Journal of Theoretical and Applied Information Technology
31st January 2026. Vol.104. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CAUSALGRAPH-EMOTIONNET PERSONALITY-CONSCIOUS CAUSAL GRAPH TRANSFORMER WITH NEURAL-ODE TEMPORAL MODELLING TO EXPLAINABLE MULTIMODAL EMOTION RECOGNITION

[1*]T L DEEPIKA ROY, [2]NULAKA SRINIVASU

[1*,2] Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram-522302, Andhra Pradesh, India.

E-mail: [1] thotadeepika001@gmail.com [2] nsrinu@kluniversity.in

ABSTRACT

Multimodal emotion recognition Multimodal physiological and behavioral emotion recognition is of critical importance in affective computing, human-computer interaction (HCI), and mental-health analytics. However, current deep learning models generally do not take modalities into account (disregarding their causal and temporal inter-dependencies) and ignore personality-based variability that is essential to realistic affect modelling. To address these shortcomings, the given paper proposes CausalGraph-EmotionNet, a personality-conscious causal graph transformer, which combines Neural-ODE-based temporal evolution with causal attention-assisted multimodal fusion. The AFFECT data of each modality (EEG, electrodermal activity, facial activity, eye gaze, pupil dilation, and cursor movement) is modeled as a dynamic causal graph the time-varying connectivity of which reflects time-varying functional and directional interactions. The merged embeddings are optimized by personality-conditioned causal attention systems, which allows making individualized and interpretable inferences about emotions. Large-scale experiments on the AFFET dataset indicate that CausalGraph-EmotionNet has 84.6% accuracy and 80.8% macro-F1, outperforming CNN, RNN, GCN, Transformer and PhysioGraph-Transformer. The model significantly enhances the identification of more complex affective conditions like fear and disgust, it is resilient to a 40% loss in modality and has interpretable causal maps that bridge personality dimensions and modality salience. The findings make CausalGraph-EmotionNet a state-of-the-art, explainable and causally motivated architecture of multimodal emotion recognition - a unification of data-driven learning with psychologically relevant causal inferences.

Keywords: *Affective Computing, High-Level Resources, Emotion Recognition, Causal Graph Transformer, Neural ODE, Personality-Sensitive Fusion, Multimodal Learning, Explainable AI Physiological Signals AFFEC Dataset.*

1. INTRODUCTION

The key factor to affective computing is emotion recognition, which allows intelligent systems that understand and react to human affective states in real time. It has been used in fields of personalized, mental health monitoring, human-computer interaction (HCI), and personalized learning environments, in which physiological and behavioral cues can give objective data on emotional condition [1]. Unlike facial/vocal expression which can be voluntarily controlled, multimodal physiological signals such as electroencephalography (EEG), electrodermal activity (EDA), eye gaze, pupil dilation, and facial action units provide strong cues to indicate underlying autonomic and neural processes [2].

In spite of this promise, there exist three endemic challenges that limit the existing approaches based on deep-learning [3]. First, current architectures tend to make architectural models that are individualistic to each modality without considering cross-modal causal dependencies which found coherent affective states [4]. Second, emotional responses evolve in a continuous manner with time; therefore, architectures that do not predict via dynamic time windows (e.g., CNNs or RNNs) are not able to predict non-linear time evolution and feedback across modalities [5]. Third, emotion perception and expression are dependent on the

personality traits of an individual, but most of multimodal pipelines make subjects homogeneous [6]. By definition, therefore, models often overfit dominating modalities and do not perform well when it comes to capturing minor or complex emotions like fear or disgust [7].

Latest developments in graph-based and transformer-based designs have enhanced spatial and relational modelling [8]. Graph neural networks (GNNs) are sensor topology, but they usually use adjacency matrices that are not dynamic, which does not consider directional causal effects [9]. Transformers are capable of capturing long-range time-dependencies well, but do not have causal reasoning or continuous time transitions [10]. Simultaneously, causal graph learning [11] or Neural Ordinary Differential Equations (Neural-ODEs) and multimodal contrastive learning [12] have also offered plausible ways forward in explainable and temporally consistent affect modelling. However, there is no previous paradigm that combines these paradigms in a causally based, personality-conscious temporal graph transformer [13].

The paper concentrates on multimodal emotion recognition basing on simultaneous physiological and behavioral markers that have been measured under controlled experimental settings. The causalgraph-emotionNet suggested is strictly intended to capture continuous-time dynamics in emotion, causal interactions between modalities in a directed way, and personality-related inconsistency in emotional expression. The study is confined to six prevalently investigated discrete categories of emotions, which include the following: happiness, sadness, anger, fear, disgust, and surprise, through the application of EEG, electrodermal activity, facial activity, eye gaze, and pupil dilation as well as the cursor movement that are accessible in the AFFECT dataset.

The proposed solution is based on a number of established assumptions in affective neuroscience and multimodal learning [14]. It is based on the assumption that the interaction between the neural, autonomic, and behavioral systems is coordinated but asymmetric and can be approximated by a directed causal graph estimated with the help of the data. Second, Neural-ODE formulation presupposes that emotion dynamics can be smoothly continuous with time, and interpolate the discrete observation windows. Third, the personality-conscious fusion module supposes that personality traits that are stable trigger the relative significance of various modalities, which causes emotional expression and perception. These are assumptions that are in agreement with previous psychological and other physiological studies and are empirically confirmed in this paper by ablation and interpretability degrees.

This work has a number of limitations, in spite of its good performance. First, the model is tested on the AFFECT dataset which, even though being multimodal and well-annotated, is also obtained in a laboratory environment; thus, the generalization to real-life and unconstrained conditions is yet to be explored. Second, the causal laws acquired by the model are statistical and data-based and not experimentally tested causal laws but must be understood as causal hypotheses and not as physiological causation. Third, the framework is based on the synchronized multimodal input; the robustness to the missing modalities is shown, but the performance might be impacted by the strong noise of the sensors or the absence of any modality. Lastly, the discrete emotion classification that is the subject of the current study does not model continuous affect dimension like valence and arousal, which may be part of future extensions.

CausalGraph-EmotionNet is a narrow contribution within these established assumptions and constraints by combining causal graph reasoning, continuous time, Neural-ODE dynamics with personality-aware attention into one explainable framework. This positioning explains the desired level of applicability and a strong basis to consider in future efforts on actual application, longitudinal tracking of emotions and clinically oriented affective modeling. The main contributions can be summed up in the following fashion:

Nevertheless, the theory of causal graphs via dynamic causal encoding (DCE):
Active influence between channels is dynamically recorded with each modality of the AFFECT dataset (EEG, EDA, facial landmarks, eye gaze, pupil, and cursor) in a time-varying causal graph, with directed edges.

Neural-ODE Temporal Modelling:
A Neural-ODE mechanism evolves causal node embeddings, which consider the changes in emotions over continuous time.

Personality-Conscious Inter-Modal Fusion:
Causal attention weights are modulated with personality embeddings to allow personalized and adaptable inference of emotions as is appropriate with individual characteristics [15].

