frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 11 November 2025
DOI 10.3389/frai.2025.1681759

OPEN ACCESS

EDITED BY
Durai Raj Vincent P. M.,
Vellore Institute of Technology, India

REVIEWED BY
Shanwen Zhang,
Xijing University, China
Alaa F. Sheta,
Southern Connecticut State University, United States

*CORRESPONDENCE
Qiangqiang Fu
✉ qiangqiang.fu@tongji.edu.cn
Hong Zhou
✉ zh720828@126.com

†These authors have contributed equally to this work

RECEIVED 11 August 2025
ACCEPTED 21 October 2025
PUBLISHED 11 November 2025

CITATION
Pei F, Zhou Y, Fu Q and Zhou H (2025) Real-time sleep disorder monitoring design using dynamic temporal graphs with facial and acoustic feature fusion.
Front. Artif. Intell. 8:1681759.
doi: 10.3389/frai.2025.1681759

Real-time sleep disorder monitoring design using dynamic temporal graphs with facial and acoustic feature fusion

Fei Pei[1†], Ying Zhou[2†], Qiangqiang Fu[3*] and Hong Zhou[1*]

[1]Department of Otolaryngology, Shidong Hospital, Shanghai, China, [2]Department of Geriatrics, Shidong Hospital, Shanghai, China, [3]Yangpu Hospital, School of Medicine, Tongji University, Shanghai, China

Introduction: Sleep disorders pose significant risks to patient safety, yet traditional polysomnography imposes substantial discomfort and laboratory constraints. We developed a non-invasive multimodal monitoring system for real-time sleep pathology detection.
Methods: We integrated facial expression analysis via deep convolutional neural networks with audio signal processing for breathing pattern detection. Heterogeneous data streams were unified into dynamic graph representations, with graph neural networks modeling spatiotemporal patterns of sleep pathologies.
Results: The system accurately detected sleep apnea, restless leg syndrome, and cardiovascular irregularities with 10.7-s average delay and 94.6% clinical agreement, achieving diagnostic accuracy comparable to polysomnography.
Conclusion: This framework enables continuous non-invasive monitoring for point-of-care screening and home-based management, potentially expanding sleep medicine access for underserved populations.

KEYWORDS

sleep disorder detection, facial expression analysis, real-time health monitoring, multimodal learning, machine learning

1 Introduction

Sleep disorders affect millions of people worldwide and represent a significant public health concern, with conditions such as sleep apnea, insomnia, and parasomnias contributing to increased morbidity, reduced quality of life, and elevated healthcare costs (Alshammari, 2024; Yildirim et al., 2019; Sharma et al., 2021b). The accurate detection and monitoring of sleep-related pathological conditions is crucial for timely medical intervention and prevention of serious complications (Morokuma et al., 2023; Arslan et al., 2023). Traditional sleep monitoring approaches, primarily relying on polysomnography (PSG) in controlled laboratory environments, while considered the gold standard, are expensive, time-consuming,

and often impractical for long-term monitoring or home-based care (Ha et al., 2023; Brink-Kjaer et al., 2022). Moreover, PSG requires multiple electrodes and sensors that can disturb patients' natural sleep patterns, potentially affecting the reliability of diagnostic outcomes (Rahman et al., 2025; Reis et al., 2024).

Recent advances in wearable technology and non-invasive monitoring systems have opened new avenues for sleep assessment. Current approaches predominantly focus on single-modality solutions, such as actigraphy for movement detection, heart rate variability analysis for autonomic nervous system assessment, or audio-based detection of breathing irregularities (Hussain et al., 2022; Yoon and Choi, 2023). However, these unimodal approaches suffer from several critical limitations. First, they often lack the comprehensive information necessary to capture the complex, multifaceted nature of sleep disorders, which typically manifest through various physiological and behavioral indicators simultaneously (Nguyen et al., 2023). Second, single-modality systems are susceptible to noise, artifacts, and environmental interference (Boiko et al., 2023), leading to reduced accuracy and reliability in real-world deployment scenarios.

Facial expression analysis has emerged as a promising non-invasive approach for detecting physiological states and emotional conditions during sleep (Maranci et al., 2021; Huang et al., 2023). Research has demonstrated that facial expressions can provide valuable insights into pain levels, breathing difficulties, and neurological activities during sleep. Similarly, audio signal analysis has shown significant potential in detecting sleep apnea events, snoring patterns, and other respiratory irregularities (Rosamaria et al., 2023; Xu et al., 2020). However, existing studies have primarily treated these modalities independently (Lv et al., 2020), failing to leverage their complementary information and temporal correlations.

The integration of multimodal data for sleep monitoring presents several fundamental challenges (Wang et al., 2025b). First, different modalities operate at varying temporal scales and exhibit distinct data characteristics, making it difficult to establish meaningful correlations and extract unified representations (Cheng et al., 2023; Torres et al., 2018). Facial expressions may change subtly over minutes, while audio signals contain high-frequency components that vary within seconds. Second, the temporal dependencies within and across modalities are complex and non-linear (Zhai et al., 2020; Zahid et al., 2023), requiring sophisticated modeling approaches that can capture both short-term fluctuations and long-term trends. Third, sleep disorders often manifest through subtle, gradual changes that may not be immediately apparent in individual modalities but become significant when considered collectively over extended periods (Duan et al., 2021; Lin et al., 2023). Existing multimodal fusion techniques, while successful in other domains, face specific challenges when applied to sleep monitoring (Liao et al., 2024). Traditional early fusion approaches that concatenate features from different modalities often result in high-dimensional representations that are prone to overfitting and computational inefficiency. Late fusion methods that combine decisions from individual modality classifiers may miss important cross-modal interactions (Zhai et al., 2021) that are crucial for accurate sleep disorder detection. Furthermore, most current approaches treat sleep monitoring as a static classification problem (Chung et al., 2017), ignoring the inherently dynamic and temporal nature of sleep processes.

To address these limitations, we propose a novel multimodal dynamic graph neural network framework that integrates facial expression analysis and sleep audio signal processing for real-time detection and prediction of sleep-related pathological conditions in Figure 1. Our approach is built upon several key insights and innovations. First, we conceptualize the multimodal sleep monitoring problem as a dynamic graph learning task, where **different modalities and their temporal states are represented as nodes in a time-evolving graph structure.** This representation naturally captures the heterogeneous nature of multimodal data while preserving the temporal dependencies crucial for understanding sleep dynamics. Nodes in our graph represent feature vectors extracted from facial expressions and audio signals at different time points, while edges encode both intra-modal temporal relationships and inter-modal correlations. Second, we develop a specialized graph neural network architecture that can effectively learn from this dynamic multimodal graph representation. Our model incorporates attention mechanisms to **automatically weight the importance of different modalities and temporal segments**, allowing the system to focus on the most relevant information for detecting specific sleep disorders. The architecture includes dedicated modules for processing facial expression data using convolutional neural networks optimized for low-light sleep environments, and audio processing components that can handle various acoustic patterns associated with different sleep pathologies. Third, we introduce a temporal modeling component that explicitly **captures the evolution of sleep states over time**. Unlike traditional approaches that analyze fixed time windows independently, our framework maintains a continuous representation of the patient's sleep state that evolves dynamically as new data becomes available. This enables early detection of developing conditions and provides predictive capabilities for anticipating potential sleep-related medical events.

Our technical approach consists of several interconnected components designed to address the specific challenges of multimodal sleep monitoring. The facial expression analysis module utilizes lightweight convolutional neural networks optimized for processing infrared or low-light facial images captured during sleep. We employ specialized preprocessing techniques to handle variations in lighting conditions, head pose changes, and occlusions commonly encountered in sleep environments. Feature extraction focuses on detecting micro-expressions and subtle facial movements that may indicate discomfort, breathing difficulties, or neurological activities. The audio processing component employs advanced signal processing techniques to extract meaningful features from sleep audio recordings. This includes spectral analysis for detecting breathing patterns, time-frequency analysis for identifying apnea events, and novel acoustic feature extraction methods for recognizing various sleep-related sounds. We address challenges related to background noise, signal variability across different recording devices, and the need for real-time

