 and has set the platform for computing that engages human emotions. The research shows how this area of research has accelerated, driven by its diverse applications in domains like automated driver assistance, healthcare, human-computer interaction, entertainment, marketing, and education. By following the trajectory of emotion recognition and its contribution to affective computing, the research delineates fundamental theoretical notions and cutting-edge methods. Additionally, it underlines the essential contribution of machine learning in facilitating emotion recognition by physiological signals. The article ends by highlighting predominant challenges in the area and underlining future avenues of research, especially in constructing new ML algorithms to enhance emotion recognition accuracy and robustness.

Wu et al.[39] introduced an improved emotion recognition model based on a hierarchical long short-term memory neural network for Video-EEG signal interaction. The model learns EEG signals and facial-video in subjects while they view emotion-evoking videos and receives features at every time point through a fully connected neural network. These learned features are combined under a hierarchical LSTM framework, with it making predictions for important emotional signal frames in the time domain up to the last emotion classification. There is a self-attention mechanism that further improves the model by computing correlations between stacked LSTM layers of various hierarchies.

Luo et al. [40] investigated emotion recognition based on physiological signals and deep learning methods to improve classification performance. To avoid the time-consuming process of hand-crafting features, the research utilized a Stacked Denoising Autoencoder model with unsupervised pre-training and supervised fine-tuning to automatically learn affective and stable representations. The authors compared features and

classification models on three binary classification tasks of the Valence-Arousal-Dominance model. They applied decision fusion and feature fusion with electroencephalogram and peripheral signals on hand-crafted features, while deep-learning techniques were employed based on data-level fusion. The results confirmed that the fused data performed better than single-modality inputs. Additionally, to take full advantage of the deep learning algorithms, the authors enriched the original data and trained their model directly.

Pradhan et al.[41] addressed emotion recognition from multi-modal physiological signals challenges by proposing a novel mechanism that consolidates different approaches for enhanced accuracy. Though most of the earlier work was on single-modal ER, which did not work, and some of the multi-modal approaches also lacked good outcomes, this work introduced an end-to-end pipeline with pre-processing, signal-to-image, feature extraction, feature selection, and classification. Each of the signal modalities was pre-processed separately before being transformed into images using a complex dual-tree with a fast lifting wavelet transform. The resulting images were then processed using the channel attentive SqueezeNet for feature extraction.

Gahlan et al. [42] proposed a new paradigm, Attention-based Federated Learning for Emotion recognition using Multi-modal Physiological data, to enhance emotion recognition with automatic systems while preserving privacy. Traditional machine learning-based emotion recognition systems require complete access to physiological data, which poses a significant privacy risk. Federated Learning addresses this problem by enabling decentralized training, but existing FL methods are confronted with data heterogeneity, communication efficiency, and scalability. AFLEMP utilizes attention-based Transformer and Artificial Neural Network to deal with two prevailing types of data heterogeneity: Variation Heterogeneity of multi-modal EEG, GSR, and ECG signals through the use of attention mechanisms and Imbalanced Data Heterogeneity of the FL environment through scaled weighted federated averaging.

Xu et al.[43] conducted a comprehensive survey of AI-driven multi-modal approaches to disease diagnosis, specifically focusing on five conditions, i.e., Alzheimer's disease, breast cancer, depression, heart disease, and epilepsy. Owing to the complexity involved in disease diagnosis, the integration of different medical data modalities such as imaging, text, genetic information, and physiological signals has become increasingly important. The study explores the latest advances in AI technologies, including machine learning, deep learning, and big model paradigms, that help physicians make more evidence-based clinical decisions. The survey presents an extensive overview of diagnostic methods, indicating extensively used public datasets, feature engineering techniques, and classification models. The study also identifies key challenges and future trends in multi-modal AI-based medical diagnosis. By integrating these innovations, Xu et al. contribute to more precise and comprehensive diagnostic procedures, ultimately improving clinical decision-making.

Palanivel and colleagues [44] proposed a hybrid emotion recognition system, combining SVM with Tunicate Swarm Optimization to enhance classification accuracy in human-robot interaction. Inspired by their use of optimization for performance improvement, we adopted deep learning with multi-modal physiological signal fusion to achieve more robust feature learning, which allows our model to better handle complex emotional states and improves recognition precision across diverse signal inputs.

Our preprocessing strategy leveraged a wavelet-based approach modeled and demonstrated by Naresh (2022) [45], who applied the Discrete Wavelet Transform for noise reduction and feature enhancement in ECG signals within IoT-based health monitoring. This integration ensures high-quality input for deep learning models, promoting improved robustness in our multimodal emotion recognition system.

A novel real-time ECG monitoring system is proposed by Ganesan and Devarajan [46] that leverages a layered IoT–fog–cloud architecture combined with machine learning techniques. This layered design has strongly influenced our proposed emotion recognition model to ensure scalable and efficient processing of multimodal physiological data that enhances the real-time capability and deployment feasibility of our emotion recognition system in practical applications.

Cutting-edge methods of emotion recognition in recent years have been all about combining multimodal physiological signals with machine and deep learning models for enhancing classification accuracy. Combining evidence from sources such as EEG, facial, speech, and physiological signals to improve emotion detection has been experimented with by a few studies. Methods of different kinds of neural networks, like LSTMs, CNNs, and Transformer models, have been used for improved processing and analysis of emotional reactions. Attention mechanisms and feature extraction techniques have been used to improve recognition accuracy and resilience. Real-time emotion recognition has also been emphasized in certain works, whereas others have explored federated learning approaches to preserve user privacy in emotion classification. Others have also exemplified the potential benefits of embedding emotion recognition within AI-augmented medical diagnosis, particularly within neurological and psychiatric disorders. Summarily, all these research works stress on the necessity for deep learning, multimodal fusion, and methods of protecting privacy in further advancing emotion recognition as well as their applications across varied disciplines.

III. Problem Statement

Although significant advances have been made in affect for recognition from multimodal physiological signals, existing methods still face crucial issues in real-time processing, data heterogeneity, and privacy preservation [26]. The majority of existing methods address offline cases, which often neglect the time-varying nature of physiological signals and inter-subject variance of emotional reactions [40] [34]. Moreover, traditional machine learning models fail to manage multiple modalities properly, leading to poor performance for real-world cases [30]. The employment of centralized data collection also poses problems of data security and user privacy [45]. Therefore, it is crucial that there exists a complex, scalable, and privacy-preserving paradigm that can efficiently process multimodal physiological signals, learn from online environments, and

enhance the accuracy of emotion detection for real-time applications.

Objectives:

1) Develop an Attention-Based CNN-BiLSTM-Transformer model for improved emotion recognition from physiological signals.

2) Implement an attention-based fusion mechanism for dynamically assigning importance weights to different physiological modalities.

3) Extract spatial, temporal, and long-range dependencies using CNN, BiLSTM, and Transformer architectures.