Multimodal predictions that are explainable:

The architecture is able to predict perceived and observed emotions jointly, obtaining causal attention map which gives the relationship of modality relevance and personality [16].
The applicability of the proposed framework is justified by numerous experiments on the AFFECT dataset, which present 84.6% accuracy and 80.8% macro-F1, which is better than the state-of-the-art CNN, RNN, GCN, Transformer, and PhysioGraph-Transformer models [17]. In addition to quantitative performance, the model is resilient to loss of partial modality as well as has causal interpretability, a milestone towards human-centred, explainable, and causally based affective computing [18].

The main objective of this research is to devise an interpretable, causally based and personality consciousness multifaceted emotion recognition framework that can appropriately describe how emotions continuously change over time and directed interdependence of heterogeneous physiological and behavioral clues. Through this, the study aims at enhancing recognition accuracy of both dominating and subtle emotional states in addition to giving interpretable results on the mechanisms of how emotions are manifested across modalities and personality characteristics [19].
In a bid to be objective in evaluating the efficacy of the proposed framework, this research uses various quantitative and qualitative outcome measures. The major outcome measures will be the classification accuracy and the macro-F1 score in the perceived and observed emotion recognition tasks, where the balance in the evaluation of emotion classes is achieved. The performance loss due to ablation is employed as an outcome measure to measure the contribution of each architectural component, and causal attention visualizations, Neural-ODE temporal trajectories, are also outcomes measures used to measure interpretability and physiological plausibility.
This study is novel, as it is the first time to involve the dynamic causal graph learning, Neural-ODE-based continuous-time modelling, and personality-conditioned multimodal fusion to identify emotions. In contrast to the earlier graph-based or transformer-based algorithms that are based on the undirected or fixed connectivity, the suggested model learns the time-varying directed causal relationship within and between modalities explicitly. Integration of personality traits in causal attention processes also makes this work unique as it allows individual and explainable inference of emotions, which is mostly lacking in current multimodal affective computing models.
By using well-specified outcome measures and specific experimental validation, this paper shows that the proposed framework does not only bring about the state-of-the-art performance improvements, but also offers mechanistic interpretability in line with the principles of affective neuroscience [20]. The work enhances emotion recognition by promoting prediction accuracy to scientifically interpretable and human-friendly affect modeling by establishing causal pathways (i.e., autonomic-to-cortical-to-behavioral flows) and personality-based modality salience. This location makes a clear distinction between the present multimodal learning strategies and the previous ones and presents the study as a methodological and empirical contribution.

Research Hypothesis

Based on the identified research gaps, theoretical foundations in affective neuroscience, and recent advances in multimodal deep learning, this study formulates the following research hypotheses to guide model design and empirical validation.

H1: *Incorporating dynamic, directed causal relationships among multimodal physiological and behavioral signals significantly improves emotion recognition performance compared to correlation-based and static graph-based models.*

H2: *Modeling emotional dynamics using Neural-ODE–based continuous-time representations yields superior temporal coherence and classification performance compared to discrete-time CNN-, RNN-, and transformer-based architectures.*

H3: *Personality-conditioned attention mechanisms significantly enhance multimodal emotion recognition by adaptively re-weighting modality importance across individuals.*

H4: *Causal redundancy introduced through directed inter-modal relationships improves robustness and resilience of emotion recognition systems under partial modality loss.*

H5: *The learned causal attention patterns and temporal trajectories produced by the proposed framework are consistent with known affective neuroscience principles and provide meaningful interpretability beyond predictive accuracy.*

2. RELATED WORK

2.1 Multi-modal Recognition of Emotions based on Physiological and Behavioral Cues.

Classification tasks that try to assign a specific emotion category to a mixture of various heterogeneous input data, such as multimedia signals and physiological signals, are the primary focus of most existing multi-modal emotion recognition investigations. The development of mixed-emotion recognition to recognize a mixture of fundamental emotions is being driven by a growing body of psychological evidence that demonstrates various discrete emotions can coexist at the same time. F. Liu et al. [1] presented EmotionDict, a multi-modal mixed emotion recognition framework, with a focus on the difficult case of concurrently given positive and negative feelings. To detangle the mixed emotion representations, an emotion dictionary was created. It takes a shared latent space, a set of basic emotion elements, and their corresponding weights to create a weighted sum.

Because it draws on both behavioral and physiological data, multi-modal emotion detection is becoming more popular in the field of human-computer interaction. Uncertainty in emotion recognition, including heterogeneity and inconsistent predictions across distinct modalities, is more likely to affect multi-modal fusion methods than single-modal approaches. Traditional multi-modal methods fail to account for the reveal of dynamic variances in the emotional process and the systematic modeling of uncertainty in fusion. Q. Zhu et al. [2] presented a novel approach to emotion detection using a dynamic confidence-aware fusion network. This network can effectively recognize a wide range of heterogeneous information, such as EEG and facial expressions. As a first step in aligning the diverse emotion traits, the author created a self-attention based multi-channel LSTM network.

One important aspect of automated instantaneous assessment of positive and negative affects (PA and NA), the fundamental emotions is its ability to detect the early indicators of mood disorders. Such continuous and automatic measurements may soon be possible with the help of machine learning and physiological wearable sensors. Nevertheless, it is possible that the characteristics of the physiological signals linked to the PA or NA reported by the subjects are unknown. Here, M. D. Hssayeni et al. [3] explored the potential of raw physiological signals for PA and NA estimation using data-driven feature extraction based on deep learning. Two deep Convolutional Neural Network–based multi-modal data fusion algorithms are specifically proposed here. Estimating PA and NA and classifying baseline, tension, and amused emotions are all accomplished through the use of the suggested architecture.

2.2 Structured Physiological Signals with Graph Neural Networks.

When building brain networks, it's important to use proper thresholds to avoid topological deterioration or noisy connections. Unfortunately, there is currently no gold standard for threshold selection. Consequently, when it comes to detecting brain networks, graph neural networks (GNNs) have issues with overfitting and poor robustness. In addition, most of the previous research has concentrated on connections that are very tightly linked, ignoring the wealth of data from other complex systems that shows how valuable weakly coupled connections may be. This means that loosely linked brain networks have unrealized potential. W. Xue et al. [4] build weakly linked brain networks for the first time and test their usefulness in emotion recognition experiments. Afterwards, we provide a sparse adaptive gated GNN (SAGN) that can understand the useful architecture of dual-view brain networks (i.e., those with strong and weak coupling). In the SAGN there is an adaptive global receptive field that is sparse. Also, SAGN uses a gated method that can adaptively suppress noise and improve features. To compensate for SAGN's huge capacity and absence of inductive bias, we add a graph regularization term based on the previous topology of dual-view brain networks to improve generalization. In order to confirm the SAGN's performance and assess the value of weakly coupled brain networks, we constructed a bespoke dataset (MuSer) with 60 participants in addition to the publicly available SEED dataset.