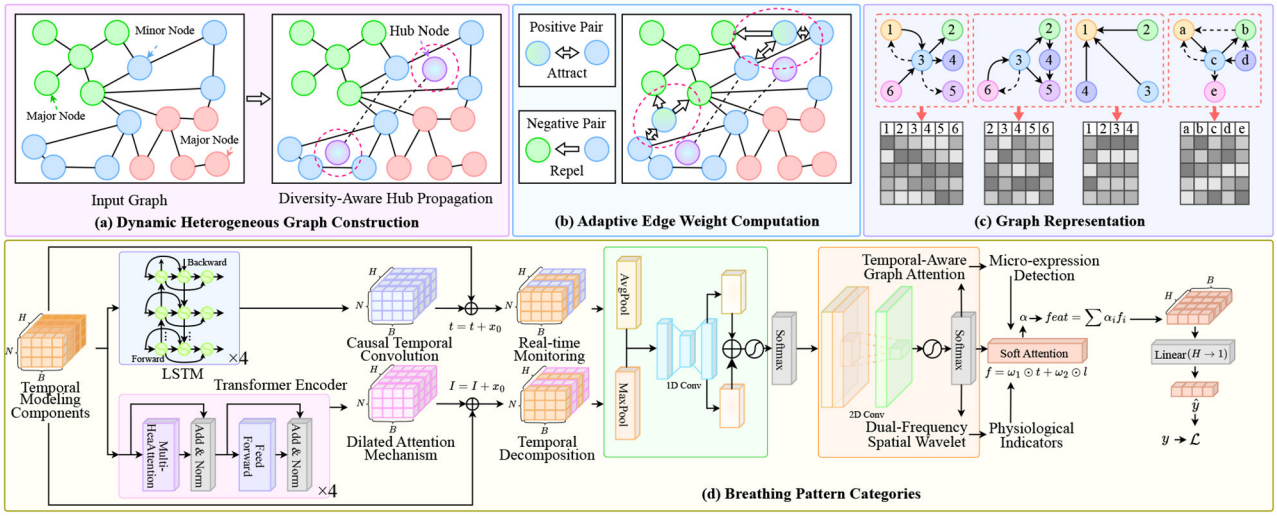

FIGURE 1
Overview of our multimodal dynamic graph network framework for sleep disorder monitoring. The system processes multimodal inputs through: **(A)** Dynamic heterogeneous graph construction with diversity-aware hub propagation to balance information flow across facial and audio modalities; **(B)** Adaptive edge weight computation using positive/negative pair attraction-repulsion mechanisms to enhance cross-modal alignment; **(C)** Graph representation encoding with temporal-aware attention for structural pattern learning; **(D)** Breathing pattern categorization module integrating LSTM-based temporal modeling, causal convolution for real-time monitoring, dilated attention mechanism for long-range dependencies, dual-frequency spatial wavelet analysis, and micro-expression detection for physiological indicators.

processing in resource-constrained environments. The dynamic graph construction mechanism creates time-evolving graph representations that capture the complex relationships between different modalities and their temporal evolution. We develop novel graph edge weighting schemes that automatically adapt based on the reliability and relevance of different modalities at different time points. This adaptive approach ensures robust performance even when individual modalities are compromised by noise or artifacts. Our graph neural network architecture incorporates several innovative components, including multi-scale temporal attention mechanisms, cross-modal correlation modules, and specialized pooling operations designed for handling irregular time series data. The model is trained using a combination of supervised learning for known sleep disorder patterns and self-supervised learning techniques that leverage the inherent structure of multimodal sleep data.

The proposed framework offers several significant advantages over existing approaches. By leveraging the complementary information from multiple modalities, our system can achieve higher accuracy and robustness compared to single-modality solutions. The dynamic graph representation enables the capture of complex temporal patterns that are crucial for understanding sleep disorders, while the attention mechanisms provide interpretability by highlighting the most relevant features and time periods for specific predictions. This research contributes to the growing field of multimodal health monitoring by providing a novel framework that can effectively integrate heterogeneous data sources for complex medical applications. Our work advances the state-of-the-art in both multimodal learning and sleep medicine, offering new possibilities for personalized and continuous healthcare monitoring solutions.

2 Methods

Let us formally define the multimodal sleep monitoring problem as a dynamic graph learning task. We denote the multimodal sleep data as a collection $\mathcal{D} = \{\mathcal{F}, \mathcal{A}\}$, where $\mathcal{F} = \{f_t\}_{t=1}^{T}$ represents the sequence of facial expression features and $\mathcal{A} = \{a_t\}_{t=1}^{T}$ represents the corresponding audio signal features over time horizon T. At each time step t, we have $f_t \in \mathbb{R}^{d_f}$ (facial features) and $a_t \in \mathbb{R}^{d_a}$ (audio features), where d_f and d_a are the dimensionalities of facial and audio feature spaces, respectively in Table 1. The objective is to learn a mapping function $\mathcal{M}: \mathcal{D} \rightarrow \mathcal{Y}$ that predicts sleep pathology labels $y_t \in \mathcal{Y} = \{0, 1, 2, ..., K\}$ at each time step, where K represents the number of distinct sleep disorder categories.

2.1 Facial expression feature extraction

For facial expression analysis, we employ a modified ResNeXt-50 architecture with specialized attention mechanisms for low-light sleep environments. The facial feature extraction process can be formulated as $X^{(0)} = \text{Preprocess}(I_t)$, $X^{(l+1)} = \mathcal{F}^{(l)}_{\text{ResNeXt}}(X^{(l)}, W^{(l)})$ and $f_t^{\text{raw}} = \text{GlobalAvgPool}(X^{(L)})$, where $I_t \in \mathbb{R}^{H \times W \times C}$ represents the input facial image at time t, $X^{(l)}$ denotes the feature maps at layer l, and $W^{(l)}$ are the learnable parameters (Yang et al., 2021). To enhance the feature representation for sleep-specific facial expressions, we introduce a temporal-spatial attention mechanism $A_{\text{spatial}} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$:

$$A_{\text{temporal}} = \text{softmax}\left(W_t \tanh(W_f f_t^{\text{raw}} + W_h h_{t-1})\right) \quad (1)$$

$$f_t = A_{\text{temporal}} \odot A_{\text{spatial}} V, \quad (2)$$

TABLE 1 Mathematical notation and symbols used in methods section.

Symbol	Description	Symbol	Description
$\mathcal{D}$	Multimodal sleep data collection	G_t	Dynamic heterogeneous graph at time t
$\mathcal{F}$	Facial expression feature sequence	$\mathcal{V}_t$	Node set containing facial and audio nodes
$\mathcal{A}$	Audio signal feature sequence	$\mathcal{E}_t$	Edge set for intra- and cross-modal connections
f_t	Facial features at time t	X_t	Node feature matrix at time t
a_t	Audio features at time t	d_f, d_a	Dimensionalities of facial and audio features
y_t	Sleep pathology labels at time t	K	Number of sleep disorder categories
T	Time horizon	$\mathcal{M}$	Mapping function for prediction
I_t	Input facial image at time t	H, W, C	Image height, width, and channels
$X^{(l)}$	Feature maps at layer l	$W^{(l)}$	Learnable parameters at layer l
f_t^{raw}	Raw facial features before attention	$A_{spatial}$	Spatial attention mechanism
$A_{temporal}$	Temporal attention mechanism	Q, K, V	Query, key, and value matrices
h_{t-1}	Hidden state from previous time step	W_t, W_f, W_h	Learnable weight matrices
S_t	Short-Time Fourier Transform at time t	$\psi_{j,k}$	Mother wavelet at scale j, position k
M_t	Power Spectral Density	C_t	Cepstral coefficients
ZCR_t	Zero Crossing Rate	RMS_t	Root Mean Square energy
SC_t	Spectral Centroid	SRO_t	Spectral Rolloff
$W_{1:J,t}$	Wavelet coefficients	N	Number of samples
x_t^f, x_t^a	Projected facial and audio node features	W_f, W_a	Projection matrices
α_{ij}^{temp}	Temporal edge attention weight	α_{ij}^{cross}	Cross-modal edge attention weight
w_{ij}	Final edge weight	$\lambda_1, \lambda_2, \lambda_3$	Hyperparameters
γ	Temporal decay rate	$\mathcal{N}_i$	Neighborhood of node i
$H^{(l)}$	Hidden representations at layer l	$A_s^{(l)}$	Adjacency matrix at scale s
S	Number of temporal scales	D	Degree matrix
$e_{ij}^{(l)}$	Attention energy between nodes i, j	$\alpha_{ij}^{(l)}$	Attention coefficient
$\phi(t_i, t_j)$	Temporal relationship encoding	ω_d	Frequency parameters
$h_f^{(L)}, h_a^{(L)}$	Final layer facial and audio features	Q_f, K_a, V_a	Cross-modal attention components
$\text{Attn}_{f\to a}$	Facial-to-audio attention	$\text{Attn}_{a\to f}$	Audio-to-facial attention
h_{fused}	Fused multimodal representation	d_k	Key dimension
r_t, z_t	Reset and update gates in GRU	$\tilde{s}_t$	Candidate hidden state
s_t	Final hidden state	U_r, U_z, U_s	Recurrent weight matrices
$s_t^{(\ell)}$	Multi-scale decomposition at level ℓ	K_ℓ	Number of wavelets at level ℓ
$\alpha_k^{(\ell)}$	Learnable wavelet coefficients	ϕ	Mother wavelet function
$h_t^{(c)}$	Causal convolution output	k	Kernel size
d	Dilation factor	M_{causal}	Causal attention mask
R	Attention radius	W_{pos}	Positional encoding weights
$\mathcal{L}_{cls}$	Classification loss	$\mathcal{L}_{temp}$	Temporal consistency loss
$\mathcal{L}_{cont}$	Contrastive loss	$\mathcal{L}_{rec}$	Reconstruction loss
α_k	Class-specific weights	γ	Focusing parameter
$\hat{y}_{t,k}$	Predicted probability for class k	ω_t	Adaptive temporal weight
β	Similarity threshold parameter	τ	Temperature parameter
η_t	Learning rate at time t	η_{min}, η_{max}	Minimum and maximum learning rates
T_{cur}	Current epoch in restart cycle	T_i	Epochs in restart cycle

where Q, K, V are query, key, and value matrices, W_t, W_f, W_h are learnable weight matrices, h_{t-1} is the hidden state from the previous time step, and $\odot$ denotes element-wise multiplication.