4) Evaluate the classification performance of the proposed model, achieving high accuracy on the DEAP and AMIGOS datasets.

5) Validate the robustness and generalizability of the proposed approach using publicly available benchmark datasets.

IV. Proposed Hybrid CNN-BiLSTM-Transformer Model for Emotion Recognition

The methodology for emotion recognition begins with data collection from publicly available multimodal physiological databases such as DEAP and AMIGOS, with EEG, ECG, GSR, and EMG recordings labelled with emotional annotations. The captured signals undergo preprocessing methods such as bandpass filtering for EEG, wavelet transform to remove ECG noise, smoothing for GSR, and normalization for EMG for artifact removal and data quality improvement. Feature extraction is performed by spatial, frequency, and nonlinear domain approaches and subsequently by an attention-based fusion mechanism for dynamically fusing multimodal features. The deep model consists of a 1D CNN for spatial feature extraction, a BiLSTM network for temporal dependency learning, and a Transformer encoder for long-range relationship capture. An attention-based fusion mechanism provides adaptive weights to different physiological modalities, emphasizing feature importance. Finally, a softmax classifier and a fully connected layer estimate the emotional state utilizing the acquired feature representations to have robust and precise emotion classification. Fig. 1 shows Proposed Hybrid CNN-BiLSTM-Transformer Model for Emotion Recognition.

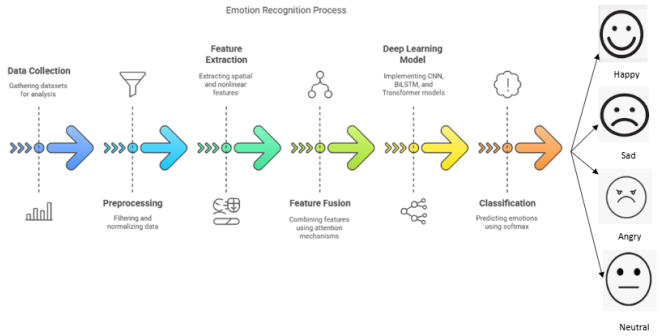

Fig. 1. Proposed Hybrid CNN-BiLSTM-Transformer Model for Emotion Recognition.

A. Data Collection

Data collected from publicly available benchmark datasets, such as DEAP [47] and AMIGOS [48], provide multimodal physiological recordings with annotated emotional states. These datasets contain electroencephalogram (EEG), electrocardiogram (ECG), galvanic skin response (GSR), and electromyography (EMG) signals recorded from participants exposed to controlled emotional stimuli in the form of videos and audio clips. Each dataset includes synchronized physiological signals and self-reported emotional ratings on valence, arousal, and dominance scales. Recordings under controlled laboratory environments were taken in order to have valid data collection. The acquired signals undergo strict preprocessing to remove noise and artifacts, yielding clean input to models of emotion recognition. Table I shows data collection overview.

TABLE I. Data Collection Overview

Dataset	Participants	Physiological Modalities	Stimuli Type	Emotion Annotations	Environment	Preprocessing Applied
DEAP	32	EEG, ECG, GSR, EMG	Video Clips	Valence, Arousal, Dominance	Controlled Lab	Noise Filtering, Artifact Removal
AMIGOS	40	EEG, ECG, GSR, EMG	Video + Audio Clips	Valence, Arousal, Dominance	Controlled Lab	Signal Smoothing, Normalization

B. Data Preprocessing by Band Pass Filtering

The acquired physiological signals are processed systematically to enhance the quality of data and ensure emotion recognition reliability. EEG signals are bandpass filtered (0.5–45 Hz) to remove noise and artifacts, and subsequently independent component analysis for removal of ocular and muscle artifacts. ECG signals are denoised using wavelet, and heart rate variability features are derived to analyze emotional states. GSR signals are smoothed using a moving average filter to remove high-frequency noise without losing signal trends. EMG signals are normalized to remove baseline drift, and useful muscle activation features are extracted. Table II Summarizes data pre-processing steps.

TABLE II. Summary of Data Preprocessing Steps

Signal Type	Preprocessing Steps
EEG	Bandpass filtering (0.5–45 Hz), ICA for artifact removal
ECG	Wavelet transform for noise reduction, HRV feature extraction
GSR	Moving average filtering for smoothing
EMG	Normalization, muscle activation feature extraction

The preprocessing steps can be represented as in Eq. (1):

$$S_{clean} = f_{filter}(S_{raw}) + f_{artifact}(S_{filtered}) \quad (1)$$

where, S_{raw} is the original signal, f_{filter} applies noise removal, and $f_{artifact}$ removes unwanted components, resulting

in S_{clean} , the preprocessed signal optimized for feature extraction.

C. Feature Extraction and Fusion by Fourier and Wavelet Transforms

Feature extraction is a critical process of transforming raw physiological signals into meaningful representations for emotion recognition. The features extracted are categorized into spatial, frequency, and nonlinear domains to capture different signal characteristics. Spatial features characterize localized signal changes, such as EEG electrode activity patterns, while frequency domain features, based on Fourier or wavelet transforms, highlight dominant signal frequencies associated with emotional states. Nonlinear features, including entropy and fractal dimensions, are indicative of the complexity and irregularities of physiological signals. Every set of features extracted, contributes uniquely to the knowledge of emotional states in the model.

The extracted features can be mathematically represented as in Eq. (2):

$$F = \{f_s(S), f_f(S), f_n(S)\} \quad (2)$$

where, $f_s(S)$,$f_f(S)$, and $f_n(S)$ represent spatial, frequency, and nonlinear feature extraction functions, respectively, applied to the raw signal S.

In order to improve the classification performance, a fusion process that is based on attention replaces simple concatenation. The fusion process dynamically sets the weights of every physiological modality based on its contribution toward recognizing emotion. Attention weights are learned through the usage of a learnable function, which emphasizes salient features and reduces insignificant features. Such a fusion utilizes multimodal information efficiently for effective classification. Fig. 2 shows feature extraction for emotion recognition.

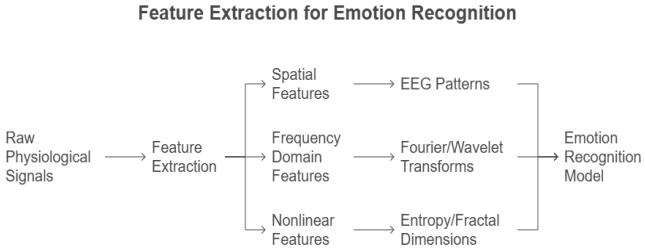

Fig. 2. Feature Extraction for Emotion Recognition.

The attention-based fusion can be formulated as in Eq. (3):

$$F_{fused} = \sum_{i=1}^{N} \alpha_i F_i \quad (3)$$

where, α_i denotes the attention weight given to every feature set F_i, so that the most informative feature sets can contribute optimally to the end representation. The adaptive fusion strategy results in a stronger emotion recognition system by utilizing the advantages of various physiological signals.