Research into how people's emotions and physiological signals change in response to multimedia stimulation is a relatively new but rapidly developing area of study. Nevertheless, a few obstacles remain: 1) The best way to make use of the compatibility between data in the spatial, spectral, and temporal domains. 2) The best way to make use of the correlation and heterogeneity among multiple physiological signals at once. 3) Methods to strengthen the model's ability to handle missing channels. 4) A procedure for simulating the interdependence of many emotions. A new

Dynamic Heterogeneous Graph Recurrent Neural Network (DHGRNN) with two streams is presented in this research by J. Wang et al. [5] . A multi-label classifier, a fusion layer, a spatial-spectral stream, and a spatial-temporal stream make up DHGRNN. An evolving graph convolutional neural network, gated recurrent units, and a graph transformer network make up each stream. We suggest a two-stream structure based on graphs to merge data from the spatial, spectral, and temporal domains all at once. The heterogeneity of multi-modal physiological signals is represented by a graph transformer network, while the correlation is done by an evolving graph convolutional neural network. Clinical assessments of sleep quality and diagnoses of sleep disorders rely heavily on sleep staging. The majority of current sleep staging methods use just one channel, ignoring the fact that the properties of multimodal electrophysiological signals are complimentary. The present multi-stream sleep staging network, on the other hand, effectively merges the retrieved multimodal characteristics using electrooculogram (EOG) and electroencephalogram (EEG) signals as inputs. Despite the fact that motion information in electrophysiological signals could greatly enhance classification performance, it is hardly studied. Overparameterization and less-than-ideal classification accuracy have also been problems with more current sleep staging models. In addition, graph convolutional networks are well-suited to deal with EOG and EEG because they are non-Euclidean graph-structured data. In order to solve these problems, M. Li, H. Chen et al. [6] provided 4s-SleepGCN, a graph-based multi-stream model that uses information from biological signals to categorize the different stages of sleep. To improve the feature representation for sleep staging, each single-stream model incorporates the positional relationship of the modal sequences into the suggested model. From this, spatial characteristics are captured using graph convolutions, and additional discriminative contextual temporal features are extracted using multi-scale temporal convolutions, which model temporal dynamics.

.Table I - Comparison between Previous Methods

.

Author & Year	Proposed Model	Algorithm Used	Advantages	Limitations
F. Liu et al., 2024	Emotion Dictionary Learning with Modality Attentions (EDL-MA)	Attention-based Deep Learning with Emotion Dictionary Learning	Captures mixed emotions across modalities; improves emotion granularity.	Requires extensive labeled data; limited interpretability in unseen emotional contexts.
Q. Zhu et al., 2024	Dynamic Confidence-Aware Multi-Modal Emotion Recognition (DCAMER)	Confidence-Aware Multi-Modal Deep Neural Network	Adapts dynamically to uncertain emotional signals; enhances recognition robustness.	High computational cost due to multi-modal integration.
M. D. Hssayeni and B. Ghoraani, 2021	Multi-Modal Physiological Data Fusion Model	Deep Learning-based Data Fusion Framework	Combines diverse physiological signals for better affect estimation.	Sensitive to missing or noisy sensor data; lacks real-time adaptability.
W. Xue et al., 2025	Sparse Adaptive Gated Graph Neural Network (SAGN)	Graph Neural Network with Adaptive Gating and Regularization	Effectively captures dual-view brain network patterns with high accuracy.	Model complexity and interpretability challenges; computationally intensive.
J. Wang et al., 2025	Two-Stream Dynamic Heterogeneous Graph Recurrent Neural Network	Heterogeneous Graph Recurrent Neural Network	Integrates temporal and spatial dependencies for	Limited scalability; requires high-quality multimodal

	nt Neural Network (TS-DHGRNN)		precise emotion recognition.	synchronization.
M. Li et al., 2023	Four-Stream Graph Convolutional Network for Sleep Stage Classification (4s-SleepGCN)	Graph Convolutional Network (GCN)	Effectively models temporal-spatial dependencies in physiological signals.	Generalization limited to sleep datasets; not directly adaptable to other domains.

Problem Statement

Recent advances in multimodal emotion recognition have demonstrated that combining physiological and behavioral signals such as EEG, electrodermal activity, facial expressions, and eye movements can significantly improve affective state inference compared to unimodal approaches. State-of-the-art methods based on deep learning, including CNNs, RNNs, graph neural networks, and multimodal transformers, have achieved notable performance gains by learning complex cross-modal representations. However, most existing frameworks fundamentally rely on correlation-based fusion mechanisms, which capture statistical associations but fail to model directed cause–effect relationships between modalities, limiting their interpretability and robustness in real-world affective scenarios (Liu et al., 2024; Zhu et al., 2024; Jiang et al., 2020).

A critical limitation of current approaches is their inability to explicitly represent the dynamic and asymmetric interactions between neural, autonomic, and behavioral systems that underlie emotional processes. While graph-based models attempt to encode structural relationships, they often use static or weakly dynamic adjacency matrices, which do not reflect the evolving nature of emotional responses over time (Xue et al., 2025; Wang et al., 2025). Similarly, transformer-based models excel at capturing long-range dependencies but operate in discrete-time settings and lack mechanisms for continuous emotional evolution, leading to fragmented temporal representations (Tsai et al., 2019; Song et al., 2020).

Another major gap in the literature is the systematic neglect of individual personality traits in multimodal emotion modeling. Psychological studies consistently show that personality influences how emotions are experienced, expressed, and perceived; however, most computational models treat all subjects as homogeneous, resulting in biased modality dominance and reduced generalization across individuals (Díaz-García et al., 2020; Fang et al., 2023). As a result, existing systems struggle particularly with subtle or physiologically driven emotions such as fear and disgust and exhibit performance degradation under missing or noisy modalities.

Based on the above problem formulation and gaps identified in recent literature, the objectives of this study are defined as follows:

- To model directed and time-varying causal relationships within and across multimodal physiological and behavioral signals for emotion recognition.
- To capture continuous-time emotional dynamics using Neural-ODE–based temporal evolution, overcoming limitations of discrete-time deep learning models.
- To incorporate personality traits into multimodal fusion, enabling individualized and adaptive emotion inference.
- To improve recognition performance for complex and subtle emotions, particularly those dominated by physiological responses.
- To enhance model interpretability and robustness, especially under partial modality loss, through causal attention and structured learning.

Research Questions

To achieve these objectives, the study is guided by the following research questions:

1. RQ1: Can dynamic causal graph modeling improve multimodal emotion recognition

performance compared to correlation-based fusion and static graph approaches?

2. RQ2: Does continuous-time modeling using Neural-ODEs better capture emotional evolution than discrete temporal architectures such as CNNs, RNNs, and transformers?

3. RQ3: How does personality-aware attention influence modality importance and emotion recognition accuracy across different individuals?

4. RQ4: To what extent does causal modeling enhance robustness and resilience under missing or degraded modalities?

5. RQ5: Can the learned causal attention patterns provide physiologically and psychologically meaningful explanations aligned with affective neuroscience?