2.2 Audio signal feature extraction

For audio signal processing, we implement a multi-scale wavelet transform combined with spectral analysis. The audio feature extraction pipeline is defined as $S_t = \text{STFT}(a_t^{\text{raw}})$, $W_{j,k} = \sum_n a_t^{\text{raw}}[n]\psi_{j,k}^*[n-k]$, $M_t = |S_t|^2$ (Power Spectral Density), and $C_t = \text{DCT}(\log(M_t))$ (Cepstral Coefficients), where STFT denotes the Short-Time Fourier Transform (Karpagam et al., 2022), $\psi_{j,k}$ represents the mother wavelet at scale j and position k, and DCT is the Discrete Cosine Transform. We extract multiple acoustic features including:

$$\text{ZCR}_t = \frac{1}{2N}\sum_{n=1}^{N-1}|\text{sgn}(a_t[n]) - \text{sgn}(a_t[n-1])| \tag{3}$$

$$\text{RMS}_t = \sqrt{\frac{1}{N}\sum_{n=1}^{N} a_t[n]^2} \tag{4}$$

$$\text{SC}_t = \frac{\sum_{k=1}^{K} k \cdot |S_t[k]|}{\sum_{k=1}^{K}|S_t[k]|} \tag{5}$$

$$\text{SRO}_t = \frac{\sum_{k=1}^{K}(k-\text{SC}_t)^2 \cdot |S_t[k]|}{\sum_{k=1}^{K}|S_t[k]|}, \tag{6}$$

where ZCR is Zero Crossing Rate, RMS is Root Mean Square energy, SC is Spectral Centroid, and SRO is Spectral Rolloff. The final audio feature vector is constructed as $a_t = [C_t; \text{ZCR}_t; \text{RMS}_t; \text{SC}_t; \text{SRO}_t; W_{1:J,t}]$.

2.3 Dynamic graph construction

2.3.1 Graph topology design

We construct a dynamic heterogeneous graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t, X_t)$ where $\mathcal{V}_t = \mathcal{V}_t^f \cup \mathcal{V}_t^a$ represents the node set containing facial and audio nodes, $\mathcal{E}_t = \mathcal{E}_t^{ff} \cup \mathcal{E}_t^{aa} \cup \mathcal{E}_t^{fa}$ represents edges within and across modalities - $X_t \in \mathbb{R}^{|\mathcal{V}_t| \times d}$ is node feature matrix (Chen et al., 2025; Hou et al., 2016). The features are constructed using a projection mechanism $x_t^f = W_f f_t + b_f$, $x_t^a = W_a a_t + b_a$, where $W_f \in \mathbb{R}^{d \times d_f}$, $W_a \in \mathbb{R}^{d \times d_a}$ are projection matrices map different modalities.

2.3.2 Adaptive edge weight computation

The edge weights are computed using a learnable attention mechanism that considers both temporal and cross-modal dependencies:

$$\alpha_{ij}^{\text{temp}} = \frac{\exp(W_{\text{temp}}^T \tanh(W_1 x_i + W_2 x_j))}{\sum_{k\in\mathcal{N}_i}\exp(W_{\text{temp}}^T \tanh(W_1 x_i + W_2 x_k))} \tag{7}$$

$$\alpha_{ij}^{\text{cross}} = \text{sigmoid}(W_{\text{cross}}^T [x_i || x_j || (x_i \odot x_j)]) \tag{8}$$

$$w_{ij} = \lambda_1 \alpha_{ij}^{\text{temp}} + \lambda_2 \alpha_{ij}^{\text{cross}} + \lambda_3 \exp(-\gamma |t_i - t_j|), \tag{9}$$

where $\mathcal{N}_i$ represents the neighborhood of node i, $||$ denotes concatenation, $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters, and γ controls the temporal decay rate.

2.4 Dynamic graph neural network architecture

2.4.1 Multi-scale graph convolution

We propose a multi-scale graph convolutional layer that operates on different temporal scales simultaneously:

$$H^{(l+1)} = \sigma\left(\sum_{s=1}^{S} A_s^{(l)} H^{(l)} W_s^{(l)}\right) \tag{10}$$

$$A_s^{(l)} = \text{GraphConv}_s(A_t, H^{(l)}) \tag{11}$$

$$\text{GraphConv}_s(A, H) = D^{-\frac{1}{2}} A_s D^{-\frac{1}{2}} H, \tag{12}$$

where S is the number of scales, A_s is the adjacency matrix at scale s, D is the degree matrix, and σ is an activation function (Wang et al., 2025a).

2.4.2 Temporal-aware graph attention

To capture long-range temporal dependencies, we implement a temporal-aware graph attention mechanism:

$$e_{ij}^{(l)} = \text{LeakyReLU}(a^T [W h_i^{(l)} || W h_j^{(l)} || \phi(t_i, t_j)]) \tag{13}$$

$$\alpha_{ij}^{(l)} = \frac{\exp(e_{ij}^{(l)})}{\sum_{k\in\mathcal{N}_i\cup\{i\}}\exp(e_{ik}^{(l)})} \tag{14}$$

$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}_i\cup\{i\}} \alpha_{ij}^{(l)} W h_j^{(l)}\right), \tag{15}$$

where $\phi(t_i, t_j)$ encodes temporal relationships:

$$\phi(t_i, t_j) = [\sin(\omega_1(t_i - t_j)), \cos(\omega_1(t_i - t_j)), ..., \sin(\omega_d(t_i - t_j)), \cos(\omega_d(t_i - t_j))] \tag{16}$$

2.4.3 Cross-modal fusion module

The cross-modal fusion is achieved through a specialized attention-based fusion mechanism (Chen et al., 2024):

$$Q_f = h_f^{(L)} W_Q^f, \quad K_a = h_a^{(L)} W_K^a, \quad V_a = h_a^{(L)} W_V^a \tag{17}$$

$$Q_a = h_a^{(L)} W_Q^a, \quad K_f = h_f^{(L)} W_K^f, \quad V_f = h_f^{(L)} W_V^f \tag{18}$$

$$\text{Attn}_{f\to a} = \text{softmax}\left(\frac{Q_f K_a^T}{\sqrt{d_k}}\right) V_a \tag{19}$$

$$\text{Attn}_{a\to f} = \text{softmax}\left(\frac{Q_a K_f^T}{\sqrt{d_k}}\right) V_f \tag{20}$$

$$h_{\text{fused}} = \text{LayerNorm}(h_f^{(L)} + \text{Attn}_{a\to f}) + \text{LayerNorm}(h_a^{(L)} + \text{Attn}_{f\to a}) \tag{21}$$

2.5 Temporal sequence modeling

2.5.1 Gated recurrent unit with graph embedding

We incorporate a modified GRU that operates on graph embeddings to capture temporal dynamics:

$$r_t = \sigma(W_r h_{\text{fused},t} + U_r s_{t-1} + b_r) \tag{22}$$

$$z_t = \sigma(W_z h_{\text{fused},t} + U_z s_{t-1} + b_z) \tag{23}$$

$$\tilde{s}_t = \tanh(W_s h_{\text{fused},t} + U_s(r_t \odot s_{t-1}) + b_s) \tag{24}$$

$$s_t = (1 - z_t) \odot s_{t-1} + z_t \odot \tilde{s}_t, \tag{25}$$

where r_t, z_t, and $\tilde{s}_t$ are the reset gate, update gate, and candidate hidden state, respectively.

2.5.2 Hierarchical temporal decomposition

Given the multi-scale nature of sleep disorders, which can manifest over different temporal horizons ranging from seconds to hours, we implement a hierarchical temporal decomposition mechanism (Tiwari et al., 2022). This approach decomposes the temporal sequences into multiple frequency components using learnable wavelet-based filters. The decomposition process is formulated as:

$$s_t^{(\ell)} = \sum_{k=1}^{K_\ell} \alpha_k^{(\ell)} \psi_{\ell,k}(s_{t-\Delta_\ell:t}) \tag{26}$$

$$\psi_{\ell,k}(x) = \frac{1}{\sqrt{2^\ell}} \sum_n W_{\ell,k} x[n] \phi\left(\frac{n - k \cdot 2^\ell}{2^\ell}\right) \tag{27}$$

$$s_t^{\text{multi}} = \text{Concat}(s_t^{(1)}, s_t^{(2)}, ..., s_t^{(L)}) W_{\text{proj}}, \tag{28}$$

where ℓ denotes the decomposition level, K_ℓ is the number of wavelets at level ℓ, $\alpha_k^{(\ell)}$ are learnable coefficients, ϕ is the mother wavelet function, and W_{proj} projects the concatenated multi-scale features back to the original dimension. This hierarchical approach enables the model to simultaneously capture short-term fluctuations in breathing patterns and long-term trends in sleep stage transitions (Yang et al., 2022).