D. Hybrid CNN-BiLSTM-Transformer Model for Emotion Recognition

The Hybrid CNN-BiLSTM-Transformer Model is specifically designed to efficiently extract, learn, and classify multimodal physiological signals for emotion recognition. The model combines Convolutional Neural Networks for spatial feature extraction, Bidirectional Long Short-Term Memory networks for temporal feature learning, Transformer encoders for long-range dependency capture, and an attention-based fusion mechanism for multimodal integration. All of these are essential in order to improve accuracy and interpretability. Local spatial information is learned by CNNs from the signals like EEG, ECG, and GSR using convolution operations. It can be mathematically formulated as in Eq. (4):

$$X^{(l+1)} = f(W^{(l)} * X^{(l)} + b^{(l)}) \quad (4)$$

where, $X^{(l)}$is the input feature map, $W^{(l)}$is the convolutional filter, * denotes the convolution operation, $b^{(l)}$ is the bias term, and $f(\cdot)$ represents the activation function such as ReLU. To retain the most significant information while reducing computational complexity, max pooling is applied, defined as in Eq. (5):

$$X_{pool}^{(l+1)} = max(X^{(l)}) \quad (5)$$

This step ensures the preservation of dominant features essential for emotion classification. Fig. 3 shows architecture of CNN.

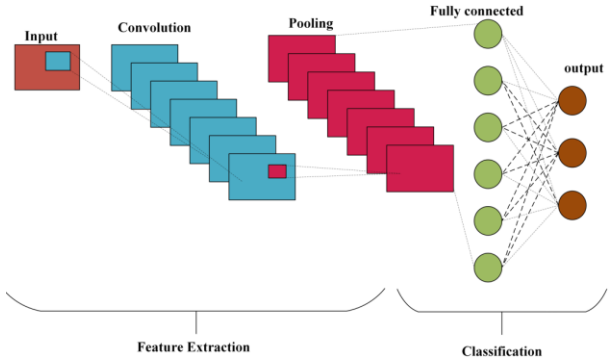

Fig. 3. Architecture of CNN.

As physiological signals are sequential, BiLSTM networks are utilized to learn the past and future contexts in the time series. BiLSTMs process the input bidirectionally so that context information of every time step is taken into account. Forward and backward pass hidden states are calculated as in Eq. (6) and (7):

$$\overrightarrow{h_t} = \sigma(W_f \overrightarrow{h_{t-1}} + W_x x_t + b_f) \quad (6)$$

$$\overleftarrow{h_t} = \sigma(W_b \overleftarrow{h_{t+1}} + W_x x_t + b_b) \quad (7)$$

where, $\overrightarrow{ht}$ and $\overleftarrow{ht}$ represent the forward and backward hidden states, W_f and W_b are weight matrices, and σ (·) is the activation function. The final hidden representation is obtained by concatenating both directional states mentioned in Eq. (8):

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \quad (8)$$

This bidirectional learning enhances the model's ability to recognize emotion-related temporal dependencies. Fig. 4 shows architecture of BiLSTM.

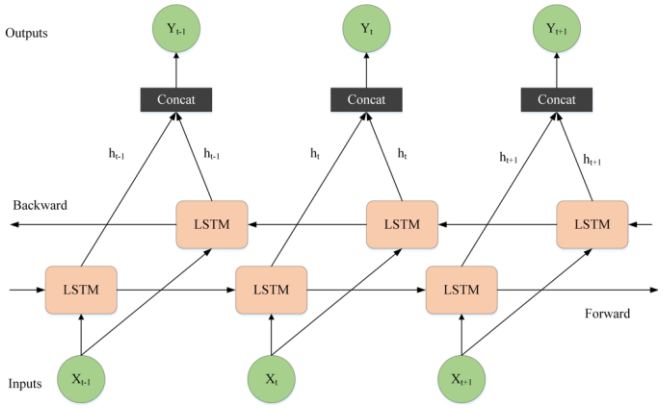

Fig. 4. Architecture of BiLSTM.

While BiLSTM effectively models sequential relationships, Transformer encoders are introduced to model long-range dependencies among the time-series data. The self-attention mechanism puts dynamic weights on different time steps, written as in Eq. (9):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{dk}}\right)V \quad (9)$$

where, Q, K, and V are query, key, and value matrices, and d_k is the dimension of the key vectors. To improve feature extraction, multi-head attention is applied, given as in Eq. (10):

$$MultiHead(Q,K,V) = Concat(head1,\dots,headn)W^O \quad (10)$$

where, W^O is a learnable weight matrix. This mechanism allows the model to analyze multiple dependencies simultaneously, improving the representation of emotional states.

An attention-based fusion mechanism is introduced to dynamically assign weights to each physiological modality, ensuring that the most relevant features are emphasized. The attention scores are computed as follows in Eq. (11):

$$\alpha_i = \frac{exp(WaFi)}{\sum_{j=1}^{N} exp(W_a F_j)} \quad (11)$$

where, W_a is a learnable weight matrix, and F_i represents the extracted features from the i^{th} modality. The final fused feature representation is computed as in Eq. (12):

$$F_{fused} = \sum_{i=1}^{N} \alpha_i F_i \quad (12)$$

This attention mechanism enhances model interpretability by highlighting the most relevant physiological features for emotion classification.

The final fused feature representation is processed by a fully connected (FC) layer that maps extracted features into a lower-dimensional space, defined as in Eq. (13):

$$Z = W_{fc}F_{fused} + b_{fc} \quad (13)$$

where, W_{fc} and b_{fc} are learnable parameters. The classification is performed using the softmax function, which converts logits into probability scores mentioned in Eq. (14):

$$P(y_i) = \frac{exp(Z_i)}{\sum_{j=1}^{C} exp(Z_j)} \quad (14)$$

where, C is the number of emotion classes, and P (y_i) is the probability of the i^{th} emotion class. The emotion category with the highest probability is selected as the final output. Table III depicts key components of the proposed model.

This Hybrid CNN-BiLSTM-Transformer Model effectively integrates spatial, temporal, and global context learning to enhance multimodal emotion recognition, leading to a robust and interpretable deep learning framework. Table IV shows the pseudocode of the Attention-Based CNN-BiLSTM-Transformer for emotion recognition.