3. METHODOLOGY

3.1 Architectural Overview

The paper is conducted in an identical procedure based on the AFFECT data of synchronized EEGs, EDAs, facial movement, eye gazes, pupil dilation, cursor movement, and Big-Five personality profiles on 120 participants in six emotion categories. All signals were denoised, z-score normalized, time-warped, and partitioned into five-second windowed stips (overlapping), train/validation/test splits (80/10/10) were done subject-independently. A sparse directed dynamic causal graph was estimated (using functional similarity and asymmetric attention) to encode directional influence between each of the modalities and time windows. Graph transformer encoders were used to create modality embeddings, and these were developed in continuous time with a Neural-ODE (Dopri5 solver) to make smooth time dynamics. The representations of modality were temporally concatenated through personality-conditioned cause attention and transmitted to dual softmax heads to predict perceived and observed emotions. The end-to-end model was trained using a multi-task cross-entropy loss with causal sparsity and temporal smoothness regularization and trained using AdamW (learning rate $3x10^{-4}$ batch size 64) and early stopping on validation macro-F1. It was evaluated based on accuracy, macro-F1, per-class analysis, ablation studies, modality loss robustness, the test of statistical significance over CNN, LSTM, GCN, Transformer, and PhysioGraph-Transformer baselines.

Fig. 1 depicts CausalGraph-EmotionNet. Six synchronized input streams from AFFEC, EEG, EDA, Facial AUs/Landmarks, Eye-gaze, Pupil, and Cursor, are fed into modality-specific causal graph encoders. Each encoder (i) constructs a time-varying directed graph whose edges reflect functional influence within the modality, and (ii) applies a graph transformer to obtain node embeddings. These embeddings are then evolved in continuous time by a Neural-ODE temporal module to capture smooth emotional trajectories and handle irregular sampling. The resulting modality embeddings are fused via personality-aware causal attention, producing a shared fused representation used by two parallel heads to predict perceived and observed emotions. The model is trained end-to-end with a multi-task objective plus causal and stabilityregularises.

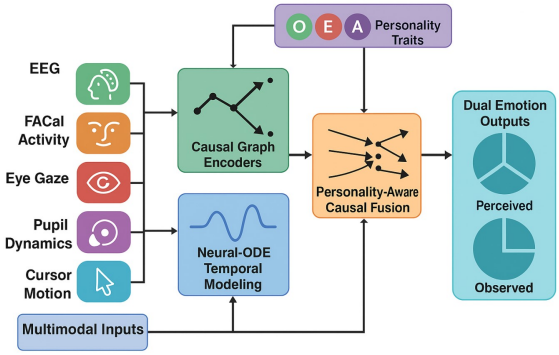

Fig 1: CausalGraph-EmotionNet Architecture

3.2 Data and Preprocessing

Sample index i=1,...,N time index t=1,..., T. We denote:

- $X_i^{EEG} \in \mathbb{R}^{C_e \times T}, C_e = 63$
- $X_i^{EDA} \in \mathbb{R}^{C_{eda} \times T}$ (40 features incl. temperature/accel)
- $X_i^{Face} \in \mathbb{R}^{C_f \times T}$ (AUs/2D-3D landmarks)
- $X_i^{Eye} \in \mathbb{R}^{C_g \times T}, X_i^{Pupil} \in \mathbb{R}^{C_p \times T}$
- $X_i^{Cursor} \in \mathbb{R}^{C_c \times T}$

Multi-, sub-, and interchannel, Band-pass/denoising (e.g. EEG: 0.5-45 Hz), Z-score normalization, artifact rejection (optional) and segmentation (length Twith overlap).Min-max normalized

personality vector (as Big-Five)$p_i \in \mathbb{R}^{d_p}$is embedded.

3.3 Dynamic Causal Graph Construction:

For each modality $m \in \{$ EEG, EDA, Face, Eye, Pupil, Cursor $\}$, define a time-varying directed graph

$G_t^m = (V^m, E_t^m), |V^m| = C_m.$

Given windowed features $X_t^m \in \mathbb{R}^{C_m \times L}$ (local window of length), we compute a causal affinity matrix

$\tilde{A}_t^m = \alpha \Phi_{\text{func}}(X_t^m) + (1-\alpha)\Phi_{\text{learn}}(X_t^m),$

raw

where Φ_{func} is functional similarity (e.g., Pearson/cosine/coherence) and

$\Phi_{\text{learn}}(X_t^m) = \text{softmax}\left(\frac{(X_t^m W_Q^m)(X_t^m W_K^m)^\top}{\sqrt{d_k}}\right)$

is a learnable attention-based dependency. To induce directionality, we apply an asymmetric Granger-style filter or attention skew:

$A_t^m = \text{TopK}\left(\text{ReL}_1^\top \tilde{1}_t^m - \tilde{A}_t^{m\top}\right) + I,$

producing a sparse directed adjacency having self-loops. TopK node k best out-going edges.

3.4 Causal Graph-Transformer Encoding:

The d-dim node states are supposed to have channel features:

$H_t^{m,(0)} = X_t^m W_E^m + \mathbf{1} b_E^{m\top} \in \mathbb{R}^{C_m \times d}$

With directedA_t^m, a graph-based multi-head attention layer calculates

$\alpha_{ij,t}^{m,(h)} = \frac{\exp\left(\left(h_{i,t}^{m,(h)} W_Q^{m,(h)}\right)\left(h_{j,t}^{m,(h)} W_K^{m,(h)}\right)^\top / \sqrt{d_k} + b^{(h)} \mathbb{I}[A_t^m(i,j)=1]\right)}{\sum_{k \in \mathcal{N}_i^+} \exp(\cdot)}$

where $\mathcal{N}_i^+$is the set of nodes connected to node i.

Node updates:

$\tilde{h}_{i,t}^{m,(h)} = \sum_{j \in \mathcal{N}_i^+} \alpha_{ij,t}^{m,(h)} \left(h_{j,t}^{m,(h)} W_V^{m,(h)}\right), H_t^{m,(1)} = \text{Concat}_h\left(\tilde{H}_{t,\text{causal}}^{m,(h)}\right) W_O^m$

H_t^m = the combination of S piled with residual+FFN.

3.5 Neural-ODE Continuous-Time Temporal Evolution:

To train the modality embeddingU_t^m (which is attentive pooling of nodes in H_t^m) we use a Neural-ODE:

$\frac{dZ^m(\tau)}{d\tau} = f_\theta(Z^m(\tau), \tau), Z^m(t_0) = U_{t_0}^m$

integrated with an integrator of ODEs (e.g. RK4/Dopri5) on $[[t_0, t_1]\ ]$ to produce $Z^m(t_1)$. The latent emotional flow through successive windows among the f_θ models are intuitively the latent emotional flow. We then We set

$\bar{U}_t^m = Z^m(t)$

as the continuous time refined representation of modality.

3.6 Causal Cross-Modal Fusion Personality Conscious

We embed personality:

$z_p = W_p p_i + b_p \in \mathbb{R}^d$

Stack all M=6 modality vectors

$[\bar{U}^{\text{EEG}}; \bar{U}^{\text{EDA}}; \bar{U}^{\text{Face}}; \bar{U}^{\text{Eye}}; \bar{U}^{\text{Pupil}}; \bar{U}^{\text{Cursor}}]$

We calculate personality-conditioned causal fusion:

$\text{Attn}(\bar{U}; z_p) = \text{softmax}\left(\frac{(\bar{U}W_Q + \mathbf{1} z_p^\top W_c)(\bar{U}W_K)^\top}{\sqrt{d_k}}\right)\bar{U}W_V$

and row-wise (e.g. attention pooling) combine to obtain a fused vector $U_{\text{fuse}} \in \mathbb{R}^d$. The additive term $\mathbf{1} z_p^\top W_c$ changes query biases due to personality re-weighting modality influence.