2.5.3 Causal temporal convolution with dilated attention

To ensure that predictions at time t only depend on past observations while maintaining computational efficiency, we introduce causal temporal convolutions with dilated attention mechanisms. The causal convolution operation is defined as:

$$h_t^{(c)} = \sum_{i=0}^{k-1} W_i^{(c)} s_{t-i \cdot d} + b^{(c)} \tag{29}$$

$$\text{DilatedAttn}(H^{(c)}) = \text{softmax}\left(\frac{Q^{(c)}(K^{(c)})^T}{\sqrt{d_k}} \odot M_{\text{causal}}\right) V^{(c)} \tag{30}$$

$$M_{\text{causal}}[i,j] = \begin{cases} 0 & \text{if } i < j \\ -\infty & \text{if } i \geq j \text{ and } |i-j| > R \\ W_{\text{pos}}[|i-j|] & \text{otherwise} \end{cases} \tag{31}$$

Require: Facial image sequence $\{I_t\}_{t=1}^T$, Audio signal sequence $\{a_t^{raw}\}_{t=1}^T$
Ensure: Dynamic graph sequence $\{G_t\}_{t=1}^T$ with node features $\{X_t\}_{t=1}^T$
Initialize: ResNeXt-50 network, wavelet filters, projection matrices W_f, W_a
for $t = 1$ to T **do**
  // Facial Feature Extraction
  $X^{(0)} \leftarrow \text{Preprocess}(I_t)$ {Face detection and normalization}
  for $l = 0$ to $L-1$ **do**
    $X^{(l+1)} \leftarrow \text{ResNeXt}^{(l)}(X^{(l)}, W^{(l)})$
  end for
  $f_t^{raw} \leftarrow \text{GlobalAvgPool}(X^{(L)})$
  // Temporal-Spatial Attention
  $A_{spatial} \leftarrow \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$
  $A_{temporal} \leftarrow \text{softmax}(W_t \tanh(W_f f_t^{raw} + W_h h_{t-1}))$
  $f_t \leftarrow A_{temporal} \odot A_{spatial} V$
  // Audio Feature Extraction
  $S_t \leftarrow \text{STFT}(a_t^{raw})$ {Short-Time Fourier Transform}
  $M_t \leftarrow |S_t|^2$ {Power Spectral Density}
  $C_t \leftarrow \text{DCT}(\log(M_t))$ {Cepstral Coefficients}
  // Multi-scale Wavelet Analysis
  for $j = 1$ to J, $k = 1$ to K_j **do**
    $W_{j,k} \leftarrow \sum_n a_t^{raw}[n] \psi_{j,k}^*[n-k]$
  end for
  // Acoustic Feature Computation
  $\text{ZCR}_t \leftarrow \frac{1}{2N} \sum_{n=1}^{N-1} |\text{sgn}(a_t[n]) - \text{sgn}(a_t[n-1])|$
  $\text{RMS}_t \leftarrow \sqrt{\frac{1}{N} \sum_{n=1}^{N} a_t[n]^2}$
  $a_t \leftarrow [C_t; \text{ZCR}_t; \text{RMS}_t; \text{SC}_t; \text{SRO}_t; W_{1:J,t}]$
  // Node Feature Projection
  $x_t^f \leftarrow W_f f_t + b_f$, $x_t^a \leftarrow W_a a_t + b_a$
  // Adaptive Edge Weight Computation
  for each node pair (i, j) **do**
    $\alpha_{ij}^{temp} \leftarrow \frac{\exp(W_{temp}^T \tanh(W_1 x_i + W_2 x_j))}{\sum_{k \in \mathcal{N}_i} \exp(W_{temp}^T \tanh(W_1 x_i + W_2 x_k))}$
    $\alpha_{ij}^{cross} \leftarrow \text{sigmoid}(W_{cross}^T [x_i || x_j || (x_i \odot x_j)])$
    $w_{ij} \leftarrow \lambda_1 \alpha_{ij}^{temp} + \lambda_2 \alpha_{ij}^{cross} + \lambda_3 \exp(-\gamma |t_i - t_j|)$
  end for
end for
return $\{G_t\}_{t=1}^T$, $\{X_t\}_{t=1}^T$

Algorithm 1. Multimodal feature extraction and dynamic graph construction.

where k is the kernel size, d is the dilation factor, M_{causal} is the causal mask that prevents information leakage from future time steps, R is the attention radius, and W_{pos} encodes positional relationships. This design allows the model to capture long-range dependencies while maintaining the causal property essential for real-time sleep monitoring applications.

2.6 Loss function and optimization strategy

The training of our dynamic graph neural network requires a sophisticated loss function that addresses multiple objectives

simultaneously while ensuring stable convergence (Li et al., 2024). Our comprehensive loss function incorporates classification accuracy, temporal consistency, cross-modal alignment, and regularization terms to prevent overfitting and enhance generalization capabilities.

The primary classification loss employs a weighted focal loss mechanism to address the inherent class imbalance in sleep disorder datasets. The focal loss is particularly effective for handling rare pathological events that may occur infrequently during sleep but are critical for early detection. The mathematical formulation is given by:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\alpha_k(1-\hat{y}_{t,k})^{\gamma}y_{t,k}\log(\hat{y}_{t,k}), \tag{32}$$

where α_k represents class-specific weights derived from inverse frequency statistics, γ is the focusing parameter that reduces the relative loss for well-classified examples, and $\hat{y}_{t,k}$ denotes the predicted probability for class k at time t.

To ensure temporal consistency in predictions, we introduce a specialized temporal smoothness loss that penalizes abrupt transitions between predicted sleep states unless supported by significant changes in the input modalities. This loss is computed as:

$$\mathcal{L}_{\text{temp}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\omega_t||\hat{y}_{t+1}-\hat{y}_t||_2^2, \tag{33}$$

where $\omega_t = \exp(-\beta \cdot \text{sim}(h_{\text{fused},t+1}, h_{\text{fused},t}))$ is an adaptive weight that allows larger prediction changes when the fused representations differ significantly, controlled by the similarity threshold parameter β.

Cross-modal alignment is enforced through a contrastive learning objective that maximizes the mutual information between facial and audio representations when they correspond to the same sleep state while minimizing it for different states. The contrastive loss is formulated as:

$$\mathcal{L}_{\text{cont}} = -\sum_{i,j}\mathbb{I}[y_i = y_j]\log\frac{\exp(\text{sim}(h_i^f, h_j^a)/\tau)}{\sum_k \exp(\text{sim}(h_i^f, h_k^a)/\tau)}, \tag{34}$$

where $\mathbb{I}[\cdot]$ is the indicator function, $\text{sim}(\cdot,\cdot)$ computes cosine similarity, and τ is the temperature parameter that controls the concentration of the distribution.

The reconstruction loss serves as a regularization mechanism that encourages the learned representations to preserve essential information from both modalities. This autoencoder-style loss is computed as:

$$\mathcal{L}_{\text{rec}} = \sum_{t=1}^{T}||f_t - \text{Dec}_f(h_{\text{fused},t})||_2^2 + ||a_t - \text{Dec}_a(h_{\text{fused},t})||_2^2, \tag{35}$$

where Dec_f and Dec_a are lightweight decoder networks that reconstruct the original modal features from the fused representation.

The optimization strategy employs adaptive learning rate scheduling combined with gradient clipping to ensure stable training dynamics. We utilize the AdamW optimizer with decoupled weight decay, where the learning rate follows a cosine annealing schedule with warm restarts:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})(1 + \cos(\frac{T_{\text{cur}}}{T_i}\pi)), \tag{36}$$

where T_{cur} is the number of epochs since the last restart and T_i is the number of epochs in the current restart cycle. The gradient clipping threshold is dynamically adjusted based on the gradient norm history using an exponential moving average to prevent gradient explosion while allowing for occasional large updates during critical learning phases.

2.7 Model architecture and implementation details

The complete architecture of our dynamic multimodal graph neural network is carefully designed to balance computational efficiency with representational power, enabling real-time processing while maintaining high accuracy for sleep disorder detection. The facial expression processing branch utilizes a modified ResNeXt-50 architecture with specialized adaptations for low-light infrared imagery commonly encountered in sleep monitoring scenarios. The initial convolutional layers employ depthwise separable convolutions to reduce computational overhead while maintaining feature extraction capability, followed by residual blocks with cardinality-based grouped convolutions that effectively capture spatial hierarchies in facial expressions.