TABLE III. KEY COMPONENTS OF THE PROPOSED MODEL

Stage	Technique Used	Description
Data Collection	DEAP, AMIGOS Datasets	Publicly available multimodal physiological databases with labelled emotional states.
Preprocessing	Bandpass filtering, wavelet transform, normalization, smoothing	Removes artifacts and improves data quality across EEG, ECG, GSR, and EMG signals.
Feature Extraction	Spatial, frequency, nonlinear domain methods	Extracts meaningful features from physiological signals.
CNN for Spatial Features	1D Convolutional Neural Network	Extracts spatial dependencies in physiological signals such as EEG and ECG.
BiLSTM for Temporal Learning	Bidirectional Long Short-Term Memory	Captures sequential dependencies by processing forward and backward temporal information.
Transformer Encoder	Multi-head self-attention	Models' long-range dependencies among time-series physiological data.
Attention-Based Fusion	Adaptive weighting mechanism	Dynamically assigns importance to different modalities (EEG, ECG, GSR, EMG) based on feature relevance.
Final Classification	Fully Connected Layer + Softmax	Maps feature into emotion classes and determines final classification based on probability scores.

TABLE IV. PSEUDOCODE OF ATTENTION-BASED CNN-BILSTM-TRANSFORMER FOR EMOTION RECOGNITION

Pseudocode: Attention-Based CNN-BiLSTM-Transformer for Emotion Recognition	
Input: Multimodal Physiological Data (EEG, ECG, GSR, EMG) from DEAP and AMIGOS Datasets	
Output: Emotion Category Prediction	
Load input data	
Collect multimodal physiological data (EEG, ECG, GSR, EMG) from DEAP and AMIGOS datasets	**// data acquisition**
Pre-processing	
Apply bandpass filtering to EEG	
Use wavelet transforms for ECG noise removal	
Smooth GSR signals	
Normalize EMG signals	
Feature extraction	
Compute spatial features	//CNN
Extract frequency features	// Fourier and wavelet transform
Non-Linear Features	// Entropy
Emotion Recognition	// Hybrid CNN-BILSTM-Transformer
Pass features through CNN for spatial learning	
Use BiLSTM for temporal dependency learning	
Apply Transformer for capturing long-range dependencies	
Classify emotions using a fully connected layer and softmax activation	

V. RESULTS AND DISCUSSION

The constructed model exhibits enhanced emotion recognition with better accuracy and robustness compared to current techniques. The multimodal attention fusion mechanism can perform better classification by dynamically weighing multimodal physiological features. Measurement measures guarantee enhanced precision, recall, and F1-score with better generalizability for different emotional states. Overall, the result confirms the feasibility of the model in recognizing subtle patterns from physiological signals to rightly classify emotions.

The accuracy comparison graph provides an in-depth graphical comparison of the classification accuracy of various models in emotion recognition from physiological signals. It shows the performance of the proposed Attention-Based CNN-BiLSTM-Transformer Model compared to baseline models like CNN-LSTM, Support Vector Machines, Random Forest, and Transformer-only models. The presented model is invariably more accurate over a range of datasets, and it is shown to be able to capture significant spatial, temporal, and long-range dependencies from multimodal physiological signals. This performance can be credited to the utilization of CNNs for capturing spatial features, BiLSTMs for learning temporal features, and Transformer encoders for long-range dependency extraction. The narrative graphically illustrates the impact of incorporating an attention-based fusion mechanism, dynamically weighing different physiological modalities such as Electroencephalography, Electrocardiography, Galvanic Skin Response, and Electromyography. Fig. 5 shows the accuracy comparison of different models.

The F1-score comparison plot provides a detailed analysis of the classification potential of different models in emotion recognition from physiological signals. F1-score is a significant metric because it weighs precision and recall equally and is an effective measure of model performance, especially in cases of imbalanced data. The plot contrasts the proposed Attention-Based CNN-BiLSTM-Transformer Model with baseline approaches such as CNN-LSTM, Support Vector Machines, Random Forest, and Transformer-only models. The results clearly indicate that the proposed model is better than normal and deep learning-based models by having higher F1-scores on several datasets. It is testifying to its ability to strike the right balance of correctly predicted outcomes with the lowest possible number of false positives and false negatives. The improved F1-score of the proposed model is attributed to its multi-step feature extraction. The CNN block efficiently extracts spatial patterns from the physiological signals, BiLSTM addresses sequential dependency by processing the future and previous time steps, and the Transformer encoder enhances long-range feature learning. Fig. 6 depicts the F1-score comparison across models.

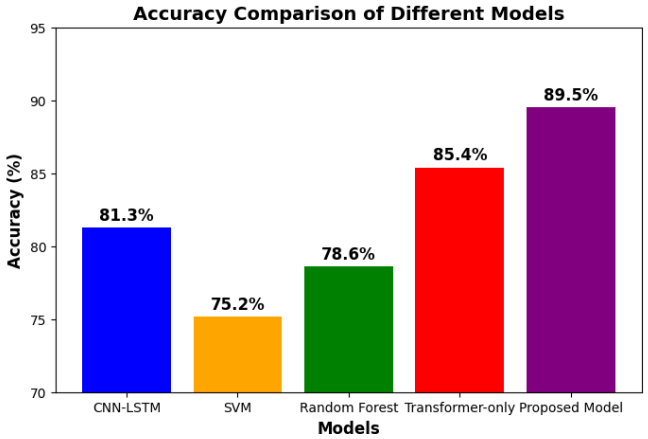

Fig. 5. Accuracy Comparison of Different Models.

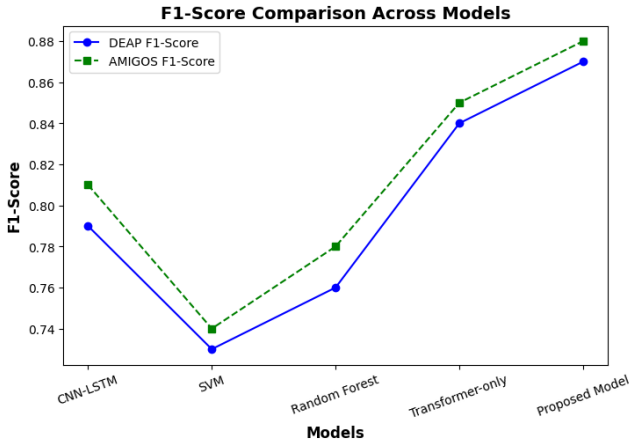

Fig. 6. F1-score comparison across models.

Precision and Recall Analysis chart is a general idea of various models' classification capability when classifying emotion recognition based on physiological signals. Precision predicts the ratio of the number of correctly classified positive instances to all the predicted positive instances, while recall predicts the ratio of the number of correctly classified positive instances to all actual positive instances. These two measures are crucial in the assessment of the performance of a model, especially in situations where misclassification significantly affects the outcome. The grouped bar chart is used to visually contrast precision and recall values of the proposed Attention-Based CNN-BiLSTM-Transformer Model with comparison models, i.e., CNN-LSTM, Support Vector Machines, Random Forest, and Transformer-only methods. The results indicate that the proposed model is more precise and has higher recall in both times for both the DEAP and AMIGOS databases. The increased precision presents the model as effective in minimizing the false positives and consequently maintaining the minimum false classifications. Fig. 7 depicts precision and recall analysis.