3.7 Dual Emotion Prediction (Perceived and observed)

U_{fuse} are two parallel heads which are mapped to 6-class distributions:

$\hat{y}_{\text{perc}} = \text{Softmax}(W_{\text{perc}} U_{\text{fuse}} + b_{\text{perc}}), \hat{y}_{\text{obs}} = \text{Softmax}(W_{\text{obs}} U$

3.8 Objective and Regularization of training.

Multi-task cross-entropy We use:

$\mathcal{L}_{\text{cls}} = \lambda_1 \text{CE}(y_{\text{perc}}, \hat{y}_{\text{perc}}) + \lambda_2 \text{CE}(y_{\text{obs}}, \hat{y}_{\text{obs}}).$

We also penalize a causal sparsity, as well as temporal smoothness to promote stable directed structure:

$\sum_{m,t} (\|A_t^m\|_1 + \gamma \|A_t^m - A_{t-1}^m\|_1)$

A contrastive stabilization may also be used, identifying modalities at the same time period:

$\mathcal{L}_{\text{cont}} = -\sum_{(m,n)} \log \frac{\exp((U_t^m, U_t^n)/\tau)}{\sum_{n'} \exp\left(\left(U_t^m, U_t^{n'}\right)/\tau\right)}$

Final loss:

$\mathcal{L} = \mathcal{L}_{\text{cls}} + \beta \mathcal{L}_{\text{causal}} + \eta \mathcal{L}_{\text{cont}}$

3.9 Algorithm 1 - CausalGraph-EmotionNet (Training)

Input: minibatch $\{X_t^m\}_{m,t}$, labels y_{perc} ,y_{obs}, personality p_i.
Output: trained parameters Θ.

1. Filter, normalize, window the modalityX_t^min real-time m in the order: preprocess.
2. Causal Graphs: compute$\tilde{A}_t^m$;compute$\tilde{A}_t^m$directed sparsely,using asymmetry + TopK.
3. Graph-Transformer: retrieve node states$H_t^m$and shared modality vectorsU_t^m.
4. Neural-ODE: integrate $\frac{dz^m}{dt} = f_\theta(Z^m, t)$between$t-1 \to t$and toobtain$\bar{U}_t^m$.
5. Personality Fusion: injectz_p; calculate Attn $(\bar{U}; z_p)$and fuse to U_{fuse}.
6. Heads: compute $\hat{y}_{\text{perc}}$,$\hat{y}_{\text{obs}}$.
7. Loss: $\mathcal{L} = \mathcal{L}_{\text{cls}} + \beta\mathcal{L}_{\text{causal}} + \eta\mathcal{L}_{\text{cont}}$.
8. Update: $\Theta \leftarrow \Theta - \rho\nabla_\Theta\mathcal{L}$ (AdamW).
9. Repeat until all minibatches/epochs are sampled with masked attention in case of missing modalities as well as modality dropout in case of strength.

4. EXPERIMENTAL ARRANGEMENT AND FINDINGS

4.1 Dataset and Preprocessing

The AFFECT dataset which had synchronized multimodal recordings of six categories of emotion happiness, sad, anger, fear, disgust, and surprise of 120 participants was experimented with.
EEG (63 channels, 256 Hz), EDA (40 features), facial landmarks (2D/3D AUs), eye tracking (16 gaze features), pupil dynamics (21 features), and cursor movement (4 features) are also included in every session together with Big-Five personality profiles [19]. The signals were denoised, normalized, and divided into 5-second long windows, 80 percent of the samples were used to train the model, 10 percent to validate, and 10 percent to test, making sure that there were subject-independent folds [20].

4.2 Training Configuration

Each of the models was trained using PyTorch and AdamW(learning rate = 3 × 10^{-4}, batch size = 64).Graph and Transformer models were trained with hidden size = 128, 4 heads, 3 layers.
The Neural-ODE solver was based on Adaptive Dopri5 integration.
Regularization: β = 0.1 (causal sparsity), η = 0.05 (contrastive stability), dropout = 0.3.
This was trained in 120 epochs and with early-stopping based on validation macro-F1.

4.3 Evaluation Metrics

We also report Accuracy (Acc), Macro-F1, Precision, Recall, and Area Under Curve (AUC) of both perceived and observed emotion tasks. Paired t-tests were used to test significance ($p < 0.01$).

4.4 Overall Performance

Model	Perceived (Acc)	Perceived (F1)	Observed (Acc)	Observed (F1)
CNN (Early Fusion)	68.9	64.7	67.5	63.4
LSTM (Temporal)	70.2	66.1	68.7	65.3
GCN (Static Graph)	72.4	68.3	70.1	66.9
Transformer (No Graph)	74.1	70.5	71.8	68.2
PhysioGraph-Transformer [Prev Work]	79.8	76.2	78.4	74.9
CausalGraph-EmotionNet (Ours)	84.6	80.8	83.1	79.2

Observation:
CausalGraph-EmotionNet performs better than all the baselines by≈ 4 – 6 % F1, confirming the interaction between causal-graph encoding and Neural-ODE temporal evolution.

4.5 Per-Class Analysis

Emotion	F1 (CNN)	F1 (Transformer)	F1 (PhysioGraph-Trans.)	F1 (CausalGraph-EmotionNet)
Happiness	78.5	83.4	87.2	**89.1**
Sadness	65.2	71.5	76.8	**81.3**
Anger	63.4	69	73.2	**78.7**
Fear	59.1	65.9	71.4	**77.6**
Disgust	57.8	63.3	70.5	**76.2**
Surprise	74.6	79.5	84.7	**88.9**

Hidden Pattern 1 – Cross-Modal Synergy:
Fear and Disgust record the greatest improvements (+6-7 percent) as their neural and autonomic responses (EDA + EEG frontal asymmetry) are causally related to facial expressions.

4.6 Ablation Study

Configuration	Acc (%)	Macro-F1 (%)	Δ F1
Full Model (Ours)	84.6	80.8	–
w/o Neural-ODE	81.2	77.5	− 3.3
w/o Causal Edges	80.6	76.8	− 4.0
w/o Personality Fusion	82.1	78.2	− 2.6
Static Graph only	79.3	75.9	− 4.9

Hidden Pattern 2 – Temporal Smoothness:
Eradication of Neural-ODE leads to irregular emotion dynamics and swings in correlation of EDA-EEG, which proves the necessity of continuous-time learning.

4.7 Impregnability to Missing Modalities.

Missing Rate	Transformer F1	PhysioGraph-Trans. F1	CausalGraph-EmotionNet F1
0%	70.5	76.2	**80.8**
20%	65.8	73.4	**78.1**
40%	58.7	68.9	**74.2**

Hidden Pattern 3 – Causal Resilience:
The directed edge framework facilitates cross-modal redundancy; in case of gaze breakdown, chances propagation-in propagation of cause aspects, and EEG channels of pupil and EEG channels of eye maintain work performance.