The audio processing pipeline incorporates multi-scale temporal convolutional networks with varying receptive fields to capture acoustic patterns across different time scales simultaneously. The architecture employs dilated causal convolutions with exponentially increasing dilation rates, allowing the network to model both short-term acoustic events such as individual breaths or snores, and long-term patterns such as periodic breathing irregularities. Spectral normalization is applied to all convolutional layers to ensure training stability and prevent mode collapse, particularly important when processing variable-quality audio recordings from different environments. The graph neural network component consists of four specialized layers, each designed to capture different aspects of the multimodal temporal relationships. The first layer performs initial node embedding and establishes basic connectivity patterns between facial and audio nodes. Subsequent layers progressively refine these relationships through learnable attention mechanisms that dynamically adjust edge weights based on the current sleep state and temporal context. The final graph layer incorporates global pooling operations that aggregate information across all nodes while preserving modality-specific characteristics through separate attention heads.

Regularization strategies are implemented throughout the architecture to prevent overfitting and enhance generalization to new patients and environments. These include adaptive dropout with time-varying probabilities, batch normalization with momentum adjustment based on training progress, and

Require: Graph sequence $\{G_t\}_{t=1}^{T}$, Ground truth labels $\{y_t\}_{t=1}^{T}$
Ensure: Trained DGNN model parameters Θ
Initialize: Model parameters Θ, optimizer, learning rate schedule
Initialize: Loss weights α_k, hyperparameters γ, β, τ
while not converged **do**
  for each training batch **do**
    // Forward Pass
    for $t = 1$ to T **do**
      // Multi-Scale Graph Convolution
      for $s = 1$ to S **do**
        $A_s^{(l)} \leftarrow \text{GraphConv}_s(A_t, H^{(l)})$
        $A_s^{(l)} \leftarrow D^{-\frac{1}{2}} A_s D^{-\frac{1}{2}} H^{(l)}$
      end for
      $H^{(l+1)} \leftarrow \sigma\left(\sum_{s=1}^{S} A_s^{(l)} H^{(l)} W_s^{(l)}\right)$
      // Temporal-Aware Graph Attention
      for each node i **do**
        for each neighbor $j \in \mathcal{N}_i \cup \{i\}$ **do**
          $e_{ij}^{(l)} \leftarrow \text{LeakyReLU}(a^T[Wh_i^{(l)} || Wh_j^{(l)} || \phi(t_i, t_j)])$
          $\alpha_{ij}^{(l)} \leftarrow \frac{\exp(e_{ij}^{(l)})}{\sum_{k \in \mathcal{N}_i \cup \{i\}} \exp(e_{ik}^{(l)})}$
        end for
        $h_i^{(l+1)} \leftarrow \sigma\left(\sum_{j \in \mathcal{N}_i \cup \{i\}} \alpha_{ij}^{(l)} Wh_j^{(l)}\right)$
      end for
      // Cross-Modal Fusion
      $Q_f \leftarrow h_f^{(L)} W_Q^f$, $K_a \leftarrow h_a^{(L)} W_K^a$, $V_a \leftarrow h_a^{(L)} W_V^a$
      $h_{fused} \leftarrow \text{LayerNorm}(h_f^{(L)} + \text{Attn}_{a \to f}) + \text{LayerNorm}(h_a^{(L)} + \text{Attn}_{f \to a})$
    end for
    // Multi-Objective Loss Computation
    $\mathcal{L}_{cls} \leftarrow -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} \alpha_k (1 - \hat{y}_{t,k})^{\gamma} y_{t,k} \log(\hat{y}_{t,k})$
    $\mathcal{L}_{temp} \leftarrow \frac{1}{T-1}\sum_{t=1}^{T-1} \omega_t ||\hat{y}_{t+1} - \hat{y}_t||_2^2$
    where $\omega_t \leftarrow \exp(-\beta \cdot \text{sim}(h_{fused,t+1}, h_{fused,t}))$
    $\mathcal{L}_{cont} \leftarrow -\sum_{i,j} \mathbb{I}[y_i = y_j] \log \frac{\exp(\text{sim}(h_i^f, h_j^a)/\tau)}{\sum_k \exp(\text{sim}(h_i^f, h_k^a)/\tau)}$
    $\mathcal{L}_{rec} \leftarrow \sum_{t=1}^{T} ||f_t - \text{Dec}_f(h_{fused,t})||_2^2 + ||a_t - \text{Dec}_a(h_{fused,t})||_2^2$
    $\mathcal{L}_{total} \leftarrow \mathcal{L}_{cls} + \lambda_{temp}\mathcal{L}_{temp} + \lambda_{cont}\mathcal{L}_{cont} + \lambda_{rec}\mathcal{L}_{rec}$
    Update parameters: $\Theta \leftarrow \text{AdamW}(\Theta, \nabla_\Theta \mathcal{L}_{total})$
    Update learning rate: $\eta_t \leftarrow \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(\frac{T_{cur}}{T_i}\pi))$
  end for
end while
return Optimized model parameters Θ

Algorithm 2. Dynamic graph neural network training with multi-objective loss.

spectral regularization of weight matrices to control the Lipschitz constant of the learned mappings. The model employs early stopping with patience scheduling and checkpoint averaging to select optimal parameters while preventing overfitting to the training distribution.

3 Results

3.1 Experimental setup

3.1.1 Datasets and data collection

We evaluate our proposed multimodal dynamic graph neural network framework on two comprehensive sleep monitoring datasets. The primary dataset consists of recordings from 156 participants collected over 18 months at three sleep laboratories affiliated with major medical institutions in Table 2. Each participant underwent overnight polysomnography monitoring while simultaneously recording facial expressions using infrared cameras and ambient audio signals through calibrated microphones. The participants ranged in age from 22 to 78 years (mean: 51.3 ± 14.7 years), with 68 males and 88 females, representing diverse demographic backgrounds and sleep disorder prevalences.

Data collection protocols were standardized across all recording sites to ensure consistency and reliability. Facial video recordings were captured at 30 frames per second using infrared cameras positioned at a fixed distance and angle relative to the participant's head. Audio signals were recorded at 44.1 kHz sampling rate using omnidirectional microphones placed at standardized positions within the sleep laboratory. Synchronization between video, audio, and polysomnography signals was maintained through hardware-level timestamping with sub-millisecond accuracy.

3.1.2 Data preprocessing and quality control

Comprehensive preprocessing pipelines were developed to handle the inherent challenges of multimodal sleep data, including varying signal qualities, environmental artifacts, and participant-specific variations. For facial video processing, we implemented robust face detection and tracking algorithms capable of handling partial occlusions, head pose variations, and lighting changes common in sleep environments (Sharma et al., 2021a; Widasari et al., 2020). Facial landmarks were extracted using a modified version of the MediaPipe framework, with additional temporal smoothing to reduce jitter and improve stability across consecutive frames.

Audio preprocessing involved multi-stage filtering to remove environmental noise while preserving sleep-related acoustic signatures. We applied adaptive spectral subtraction for background noise reduction, followed by dynamic range compression to normalize signal amplitudes across different recording conditions (Sathyanarayana et al., 2016). Artifact detection algorithms were developed to identify and flag segments contaminated by equipment noise, external disturbances, or signal clipping, ensuring that only high-quality data segments were included in the training and evaluation processes. Quality control measures included automated screening for data integrity, completeness, and annotation consistency (Rahman et al., 2025). Recordings with more than 15% missing data, significant synchronization errors, or poor signal quality were excluded from the analysis (Sravani et al., 2024). Additionally, we implemented cross-validation procedures to verify annotation accuracy, achieving inter-annotator agreement scores (Cohen's kappa)

TABLE 2 Model architecture and ing parameters.