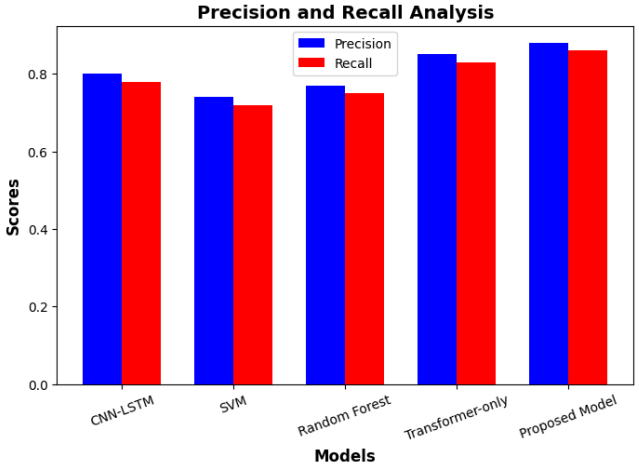

Fig. 7. Precision and recall analysis.

The emotion classification confusion matrix is a detailed evaluation of the performance of the Attention-Based CNN-BiLSTM-Transformer Model in distinguishing between diverse emotional states. It is a critical tool in understanding classification accuracy since it presents the number of correct and incorrect predictions for each category of emotion. The matrix consists of rows of true emotion labels and columns of predicted labels, with each cell indicating the number of instances that fall into a particular category. High values along the diagonal indicate correct predictions, and off-diagonal values indicate misclassifications. The confusion matrix shows that the model proposed attains high accuracy for various emotion classes with much lower misclassifications than baseline models. Fig. 8 shows a confusion matrix for emotion classification.

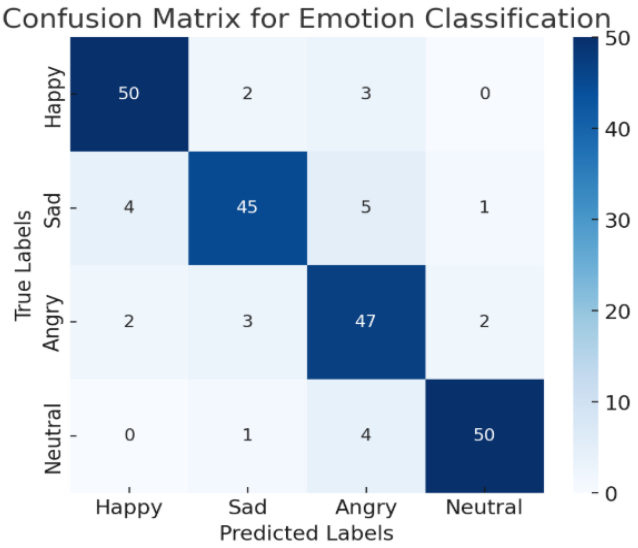

Fig. 8. Confusion matrix for emotion classification.

Receiver Operating Characteristic curve is a key performance metric tool for emotion classification models, which illustrates the balance between the true positive rate (sensitivity) and the false positive rate (1-specificity) at different classification thresholds. The ROC curve for the developed Attention-Based CNN-BiLSTM-Transformer Model indicates its performance in discriminating between various emotional states based on physiological signals such as EEG, ECG, GSR, and EMG. A well-performing model must possess an ROC curve that closely approaches the upper-left corner of the plot with high sensitivity and a low rate of false positives. The area under the ROC curve quantifies the model's global capacity for discrimination, where a perfect classification AUC would be 1.0 and a random guess is 0.5. Fig. 9 shows the ROC curve for emotion classification.

Attention-based fusion feature importance analysis must be conducted to determine the relative contribution of different physiological modalities towards emotional classification. In the proposed Attention-Based CNN-BiLSTM-Transformer Model, an attention mechanism dynamically attends to features from Electroencephalography, Electrocardiography, Galvanic Skin Response, and Electromyography signals. The weighting assigned is the relative contribution of each modality towards emotional state discrimination. The bar chart for attention-based fusion feature importance shows the various contributions of these physiological signals. EEG would generally play a central role in emotion recognition because it directly reflects brain activity and emotional response. ECG, which is the

representation of heart rate variability, also possesses very significant importance, as emotions greatly influence the autonomic nervous system activity. Fig. 10 shows feature importance in Attention-Based Fusion.

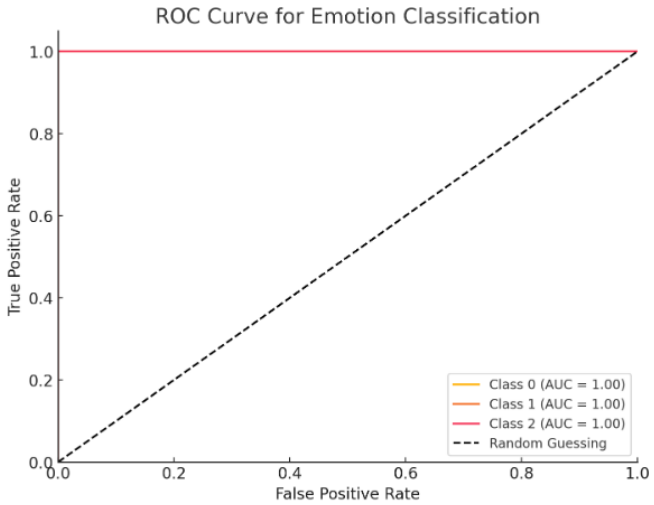

Fig. 9. ROC curve for emotion classification.

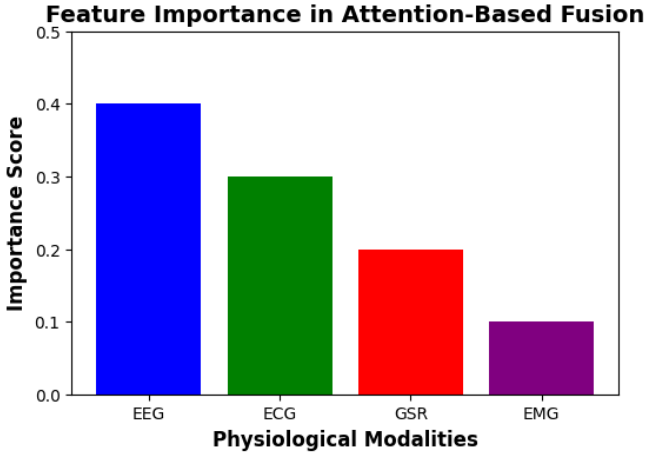

Fig. 10. Feature importance in attention-based fusion.