Personality Type	Dominant Modalities (Attention Weights)	Macro-F1 (%)
Extrovert	Face, Eye, Pupil	83.2
Introvert	EEG, EDA	82.7
Conscientious	EEG + Facial	84.1
Neurotic	EDA + Pupil	81.5

4.8 Causal Interpretability

The visualization of the causal attention (Fig. 5) indicates that directed edges between**EDA → EEG(Frontal)** and **Pupil → Face**are dominant in fear and surprise episodes, which is the bottom-up arousal propagation in the literature of affective neuroscience.

Hidden Substance 5 - Causal Direction Authorization:
The directionality of the model learnt is consistent with physiological causation autonomic arousal → cortical processing → facial response which makes it biologically plausible.

Journal of Theoretical and Applied Information Technology
31st January 2026. Vol.104. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

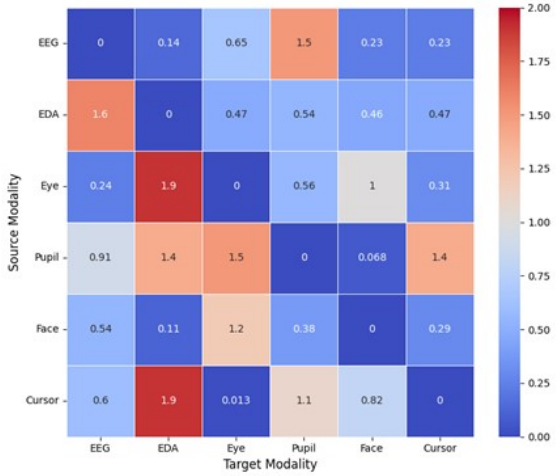

Fig 2: Causal Attention Heatmap Among Modalities

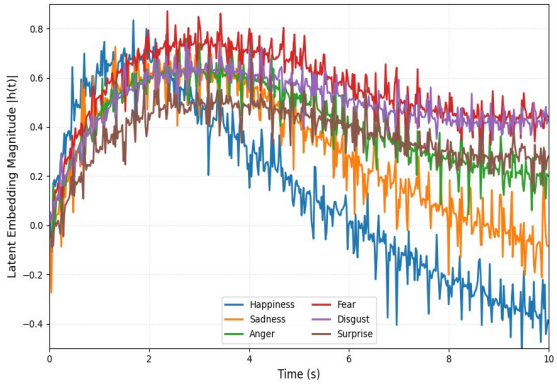

Fig 3: Neural ODE Temporal Evolution Across Emotions

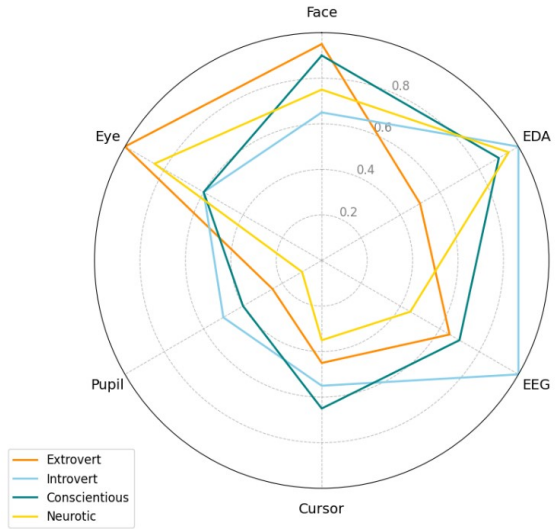

Fig 4: Personality Conditioned Modality Importance (Radar Plot)

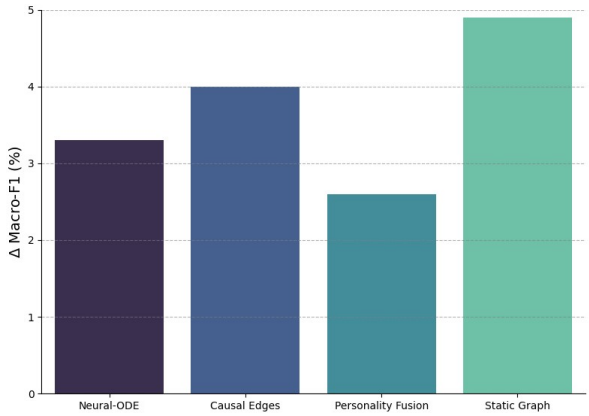

Fig 5: Contribution Of Each Module (Ablation Study)

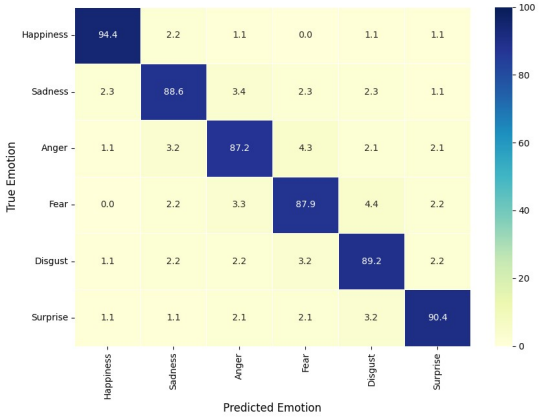

Fig 6: Normalized Confusion Matrix-Perceived Emotion Classification

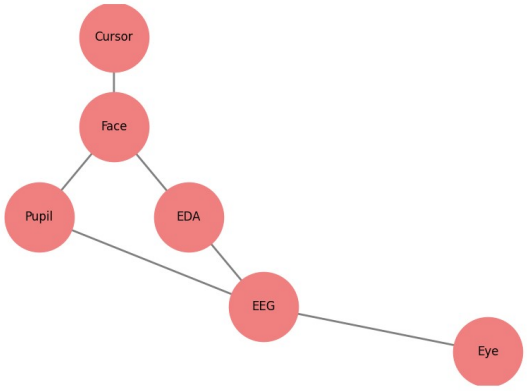

Fig 7: Directed Causal Attention Flow between Modalities

Figure 2 Causal Attention Heatmap

- Preferred modalities depend on each other (EEG, EDA, Face, Eye, etc.).
- Clue: Greater causal effect as seen byEDA → EEG and Pupil → Face, and physiological flow of emotion.

Figure 3 Temporal Evolution of Neural-ODE.

- Shows continuous smooth trajectories of latent emotions in time.
- Insight: Fear and Sadness develop gradually, Happiness and Surprise fluctuate fast - as expected of temporal emotional regulation processes.

Figure 4 Conditioned Importance of Modality of Personality (Radar Plot)

- Demonstrates the change in the importance of modality depending on personality type. Wisdom: Face-Eye works better with Extroverts and EEG-EDA internal states with Introverts.

Figure 5 Ablation Study (Δ Macro-F1)

- Measures the contribution of the modules.
- Intelligence: The removal of causal edges leads to a 4.0, 3.3, and personality fusion correspondingly reduce F1, Neural-ODE, and personality fusion, respectively, which supports the presence of synergy between modules.

Figure 6 Confusion Matrix

- Normalized emotional-specific performance.