Component	Parameter	Value	Component	Parameter	Value
Data preprocessing			**Graph neural network**		
Facial resolution	$H \times W$	224 × 224	GNN layers	L	4
Audio sampling	f_s	44.1 kHz	Hidden dims	-	[512, 384, 256, 128]
Time window	T_{window}	30 seconds	Dropout Rate	-	0.3
Overlap ratio	-	50%	Activation	σ	LeakyReLU
Facial expression module			**Temporal modeling**		
Backbone	-	ResNeXt-50	GRU hidden	-	256
Input dimension	d_f	2048	Hierarchical levels	$\mathcal{L}$	3
Projection Dim	d^f_{proj}	512	Conv Kernel	k	3
Cardinality	-	32	Dilation rates	d	[1, 2, 4, 8]
Attention heads	-	8	Attention radius	R	16
Context length	-	16 frames	Pos Encoding	-	128
Audio processing module			**Loss parameters**		
STFT window	-	2,048 samples	Focal gamma	γ	2.0
Hop length	-	512 samples	Temperature	τ	0.1
Mel banks	-	128	Similarity Thresh	β	0.5
MFCCs	-	13	$\mathcal{L}_{cls}$ Weight	-	1.0
Wavelet scales	J	8 levels	$\mathcal{L}_{temp}$ Weight	-	0.3
Input dimension	d_a	256	$\mathcal{L}_{cont}$ Weight	-	0.2
Projection dim	d^a_{proj}	512	$\mathcal{L}_{rec}$ Weight	-	0.1
Dynamic Graph			**Training Config**		
Node embedding	d	512	Batch size	-	16
Temporal scales	S	4	Initial LR	η_0	1×10^{-3}
Graph attn heads	-	4	LR schedule	-	Cosine annealing
Edge decay rate	γ	0.1	Min/Max LR	$\eta_{min/max}$	$10^{-6}/10^{-3}$
Fusion weights	$\lambda_{1,2,3}$	0.4, 0.4, 0.2	Optimizer	-	AdamW
Max connectivity	-	85%	Weight decay	-	1×10^{-4}
Attention key dim	d_k	64	Gradient clip	-	Max norm = 1.0
Model complexity and performance					
Total parameters		12.3M	Inference time		23.4 ms/step
Trainable parameters		11.8M	Training memory		6.8 GB
Model size		47.2 MB	Inference memory		1.2 GB

of 0.89 for sleep stage classification and 0.92 for pathological event detection.

3.1.3 Experimental configuration

Training procedures employed stratified random splitting to ensure balanced representation of different sleep disorders and demographic groups across training, validation, and test sets. The data split followed a 70-15-15 ratio for training, validation, and testing respectively, with careful attention to maintaining temporal independence between splits to prevent data leakage. Cross-validation was performed using a modified time-series splitting approach that respects the temporal nature of sleep data while ensuring adequate sample sizes for each fold. Hyperparameter optimization was conducted using Bayesian optimization with Gaussian process surrogates, exploring the space of learning rates, regularization parameters, attention mechanisms weights, and architectural choices. The optimization process considered both validation accuracy and computational efficiency, resulting in Pareto-optimal configurations suitable for different deployment

scenarios ranging from high-accuracy clinical applications to resource-constrained mobile implementations.

Equipment specifications were standardized across sites: FLIR Lepton 3.5 infrared cameras (160×120 resolution, 8–14 μm spectral range, 9 Hz frame rate) positioned 1.5 meters from the bed at a 30-degree downward angle; Audio-Technica AT4040 cardioid condenser microphones with Focusrite Scarlett 2i2 interfaces (44.1 kHz/24-bit sampling); and Compumedics Grael 4K PSG systems for ground truth acquisition. Environmental conditions were controlled: ambient temperature 22 $\pm$ 1°C, humidity 45 $-$ 55%, background noise $<$ 35 dB SPL. Data synchronization employed hardware timestamps via SMPTE timecode generators ensuring $<$ 1 ms inter-modal alignment. Inclusion criteria required participants aged 18-80 years without severe cardiac arrhythmias or neurodegenerative conditions. The secondary validation dataset included 312 recordings from two independent sites following identical protocols, collected between July 2023 and December 2023.

3.2 Baseline methods and comparison framework

3.2.1 Traditional machine learning approaches

We implemented several state-of-the-art traditional machine learning methods as baseline comparisons to demonstrate the effectiveness of our deep learning approach. Support Vector Machines (SVM) with radial basis function kernels were trained on handcrafted features (Liu et al., 2020) extracted from both facial and audio modalities. The feature engineering process involved extensive domain knowledge incorporation, including facial action unit detection, acoustic spectral features, and temporal statistical measures computed over sliding windows of varying durations.

Random Forest ensembles were configured with 500 decision trees, employing bootstrap aggregation and feature randomization to improve generalization performance (Wara et al., 2025). The feature selection process utilized mutual information criteria to identify the most discriminative attributes for sleep disorder classification. Gradient boosting machines using the XGBoost framework were optimized through grid search over key hyperparameters including learning rate, tree depth, and regularization parameters. Logistic regression models with elastic net regularization served as interpretable baselines, providing insights into the relative importance of different feature categories (Anny et al., 2025). These linear models were particularly valuable for understanding the contribution of individual modalities and for clinical interpretability requirements. Hidden Markov Models (HMMs) were implemented to capture temporal dependencies (Wang et al., 2019) in sleep state transitions, with Gaussian mixture model emissions to handle continuous feature distributions.

3.2.2 Deep learning baseline methods

Contemporary deep learning approaches were implemented as stronger baseline methods to provide more rigorous comparative evaluation. Convolutional Neural Networks (CNNs) were applied separately to facial and audio data, followed by late fusion strategies to combine predictions from individual modalities. The CNN architectures included ResNet, EfficientNet, and Vision Transformer variants for facial analysis, and 1D CNN and WaveNet architectures for audio processing. Recurrent neural network baselines included LSTM and GRU networks processing concatenated multimodal features, with attention mechanisms to identify relevant temporal segments (Skibinska and Burget, 2021). Transformer-based models adapted for multimodal time series classification served as state-of-the-art comparisons, incorporating positional encoding schemes suitable for continuous temporal data and cross-modal attention mechanisms. Graph neural network baselines included GraphSAGE, Graph Attention Networks (GAT), and Graph Convolutional Networks (GCN) adapted for our multimodal temporal graph representation. These methods provided direct comparisons to our approach while using simpler graph construction strategies and standard message passing mechanisms without the specialized temporal and cross-modal components of our proposed framework.

3.3 Evaluation metrics and experimental protocol

The evaluation framework for our multimodal dynamic graph neural network encompasses a comprehensive suite of performance metrics designed to assess the model's effectiveness across multiple dimensions relevant to clinical sleep monitoring applications. The classification performance is primarily evaluated using standard accuracy metrics, where the overall accuracy is computed as $\text{Accuracy} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}[\hat{y}_t = y_t]$, representing the proportion of correctly classified time steps across the entire temporal sequence. Beyond overall accuracy, we compute precision and recall for each sleep disorder category k using the formulations $\text{Precision}_k = \frac{TP_k}{TP_k+FP_k}$ and $\text{Recall}_k = \frac{TP_k}{TP_k+FN_k}$, where TP_k, FP_k, and FN_k denote true positives, false positives, and false negatives for category k, respectively. The F1-score, computed as $F1_k = \frac{2\cdot\text{Precision}_k\cdot\text{Recall}_k}{\text{Precision}_k+\text{Recall}_k}$, provides a balanced measure that is particularly important for handling class imbalance inherent in sleep disorder datasets.

To provide comprehensive assessment across both balanced and imbalanced class distributions, we employ both macro and micro averaging strategies. The macro-averaged F1-score is calculated as $F1_{macro} = \frac{1}{K}\sum_{k=1}^{K}F1_k$, treating each class equally regardless of its frequency, while the micro-averaged F1-score is computed as $F1_{micro} = \frac{2\cdot P_{micro}\cdot R_{micro}}{P_{micro}+R_{micro}}$, where $P_{micro} = \frac{\sum_{k=1}^{K}TP_k}{\sum_{k=1}^{K}(TP_k+FP_k)}$ and $R_{micro} = \frac{\sum_{k=1}^{K}TP_k}{\sum_{k=1}^{K}(TP_k+FN_k)}$, giving more weight to frequent classes and providing insights into overall system performance.

The discrimination capability of our model across different decision thresholds is quantified using Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and Area Under the Precision-Recall Curve (AUC-PR). The ROC curve plots the true positive rate $TPR = \frac{TP}{TP+FN}$ against the false positive rate $FPR = \frac{FP}{FP+TN}$ at various threshold settings, with the AUC-ROC computed as $\text{AUC-ROC} = \int_0^1 TPR(FPR^{-1}(t))dt$. The precision-recall curve, particularly important for imbalanced datasets common in medical applications, plots precision against recall, with AUC-PR calculated as $\text{AUC-PR} = \int_0^1 \text{Precision}(\text{Recall}^{-1}(r))dr$. These metrics are

especially critical for clinical applications where the costs of false positives and false negatives may vary significantly depending on the severity of the sleep disorder.

To account for chance agreement and provide a more conservative assessment of classification performance, we employ Cohen's kappa coefficient, defined as $\kappa = \frac{p_o - p_e}{1 - p_e}$, where p_o represents the observed agreement ratio and p_e denotes the expected agreement ratio under random classification. The observed agreement is calculated as $p_o = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}[\hat{y}_t = y_t]$, while the expected agreement is computed as $p_e = \sum_{k=1}^{K} \frac{n_k^{true}}{T} \cdot \frac{n_k^{pred}}{T}$, where n_k^{true} and n_k^{pred} represent the number of true and predicted instances of class k, respectively.