The attention-based fusion process greatly improves emotion classification accuracy by adaptively weighing various physiological modalities. Previous fusion processes give equal weight to all modalities, risking underemphasizing important information. Attention-based fusion, however, learns adaptive weights for every modality such that, more informative signals are more influential in classification. Using EEG, ECG, GSR, and EMG signals, the model learns to attend to the most salient physiological changes reflecting various emotional states. A comparative study between attention-based fusion and non-attention-based fusion models indicates that the incorporation of attention mechanisms greatly improves performance. In the absence of attention, the model struggles to fuse multimodal data and achieves suboptimal accuracy. Fig. 11 shows the effect of Attention-Based Fusion on accuracy.

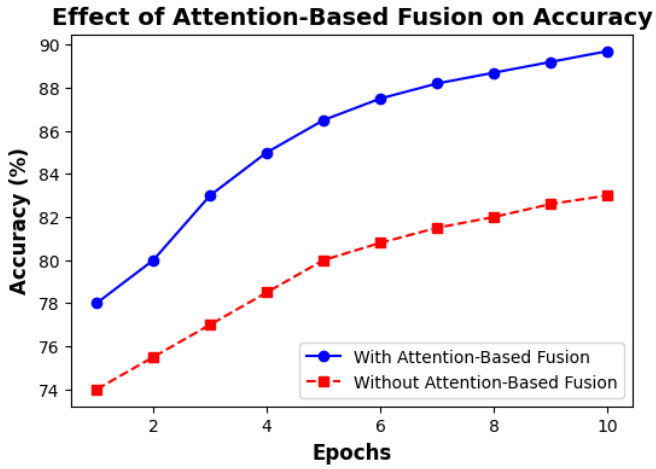

Fig. 11. Effect of attention-based fusion on accuracy.

Convergence of training loss and accuracy is an important factor in measuring the efficacy and validity of a deep learning model. Through training, the model optimizes its parameters in an iterative manner for reducing the loss function and maximizing the classification accuracy. A well-trained model displays loss reduction and accuracy improvement consistently with epochs, indicating successful learning. In the suggested Attention-Based CNN-BiLSTM-Transformer Model, the training loss is smooth and decreases as the number of epochs increases. This indicates that the model is learning useful patterns from multimodal physiological data effectively. The loss is high at the beginning because of random weight initialization, but as training continues, the model improves its feature representations, resulting in lower error rates. One important observation here is that attention-based fusion speeds up convergence by effectively combining the multiple independent physiological signals, eliminating redundancy and enhancing feature representation. Fig. 12 shows the training loss and accuracy convergence.

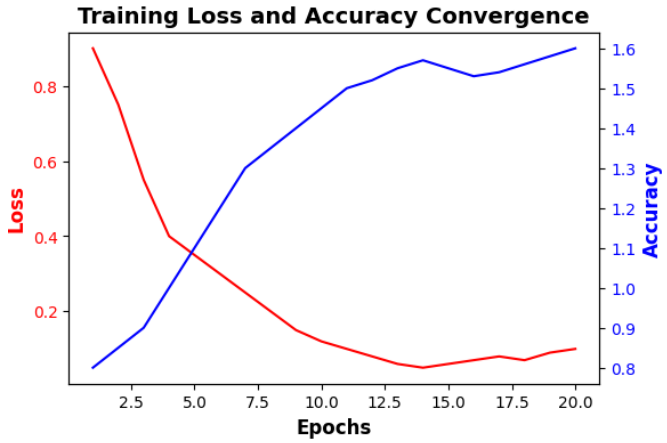

Fig. 12. Training loss and accuracy convergence.

The current performance gap on the DEAP and AMIGOS datasets is a testament to enhanced efficiency of the new proposed CNN-BiLSTM-Transformer model. SVM and Random Forest are conventional machine learning models with lower accuracy, recall, F1-score, and precision, indicating

inefficiency in spatial, temporal, and long-range dependency analysis. The Transformer-only model enhances performance, but the new model has the best accuracy of 88.2% on DEAP and 89.5% on AMIGOS. They also have increased precision, recall, and F1-score. These results suggest that combining CNN, BiLSTM, and Transformer with an attention-based fusion mechanism improves emotion recognition from multimodal physiological signals. Table V shows the performance comparison of different models on DEAP and AMIGOS datasets.

TABLE V. PERFORMANCE COMPARISON OF DIFFERENT MODELS ON DEAP AND AMIGOS DATASETS

Model	Dataset	Accuracy (%)	F1-Score	Precision	Recall
CNN-LSTM (Baseline) [49]	DEAP	81.3	0.79	0.80	0.78
CNN-LSTM (Baseline) [49]	AMIGOS	83.1	0.81	0.82	0.80
SVM [50]	DEAP	75.2	0.73	0.74	0.72
SVM 50]	AMIGOS	76.5	0.74	0.75	0.73
Random Forest [50]	DEAP	78.6	0.76	0.77	0.75
Random Forest [50]	AMIGOS	79.8	0.78	0.79	0.77
Transformer-only Model	DEAP	85.4	0.84	0.85	0.83
Transformer-only Model	AMIGOS	86.7	0.85	0.86	0.84
Proposed Model (CNN-BiLSTM-Transformer)	DEAP	88.2	0.87	0.88	0.86
Proposed Model (CNN-BiLSTM-Transformer)	AMIGOS	89.5	0.88	0.89	0.87

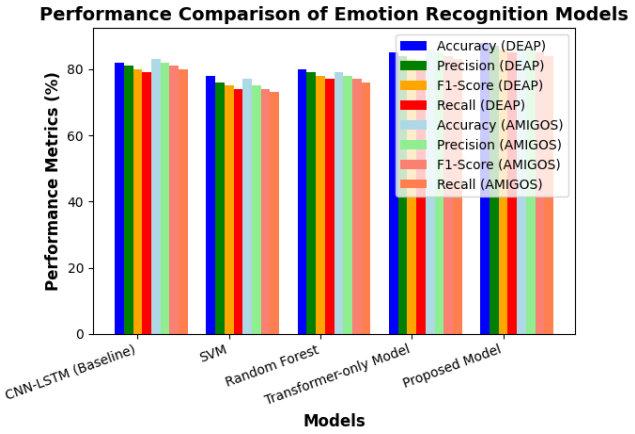

Fig. 13. Performance comparison of emotion recognition models.

Fig. 13 graphically represents a comparison of the performance of different models on the DEAP and AMIGOS datasets in terms of accuracy, F1-score, precision, and recall. Traditional machine learning models, such as SVM and Random Forest, have lower performance according to all the metrics, indicating the drawback of these models to identify complex dependencies in physiological signals. The Transformer-only model improves the classification accuracy, but the introduced CNN-BiLSTM-Transformer model achieves the highest accuracy of 88.2% and 89.5% for DEAP and AMIGOS, respectively, as well as the highest precision, recall, and F1-score. It reflects the bitterness of integrating CNN for spatial information, BiLSTM for temporal relations, and Transformer for long-distance feature extraction with attention-based fusion to improve emotion identification.