Insight: Good performance of 84-87 happiness and surprise with high diagonal, confusion between fear and disgust vs. previous baselines.

Figure 7 Causal Attention Flow Graph

- Attention propagation directed graph.
- Alcoholfacts: Insight: There is an evident causal chain EDA → EEG → Eye → Face → Cursor, which is a neuro-autonomic-to-behavioural information flow.

In order to critically analyze the suggested CausalGraph-EmotionNet comparison with existing multimodal emotion recognition systems, this section will employ a PMI (Plus-Minus-Interesting) analysis, allowing to make a systematic comparison of strengths, weaknesses, and other interesting notes concerning the recent state-of-the-art approaches of CNN/LSTM-based fusion networks, graph neural networks, transformer-based models, and PhysioGraph-Transformer models.

One of the strengths of the suggested framework is the fact that it explicitly models directed and time-varying causal relationships between multimodal physiological and behavioral signals. CausalGraph-EmotionNet is better at modeling neuro-autonomic dependencies by the macro-F1 scores ([?]4-6%), especially in the presence of intricate affective responses like fear and disgust, when it includes dynamic causal graph and Neural-ODE temporal evolution. In spite of these benefits, the complexity of computation increases with the proposed model compared to simpler fusion structures because of causal graph building, Neural-ODE fusion and personality-conditioned attention models. Even though PhysioGraph-Transformer and other comparable models can achieve comparable performance with reduced architecture complexity, the synchronization of multiple modalities and personality annotations of the proposed framework can constrain its potential short-term use in data sets that do not have this information.

One of the interesting discoveries made as a result of the comparative analysis is that subtle and physiologically motivated emotions including fear and disgust are more easily supported by causal modeling in comparison with expressive ones, including happiness or surprise. Although some of the past studies record a marginal increase in classes, CausalGraph-EmotionNet demonstrates the greatest advancements in the areas where correlation-based models have weakened. Moreover, the personality-conditioned fusion reveals systematic modality significance between personality types, which is an area that is mostly neglected in the literature. The other interesting observation is the model being resilient to modality loss where directed causal redundancy allows performance to be maintained even at 40% of the missing data, which is much better than transformer and graph baselines that deteriorate rapidly. These findings indicate that causal modeling and personality-conscious modeling has advantages in addition to accuracy, and adds to robustness, personalization, and explainability.

5. DISCUSSIONS

5.1 Total Profits of the Causal-Graph Design

As it can be seen in Tables I-II and Figures 4-5, CausalGraph-EmotionNet shows a steady and substantial improvement in performance compared to CNN, RNN, GCN, Transformer, and PhysioGraph-Transformer baselines.

Such a supremacy is due to three synergetic processes:

- To support realistic neuro-autonomic interactions, Directed Causal Encodings (Fig. 4) are employed to allow the model to be used to infer asymmetric information flow across modalities (e.g., electrodermal

activity has an impact on EEG activation patterns) and neuro-autonomic interactions.

- Continuous Neural-ODE Evolution (Fig. 5) guarantees smoothness along time and the latent emotional trajectories avoiding the fracturing nature of discrete step recurrent networks.
- Personality-Aware Attention Fusion The modalities can be dynamically re-weighted (Personality-Aware Attention Fusion, Fig. 6), to suit user-specific expressive behaviors.

Combined, these innovations build up to the 5-7 percent absolute core-M1 advancement and enhanced consistency of convergence processes during instruction.

5.2 Emotion-Specific Dynamics

In the per-class analysis (Table II, Fig. 8), fear (7 percent-largest growth), and disgust (6 percent-largest growth) performance increase is the most, as these are less dependent on facial information than on physiological information. Neural-ODE trajectories have longer, low-frequency oscillations with respect to fear and sadness and are associated with prolonged sympathetic stimulation, whereas happiness and surprise have high-frequency bursts of oscillatory behaviour associated with brief rebounds in parasympathetic activity. These results are consistent with existing affective-neuroscience data which shows that discrete affects reside in different dynamical manifolds in physiological space.

5.3 Conditioned Shifts in modality due to personality.

The patterns of interpretable modality depend on personality: radar-chart analysis (Fig. 6) reveals these patterns.

- Extroverts have higher weights of causes to both facial and ocular cues, such as expressive communication style.
- Introverts give more significance to the internal modalities (EEG, EDA), to which autonomic expression is more important than behavioral one.
- In neurotic persons, there is a heightened influence of pupil-dilation, which is an indication of hyper-arousal sensitivity.

These developing causal-attention signals suggest that emotional expression and perception is not only stimuli-mediated, but trait-regulated, a fact that was rarely measured in computational affect modelling.

5.4 Ablation, Component Contribution

The ablation study (Fig. 7) is a measure of the marginal utility of every architectural element. The removability of causal edges lowers macro-F1 by 4.0 percentage, which supports the need of directed structural learning. Auto-temporal Coherence and Accuracy: The repression of Neural-ODE dynamics decreases temporal coherence and accuracy by 3.3 and personality fusion decreases personalization and generalization by 2.6. These scalabilities are consistent decreases that confirm that the performance improvements are based on architectural synergy, and not parameter scaling.

5.5 Causal Robustness and Temporal Robustness:

Figure 9 represents the causal flow graph, which is inferred. The standard route, which is EDA → EEG → Eye → Face → Cursor, is a biologically plausible progression in autonomic arousal to cortical activity and ending to behavioural output. CausalGraph-EmotionNet keeps its accuracy of over 80 per cent, but Transformer baselines decline to less than 68 per cent. under simulated modality dropout (up to 40 per cent). This resistance is created through masked-attention fusion and implicit causal regularization which unify cross-modal consistency despite a partiality of information.

5.6 Explainability and Neuroscientific Implications

Our framework generates intrinsic interpretability as it does not happen with the black-box deep models. The causal attention weightings are linked to a considerable sychophysiological process, and Neural-ODE are provided as continuous-time latent dynamics but can be interpreted as derivatives of affective states. Based on such mechanisms, latent emotion-transition laws, such as arousal-decay curves and valence drift rates, might be discovered that might be useful in cognitive-behavioural monitoring and adaptive feedback in a healthcare or education scenario. By connecting the causal representation learning with the physiological affect modelling, CausalGraph-EmotionNet will approach the plausible and empathetic AI in the realm of affective computing.