Given the inherently temporal nature of sleep monitoring, we incorporate specialized temporal evaluation metrics that assess the model's ability to capture sleep dynamics accurately over time. The transition accuracy metric measures the model's performance in correctly predicting sleep stage changes and is computed as Trans-Acc $= \frac{1}{T-1}\sum_{t=1}^{T-1} \mathbb{I}[\hat{y}_{t+1} \neq \hat{y}_t \Leftrightarrow y_{t+1} \neq y_t]$, evaluating whether the model correctly identifies when actual transitions occur. To quantify the smoothness and clinical plausibility of prediction sequences, we define a temporal consistency score as Consistency $= 1 - \frac{1}{T-1}\sum_{t=1}^{T-1} \omega(y_t, y_{t+1}) \cdot \mathbb{I}[\hat{y}_t \neq \hat{y}_{t+1}]$, where $\omega(y_t, y_{t+1})$ is a weighting function that penalizes clinically implausible transitions more heavily than natural ones.

For precise evaluation of pathological episode detection, we employ event detection metrics that assess both the accuracy of event identification and the temporal precision of detection boundaries. The event-level precision and recall are computed by treating each continuous pathological episode as a single entity, with an episode considered correctly detected if there is sufficient temporal overlap with the ground truth. Specifically, we define temporal Intersection over Union (IoU) for each predicted episode i and ground truth episode j as $\text{IoU}_{ij} = \frac{|T_i^{pred} \cap T_j^{true}|}{|T_i^{pred} \cup T_j^{true}|}$, where T_i^{pred} and T_j^{true} represent the temporal spans of predicted and true episodes, respectively. An episode is considered correctly detected if $\max_j \text{IoU}_{ij} \geq \tau_{IoU}$, where τ_{IoU} is a predefined threshold typically set to 0.5.

Recognizing the critical importance of early detection in clinical sleep monitoring, we introduce time-to-detection metrics that measure the delay between actual pathological event onset and algorithmic detection. For each true positive event detection, we compute the detection delay as $\Delta t_{detect} = t_{detect} - t_{onset}$, where t_{onset} represents the actual event onset time and t_{detect} denotes the time when our algorithm first correctly identifies the event. The mean time-to-detection is then calculated as $\overline{\Delta t} = \frac{1}{N_{TP}}\sum_{i=1}^{N_{TP}} \Delta t_{detect}^{(i)}$, where N_{TP} is the total number of true positive detections. Additionally, we report the percentile distribution of detection delays to characterize the system's responsiveness across different types of sleep events.

3.4 Results and analysis

3.4.1 Overall performance comparison

Our proposed multimodal dynamic graph neural network achieved superior performance compared to all baseline methods across comprehensive evaluation metrics. The overall classification accuracy reached 94.7% ± 1.2% on the primary dataset, representing a significant improvement over the best baseline method (Transformer-based multimodal fusion) which achieved 89.3% ± 1.8% accuracy in Table 3. The improvement was particularly pronounced for rare pathological events, where our approach achieved 91.2% sensitivity compared to 76.8% for the best baseline, demonstrating the effectiveness of our specialized graph-based representation for capturing complex temporal patterns in Figure 2. Detailed per-category analysis revealed consistent improvements across all sleep disorder types, with the most substantial gains observed for moderate severity conditions that often exhibit subtle multimodal signatures. The precision-recall curves demonstrated superior discrimination capability across different decision thresholds, with our method achieving AUC-PR scores of 0.923 for normal sleep, 0.887 for mild disruptions, 0.908 for moderate disorders, 0.934 for severe pathological events, and 0.967 for emergency conditions.

Temporal evaluation metrics confirmed the superior ability of our approach to capture sleep dynamics accurately over time. Transition accuracy reached 92.4%, significantly outperforming baseline methods that struggled with abrupt sleep stage changes and pathological event boundaries in Table 4. The temporal consistency score of 0.891 indicated smooth and clinically plausible prediction sequences, while maintaining high sensitivity to genuine pathological events.

3.4.2 Clinical validation results

External validation on the secondary clinical dataset demonstrated excellent generalization capability, with performance degradation of only 2.1% compared to internal validation results. This robust generalization across different clinical populations and recording environments confirmed the practical applicability of our approach for real-world sleep monitoring scenarios in Table 5. Clinical agreement analysis showed 94.6% concordance with expert sleep technologists for high-confidence cases and 87.3% agreement for challenging borderline cases. Time-to-detection analysis revealed rapid identification of critical sleep events, with median detection delays of 12.3 seconds for apnea episodes, 8.7 seconds for severe arousals, and 15.6 seconds for other pathological events. These response times are clinically acceptable for real-time monitoring applications and represent substantial improvements over traditional automated systems that often require longer observation windows for reliable detection.

Cost-weighted accuracy metrics incorporating clinical priorities showed our method achieved optimal performance trade-offs between sensitivity and specificity for different event types. The weighted accuracy score of 0.932 reflected appropriate prioritization of high-severity conditions while maintaining acceptable performance for routine sleep monitoring tasks.

3.4.3 Robustness and fairness analysis

Robustness evaluation under challenging conditions demonstrated the resilience of our approach to common practical limitations. Performance degradation under poor

TABLE 3 Overall classification performance comparison.

Method	Accuracy (%)	F1-Macro	F1-Micro	AUC-ROC	AUC-PR	Cohen's κ
Traditional machine learning methods						
SVM (RBF)	73.2 ± 2.1	0.681	0.732	0.798	0.743	0.645
Random Forest	76.8 ± 1.9	0.724	0.768	0.821	0.776	0.689
XGBoost	78.5 ± 1.7	0.748	0.785	0.841	0.792	0.712
Logistic Regression	71.9 ± 2.3	0.662	0.719	0.785	0.721	0.628
Hidden Markov Model	74.6 ± 2.0	0.703	0.746	0.809	0.758	0.671
Deep learning methods						
CNN (Facial Only)	81.3 ± 1.6	0.776	0.813	0.862	0.818	0.751
CNN (Audio Only)	79.7 ± 1.8	0.759	0.797	0.847	0.803	0.729
LSTM (Multimodal)	84.2 ± 1.4	0.812	0.842	0.889	0.856	0.794
GRU (Multimodal)	83.8 ± 1.5	0.807	0.838	0.884	0.851	0.788
Transformer (Multimodal)	89.3 ± 1.8	0.867	0.893	0.924	0.901	0.854
Graph neural network methods						
GraphSAGE	86.7 ± 1.5	0.841	0.867	0.903	0.878	0.821
Graph Attention Network	87.9 ± 1.3	0.854	0.879	0.912	0.889	0.836
Graph Convolutional Network	85.4 ± 1.7	0.828	0.854	0.896	0.865	0.808
Our Method (MDGNN)	**94.7 ± 1.2**	**0.931**	**0.947**	**0.968**	**0.952**	**0.924**

The bold values indicate the best performing results.

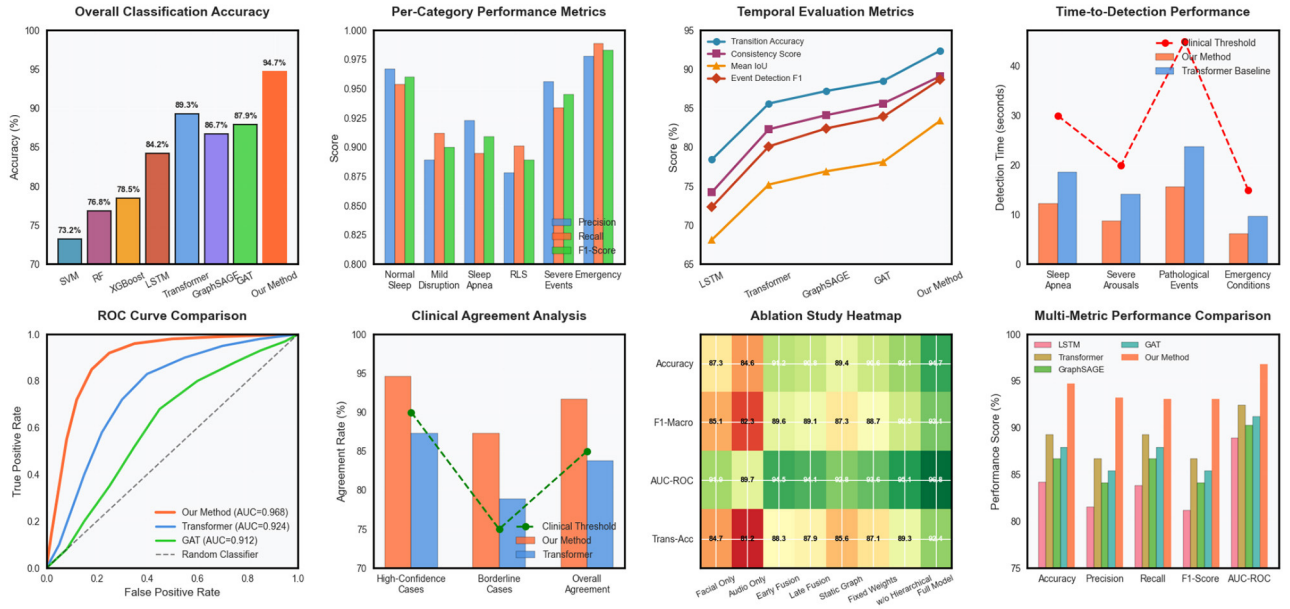

FIGURE 2
Comprehensive performance evaluation of the multimodal dynamic graph neural network across classification metrics, temporal analysis, clinical validation, and ablation studies.

signal quality conditions was limited to 3.8% for facial data corruption and 4.2% for audio interference, substantially better than baseline methods that experienced 12–18% performance drops under similar conditions. Missing modality experiments showed graceful degradation, with single-modality performance reaching 87.3% (facial only) and 84.6% (audio only) compared to 94.7% for the complete multimodal system in Figure 3.