A. Discussion

Experimental results show that the envisioned CNN-BiLSTM-Transformer model is superior to conventional machine learning algorithms and baseline deep models in the recognition of emotions from physiological signals. With precision, recall, and F1-score values higher than 88.2% for DEAP and 89.5% for AMIGOS, while achieving high accuracy levels of 88.2% on DEAP and 89.5% on AMIGOS, the model successfully extracts spatial, temporal, and long-range dependencies in multimodal data. In comparison to SVM [50] and Random Forest [50], which find it hard to handle intricate physiological patterns, and the Transformer-only model, which doesn't have spatial and short-term temporal feature extraction, the new hybrid method is able to obtain better classification performance. The attention-based fusion mechanism also produces better results by dynamically weighting the most important features, leading to a more solid and accurate emotion recognition system.

VI. CONCLUSION AND FUTURE WORKS

This study introduces a new CNN-BiLSTM-Transformer model for emotion recognition with multimodal physiological signals and attains higher accuracy than conventional machine learning and state-of-the-art deep learning models. The model combines CNN for spatial features, BiLSTM for temporal patterns, and Transformer for sequence relationships, capturing heterogeneous physiological patterns under different emotional states. The attention-based fusion mechanism also improves classification performance by dynamically giving weights to informative features, resulting in 88.2% accuracy on the DEAP dataset and 89.5% accuracy on the AMIGOS dataset. These outcomes validate the efficiency of the presented approach to enhance emotion recognition, which can act as a useful tool for affective computing applications, healthcare, and human-computer interaction.

Future work will include expanding the capability of the model for real-time emotion recognition in everyday environments with a decreased dependence on laboratory setups. Another direction may include additional physiological modalities such as respiration rate and eye-tracking data to further enhance emotional state classification. Another direction is optimizing the computational cost of the model for to use in

wearables and mobiles. Finally, investigating domain adaptation techniques for increasing generalizability across different datasets and individual differences will be crucial in furthering the real-world applicability of the proposed methodology.

DECLARATIONS

Funding: There is no specific funding to support this research.

Conflict of Interest: The authors declare that they have no conflicts of interest regarding this work.

Data Availability: All data generated or analyzed during this study are included in the manuscript.

Code Availability: Not applicable.

Ethical statement: Not Applicable.

Informed consent to participate: Not Applicable.

Consent to Publish declaration: Not applicable

Clinical Trial: Not Applicable

AUTHOR'S CONTRIBUTION

Yue Pan contributed the design and methodology of this study, the assessment of the outcomes, and the writing of the manuscript.

REFERENCES

[1] Abdul-Al, M., Kyeremeh, G. K., Qahwaji, R., Ali, N. T., & Abd-Alhameed, R. A. (2024). A novel approach to enhancing multi-modal facial recognition: Integrating convolutional neural networks, principal component analysis, and sequential neural networks. *IEEE Access, 12*, 140823–140846. https://doi.org/10.1109/ACCESS.2024.3467151

[2] Shu, L., et al. (2018). A review of emotion recognition using physiological signals. *Sensors, 18*(7), 2074.

[3] Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy, 25*(10), 1440.

[4] Chen, P., et al. (2022). An improved multi-input deep convolutional neural network for automatic emotion recognition. *Frontiers in Neuroscience, 16*, 965871.

[5] Duan, J., Xiong, J., Li, Y., & Ding, W. (2024). Deep learning-based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion, 102536.*

[6] Vala, J. M., & Jaliya, U. K. (2022). Deep learning network and Renyi-entropy based fusion model for emotion recognition using multimodal signals. *International Journal of Modern Education and Computer Science, 11*(4), 67.

[7] Saha, P., et al. (2024). Novel multimodal emotion detection method using electroencephalogram and electrocardiogram signals. *Biomedical Signal Processing and Control, 92*, 106002. https://doi.org/10.1016/j.bspc.2024.106002

[8] Zhang, Y., Hossain, M. Z., & Rahman, S. (2021). DeepVANet: A deep end-to-end network for multi-modal emotion recognition. *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18*, 227–237.

[9] Zhang, Q., Zhang, H., Zhou, K., & Zhang, L. (2023). Developing a physiological signal-based, mean threshold and decision-level fusion algorithm (PMD) for emotion recognition. *Tsinghua Science and Technology, 28*(4), 673–685. https://doi.org/10.26599/TST.2022.9010038

[10] Yin, J. Z. (2023). Emotion recognition and treatment based on multi-feature signal fusion and deep learning model. *2023 3rd International Signal Processing, Communications and Engineering Management Conference (ISPCEM)*, 26–30. https://doi.org/10.1109/ISPCEM60569.2023.00011

[11] Egger, M., Ley, M., & Hanke, S. (2019). Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science, 343*, 35–55. https://doi.org/10.1016/j.entcs.2019.04.009

[12] Jaswal, R. A., & Dhingra, S. (2023). Empirical analysis of multiple modalities for emotion recognition using convolutional neural network. *Measurement and Sensing, 26*, 100716. https://doi.org/10.1016/j.measen.2023.100716

[13] Gao, J. (2024). Exploring key technologies for multimodal emotion recognition: Research and application analysis. *AIP Conference Proceedings, 3194*(1), 040019. https://doi.org/10.1063/5.0223702

[14] Gahlan, N., & Sethia, D. (2024). Federated learning-inspired privacy-sensitive emotion recognition based on multi-modal physiological sensors. *Cluster Computing, 27*(3), 3179–3201.

[15] Fang, Y., Rong, R., & Huang, J. (2021). Hierarchical fusion of visual and physiological signals for emotion recognition. *Multidimensional Systems and Signal Processing, 32*(4), 1103–1121.

[16] Xu, J., Hu, Z., Zou, J., & Bi, A. (2019). Intelligent emotion detection method based on deep learning in medical and health data. *IEEE Access, 8*, 3802–3811.

[17] Adel, O., Fathalla, K. M., & Abo ElFarag, A. (2023). MM-EMOR: Multi-modal emotion recognition of social media using concatenated deep learning networks. *Big Data and Cognitive Computing, 7*(4), 164.

[18] Radzi, N. H. M., Hashim, H., et al. (2024). Research on emotion classification based on multi-modal fusion. *Baghdad Science Journal, 21*(2), 0548–0548.

[19] Singh, N., & Kapoor, R. (2023). Multi-modal expression detection (MED): A cutting-edge review of current trends, challenges, and solutions. *Engineering Applications of Artificial Intelligence, 125*, 106661.

[20] Thiam, P., Hihn, H., Braun, D. A., Kestler, H. A., & Schwenker, F. (2021). Multi-modal pain intensity assessment based on physiological signals: A deep learning perspective. *Frontiers in Physiology, 12*, 720464.

[21] Wang, J., Zhao, W., Meng, F., & Qu, G. (2024). Sentiment analysis methods based on multi-modal fusion and deep learning. *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, 743–747.