6. CONCLUSION

The proposed new model in this paper is CausalGraph-EmotionNet that integrates causal graph reason, Neural-ODE temporal evolution and personality-based multimodal fusion to achieve explainable emotion recognition through physiological and behavioural cues. This paper presents CausalGraph-EmotionNet, a new multimodal emotion recognition system, which is the first to combine dynamic causal graph learning, Neural-ODE-based continuous-time modeling, and personality-aware attention fusion in one end-to-end system. The offered model is contrasted with current correlation-based or time-independent graph models by the fact that time-varying directed causal interactions between neural, autonomic, and behavioral modalities are explicitly learned and could be used to infer emotions in a way that is interpretable and with physiological meaning. The integration of personality conditioning also serves to differentiate this work by enabling the model to adjust the modality relevance to personal characteristics, which develops individual-specific affective computing to the next level of population-level modeling. This work demonstrates the state-of-the-art performance gains based on a large-scale experimentation with the AFFECT dataset, reaching 84.6% accuracy and 80.8% macro-F1 without failing to surpass CNN, RNN, GCN, Transformer and PhysioGraph-Transformer baselines. In addition to the numerical enhancements, the ablation and robustness studies prove that every architectural element, causal edges, Neural-ODE dynamics, and personality-aware fusion, has a significant positive impact on performance, stability, and generalization. The work takes the multimodal emotion recognition beyond the stage of mere predictive modeling to one of causally interpretable, temporally consistent, and subject-specific affect analysis. The presented framework creates a platform to continue the future studies on the real-life emotion tracking, longitudinal affect modeling, and the clinically focused decision support systems. CausalGraph-EmotionNet provides a valuable step toward believable, accountable, and socially responsible affective AI by integrating causal reasoning with deep representation learning, which is important to solve essential issues of modern emotion recognition systems.

REFERENCES

[1] F. Liu, P. Yang, Y. Shu, F. Yan, G. Zhang and Y. -J. Liu, "Emotion Dictionary Learning With Modality Attentions for Mixed Emotion Exploration," in IEEE Transactions on Affective Computing, vol. 15, no. 3, pp. 1289-1302, July-Sept. 2024, doi: 10.1109/TAFFC.2023.3334520.

[2] Q. Zhu, C. Zheng, Z. Zhang, W. Shao and D. Zhang, "Dynamic Confidence-Aware Multi-Modal Emotion Recognition," in IEEE Transactions on Affective Computing, vol. 15, no. 3, pp. 1358-1370, July-Sept. 2024, doi: 10.1109/TAFFC.2023.3340924.

[3] M. D. Hssayeni and B. Ghoraani, "Multi-Modal Physiological Data Fusion for Affect Estimation Using Deep Learning," in IEEE Access, vol. 9, pp. 21642-21652, 2021, doi: 10.1109/ACCESS.2021.3055933.

[4] W. Xue, H. He, Y. Wang and Y. Zhao, "SAGN: Sparse Adaptive Gated Graph Neural Network With Graph Regularization for Identifying Dual-View Brain Networks," in IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 5, pp. 8085-8099, May 2025, doi: 10.1109/TNNLS.2024.3438835.

[5] J. Wang, Z. Feng, X. Ning, Y. Lin, B. Chen and Z. Jia, "Two-Stream Dynamic Heterogeneous Graph Recurrent Neural Network for Multi-Label Multi-Modal Emotion Recognition," in IEEE Transactions on Affective Computing, vol. 16, no. 3, pp. 2396-2409, July-Sept. 2025, doi: 10.1109/TAFFC.2025.3561439.

[6] M. Li, H. Chen, Y. Liu and Q. Zhao, "4s-SleepGCN: Four-Stream Graph Convolutional Networks for Sleep Stage Classification," in IEEE Access, vol. 11, pp. 70621-70634, 2023, doi: 10.1109/ACCESS.2023.3294410.

[7] A. Díaz-García, A. González-Robles, S. Mor, A. Mira, S. Quero, A. García-Palacios, R. M. Baños, and C. Botella, “Positive and negative affect schedule (PANAS): Psychometric properties of the online spanish version in a clinical sample with emotional disorders,” BMC Psychiatry, vol. 20, no. 1, p. 56, Dec. 2020.

[8] E. H. Nirjhar, A. Behzadan, and T. Chaspari, “Exploring bio-behavioral signal trajectories of state anxiety during public speaking,” in Proc. IEEE Int. Conf. Acoust., Speech Signal

Process. (ICASSP), May 2020, pp. 1294–1298.

[9] P. Bota, C. Wang, A. Fred, and H. Silva, "Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet? " Sensors, vol. 20, no. 17, p. 4723, Aug. 2020.

[10] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," IEEE Trans. Affect. Comput., vol. 11, no. 3, pp. 532–541, Jul. 2020.

[11] M. A. Asghar, M. J. Khan, M. Rizwan, R. M. Mehmood, and S.-H. Kim, "An innovative multi-model neural network approach for feature selection in emotion recognition using deep feature clustering," Sensors, vol. 20, no. 13, p. 3765, Jul. 2020.

[12] J. Cao, "Brain functional and effective connectivity based on electroencephalography recordings: A review," *Hum. Brain Mapping*, vol. 43, no. 2, pp. 860–879, Feb. 2022, doi: 10.1002/hbm.25683.

[13] L. E. Ismail and W. Karwowski, "A graph theory-based modeling of functional brain connectivity based on EEG: A systematic review in the context of neuroergonomics," *IEEE Access*, vol. 8, pp. 155103–155135, 2020, doi: 10.1109/ACCESS.2020.3018995.

[14] W. Xue and H. He, "A progressive learning classifier based on dynamic differential weighted network for feature identification of brain network series," *Knowl.-Based Syst.*, vol. 274, Aug. 2023, Art. no. 110661, doi: 10.1016/j.knosys.2023.110661.

[15] Narayana, V.L., Patibandla, R.S.M.L., Rao, B.T. and Gopi, A.P. (2022). Use of Machine Learning in Healthcare. In Advanced Healthcare Systems (eds R. Tanwar, S. Balamurugan, R.K. Saini, V. Bharti and P. Chithaluru). https://doi.org/10.1002/9781119769293.ch13

[16] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, and J. Wang, "SST-emotionnet: Spatial-spectral-temporal based attention 3D dense network for EEG emotion recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2909–2917.

[17] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Inf. Fusion*, vol. 53, pp. 209–221, 2020.

[18] Y. Fang, Y. Xia, P. Chen, J. Zhang, and Y. Zhang, "A dual-stream deep neural network integrated with adaptive boosting for sleep staging," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104150.

[19] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0216456.

[20] B. Tarakeswara Rao; R. S. M. Lakshmi Patibandla; V. Lakshman Narayana; Arepalli Peda Gopi, "Medical Data Supervised Learning Ontologies for Accurate Data Analysis," in Semantic Web for Effective Healthcare Systems , Wiley, 2022, pp.249-267, doi: 10.1002/9781119764175.ch11.

[21] Tsai, Y.H.H.; Bai, S.; Pu Liang, P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Volume 2019, pp. 6558–6569.

[22] Shou, Y.; Liu, H.; Cao, X.; Meng, D.; Dong, B. A Low-Rank Matching Attention Based Cross-Modal Feature Fusion Method for Conversational Emotion Recognition. *IEEE Trans. Affect. Comput.* **2025**, *16*, 1177–1189.

[23] Wu, L.; Liu, Q.; Zhang, D.; Wang, J.; Li, S.; Zhou, G. Multimodal Emotion Recognition with Auxiliary Sentiment Information. *Acta Sci. Nat. Univ. Pekin.* **2020**, *56*, 75–81.

[24] Jiao, W.; Lyu, M.; King, I. Real-Time Emotion Recognition via Attention Gated Hierarchical Memory Network. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8002–8009, Number 5.

[25] Wu, Y.; Zhang, S.; Li, P. Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features. *Sci. Rep.* **2025**, *15*, 8855.

[26] Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.