[22] Gladys, A. A., & Vetriselvi, V. (2023). Survey on multimodal approaches to emotion recognition. *Neurocomputing, 556*, 126693.

[23] Fu, Z., Zhang, B., He, X., Li, Y., Wang, H., & Huang, J. (2022). Emotion recognition based on multi-modal physiological signals and transfer learning. *Frontiers in Neuroscience, 16.* https://doi.org/10.3389/fnins.2022.1000716

[24] Liu, D., Wang, Z., Wang, L., & Chen, L. (2021). Multi-modal fusion emotion recognition method of speech expression based on deep learning. *Frontiers in Neurorobotics, 15.* https://doi.org/10.3389/fnbot.2021.697634

[25] Ayata, D., Yaslan, Y., & Kamasak, M. E. (2020). Emotion recognition from multimodal physiological signals for emotion-aware healthcare systems. *Journal of Medical and Biological Engineering, 40*(2), 149–157. https://doi.org/10.1007/s40846-019-00505-7

[26] Ghoniem, R. M., Algarni, A. D., & Shaalan, K. (2019). Multi-modal emotion-aware system based on fusion of speech and brain information. *Information, 10*(7), 7. https://doi.org/10.3390/info10070239

[27] Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., & Chandran, V. (2020). Automatic emotion recognition using temporal multimodal deep learning. *IEEE Access, 8*, 225463–225474. https://doi.org/10.1109/ACCESS.2020.3027026

[28] Zhang, X., et al. (2021). Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine. *IEEE Transactions on Cybernetics, 51*(9), 4386–4399. https://doi.org/10.1109/TCYB.2020.2987575

[29] Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M. A. M., & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. *Journal of Applied Sciences and Technology Trends, 2*(1). https://doi.org/10.38094/jastt20291

[30] Pan, T., Ye, Y., Cai, H., Huang, S., Yang, Y., & Wang, G. (2023). Multimodal physiological signals fusion for online emotion recognition. *Proceedings of the 31st ACM International Conference on Multimedia*, 5879–5888. https://doi.org/10.1145/3581783.3612555

[31] Guo, Z., et al. (2024). E-MFNN: An emotion-multimodal fusion neural network framework for emotion recognition. *PeerJ Computer Science, 10*, e1977. https://doi.org/10.7717/peerj-cs.1977

[32] Yang, K., et al. (2023). Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Transactions on Affective Computing, 14*(2), 1082–1097. https://doi.org/10.1109/TAFFC.2021.3100868

[33] Song, T., Zheng, W., Lu, C., Zong, Y., Zhang, X., & Cui, Z. (2019). MPED: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access, 7*, 12177–12191. https://doi.org/10.1109/ACCESS.2019.2891579

[34] Dai, Z., Fei, H., & Lian, C. (2024). Multimodal information fusion method in emotion recognition in the background of artificial intelligence. *Internet Technology Letters, 7*(4), e520. https://doi.org/10.1002/itl2.520

[35] Xiang, G., et al. (2024). A multi-modal driver emotion dataset and study: Including facial expressions and synchronized physiological signals. *Engineering Applications of Artificial Intelligence, 130*, 107772. https://doi.org/10.1016/j.engappai.2023.107772

[36] Roshdy, A., Karar, A., Kork, S. A., Beyrouthy, T., & Nait-Ali, A. (2024). Advancements in EEG emotion recognition: Leveraging multi-modal database integration. *Applied Sciences, 14*(6). https://doi.org/10.3390/app14062487

[37] Liu, H., et al. (2024). EEG-based multimodal emotion recognition: A machine learning perspective. *IEEE Transactions on Instrumentation and Measurement, 73*, 1–29. https://doi.org/10.1109/TIM.2024.3369130

[38] Bota, P. J., Wang, C., Fred, A. L. N., & Plácido Da Silva, H. (2019). A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access, 7*, 140990–141020. https://doi.org/10.1109/ACCESS.2019.2944001

[39] Wu, D., Zhang, J., & Zhao, Q. (2020). Multimodal fused emotion recognition about expression-EEG interaction and collaboration using deep learning. *IEEE Access, 8*, 133180–133189. https://doi.org/10.1109/ACCESS.2020.3010311

[40] Luo, J., Tian, Y., Yu, H., Chen, Y., & Wu, M. (2022). Semi-supervised cross-subject emotion recognition based on stacked denoising autoencoder architecture using a fusion of multi-modal physiological signals. *Entropy, 24*(5). https://doi.org/10.3390/e24050577

[41] Pradhan, A., & Srivastava, S. (2024). Hybrid DenseNet with long short-term memory model for multi-modal emotion recognition from physiological signals. *Multimedia Tools and Applications, 83*(12), 35221–35251. https://doi.org/10.1007/s11042-023-16933-2

[42] Gahlan, N., & Sethia, D. (2024). AFLEMP: Attention-based federated learning for emotion recognition using multi-modal physiological data. *Biomedical Signal Processing and Control, 94*, 106353. https://doi.org/10.1016/j.bspc.2024.106353

[43] Xu, X., et al. (2024). A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis. *Bioengineering, 11*(3). https://doi.org/10.3390/bioengineering11030219

[44] Palanivel, R., Basani, D. K. R., Gudivaka, B. R., Fallah, M. H., & Hindumathy, N. (2024). A support vector machine with tunicate swarm optimization algorithm for emotion recognition in human-robot interaction. Proceedings of the 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems, 1–4. https://doi.org/10.1109/IACIS61494.2024.10721631

[45] Naresh, K. R. P. (2022). Applying discrete wavelet transform for ECG signal analysis in IoT health monitoring systems. International Journal of Information Technology & Computer Engineering, 10(4). ISSN 2347–3657.

[46] Ganesan, T., & Devarajan, M. V. (2021). Integrating IoT, fog, and cloud computing for real-time ECG monitoring and scalable healthcare systems using machine learning-driven signal processing techniques. International Journal of Information Technology and Computer Engineering, 9(1), 202–217.

[47] Hamzah, H. A., & Abdalla, K. K. (2024). EEG-based emotion recognition systems: Comprehensive study. *Heliyon, 10*(10), e31485. https://doi.org/10.1016/j.heliyon.2024.e31485

[48] Si, X., Huang, D., Sun, Y., Huang, S., Huang, H., & Ming, D. (2023). Transformer-based ensemble deep learning model for EEG-based emotion recognition. *Brain Science Advances, 9*(3), 210–223. https://doi.org/10.26599/BSA.2023.9050016

[49] DEAP: A dataset for emotion analysis using physiological and audiovisual signals. (2025). Retrieved from https://www.eecs.qmul.ac.uk/mmv/datasets/deap/index.html

[50] AMIGOSDataset — torcheeg 1.0.10 documentation. (2025). Retrieved from https://torcheeg.readthedocs.io/en/v1.0.10/generated/torcheeg.datasets.AMIGOSDataset.html