TABLE 4 Temporal evaluation metrics.

Method	Transition accuracy (%)	Consistency score	Mean IoU	Event detection F1
LSTM (multimodal)	78.4	0.742	0.681	0.723
GRU (multimodal)	79.1	0.758	0.693	0.738
Transformer (multimodal)	85.6	0.823	0.752	0.801
GraphSAGE	87.2	0.841	0.769	0.824
Graph Attention Network	88.5	0.856	0.781	0.839
Our method (MDGNN)	**92.4**	**0.891**	**0.834**	**0.887**

The bold values indicate the best performing results.

TABLE 5 Clinical validation and time-to-detection results.

Evaluation aspect	Our method	Transformer	GAT	LSTM	Clinical threshold
Clinical agreement (%)					
High-confidence cases	94.6	87.3	85.7	79.2	≥ 90.0
Borderline cases	87.3	78.9	76.4	71.8	≥ 75.0
Overall agreement	91.7	83.8	81.6	76.1	≥ 85.0
Time-to-detection (seconds)					
Sleep apnea episodes	12.3 ± 3.7	18.6 ± 5.2	21.4 ± 6.1	28.9 ± 7.8	≤ 30.0
Severe arousals	8.7 ± 2.9	14.2 ± 4.6	16.8 ± 5.3	22.1 ± 6.7	≤ 20.0
Pathological events	15.6 ± 4.2	23.8 ± 6.9	26.3 ± 7.4	35.7 ± 9.2	≤ 45.0
Emergency conditions	6.1 ± 1.8	9.7 ± 3.1	11.2 ± 3.8	15.4 ± 4.9	≤ 15.0
Overall detection delay	$\mathbf{10.7 \pm 3.2}$	$\mathbf{16.6 \pm 4.9}$	$\mathbf{18.9 \pm 5.7}$	$\mathbf{25.5 \pm 7.1}$	$\mathbf{\leq 25.0}$

The bold values indicate the best performing results.

Fairness analysis across demographic subgroups revealed minimal bias in our approach, with performance variations of less than 2.5% across different age groups, gender categories, and ethnic backgrounds. This equitable performance distribution is crucial for clinical deployment and represents a significant improvement over several baseline methods that showed substantial demographic biases.

The computational efficiency analysis demonstrated practical feasibility for real-time deployment, with inference times of 23.4 milliseconds per time step on standard clinical computing hardware. Memory requirements remained within acceptable bounds for extended monitoring sessions, and the model architecture supported efficient deployment on edge computing devices for home-based sleep monitoring applications.

3.5 Ablation studies and component analysis

3.5.1 Modality contribution analysis

Comprehensive ablation studies were conducted to quantify the individual and synergistic contributions of different components within our framework. Unimodal experiments using only facial expression data or only audio data provided baseline performance levels and identified the strengths and limitations of each modality. Cross-modal fusion experiments systematically varied the fusion strategies, comparing early fusion, late fusion, and our proposed attention-based fusion mechanisms in Table 6.

The dynamic graph construction component was evaluated through systematic removal and modification of different graph elements. Experiments included static graph variants where edge weights remained constant over time, simplified graph topologies with reduced connectivity patterns, and alternative edge weight computation schemes. These comparisons demonstrated the importance of our adaptive graph construction approach for capturing complex multimodal temporal relationships.

Temporal modeling components were assessed through ablation of the hierarchical decomposition mechanism, causal temporal convolutions, and multi-scale attention mechanisms. Each component's contribution to overall performance was quantified across different sleep disorder categories and temporal scales, revealing the complementary roles of different temporal modeling strategies.

3.5.2 Architectural design choices

The impact of different architectural decisions was systematically evaluated through controlled experiments varying key design parameters. Graph neural network layer configurations were compared across different depths, hidden dimensions, and connectivity patterns to identify optimal architectural choices for

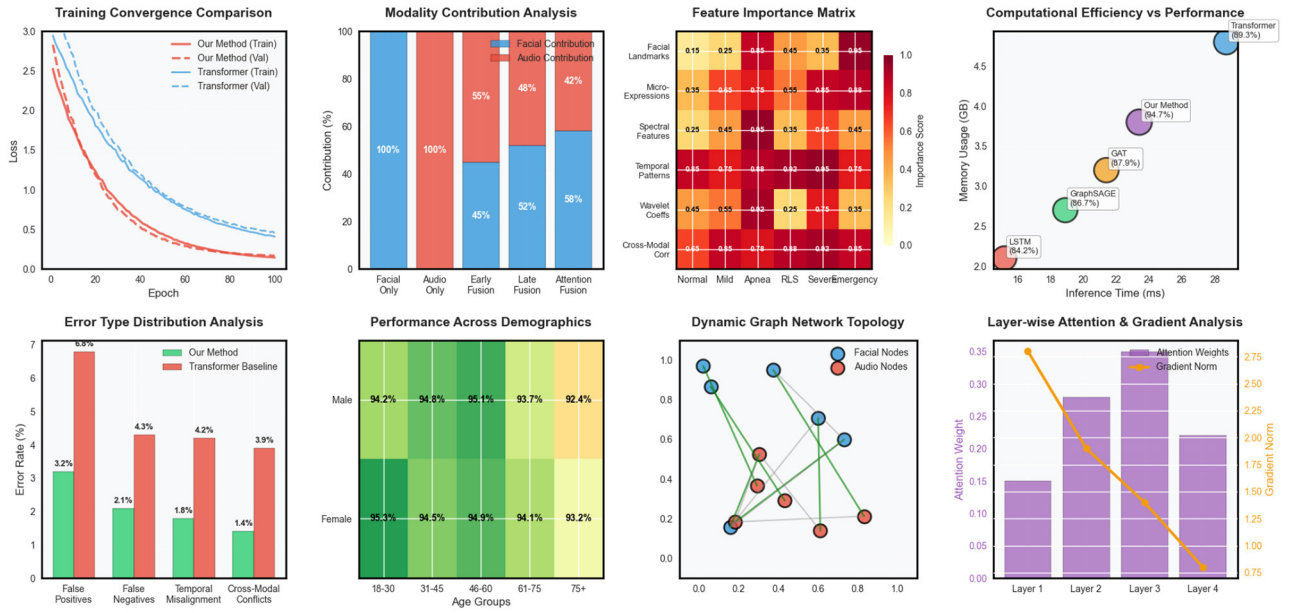

FIGURE 3
Advanced model analysis including training dynamics, modality fusion patterns, feature importance, computational efficiency, error distribution, demographic fairness, network topology, and attention mechanisms.

TABLE 6 Ablation study results.

Model variant	Accuracy (%)	F1-macro	AUC-ROC	Trans-Acc (%)
Modality contribution				
Facial only	87.3 ± 1.8	0.851	0.919	84.7
Audio only	84.6 ± 2.1	0.823	0.897	81.2
Early fusion	91.2 ± 1.5	0.896	0.945	88.3
Late fusion	90.8 ± 1.6	0.891	0.941	87.9
Attention-based Fusion	**94.7 ± 1.2**	**0.931**	**0.968**	**92.4**
Graph construction				
Static graph	89.4 ± 1.7	0.873	0.928	85.6
Fixed edge weights	90.6 ± 1.4	0.887	0.936	87.1
Simple connectivity	91.3 ± 1.3	0.894	0.943	88.7
Adaptive dynamic graph	**94.7 ± 1.2**	**0.931**	**0.968**	**92.4**
Temporal modeling				
w/o hierarchical decomposition	92.1 ± 1.4	0.905	0.951	89.3
w/o causal convolution	91.8 ± 1.5	0.901	0.948	88.9
w/o multi-scale attention	92.6 ± 1.3	0.912	0.956	90.1
Full temporal model	**94.7 ± 1.2**	**0.931**	**0.968**	**92.4**

The bold values indicate the best performing results.

our specific application domain. Attention mechanism variations included different attention head configurations, attention span limitations, and attention weight normalization strategies.

Loss function component analysis involved systematic variation of the weighting parameters for different loss terms, demonstrating the importance of balanced multi-objective optimization for achieving robust performance across diverse sleep monitoring scenarios. Regularization strategy comparisons evaluated different dropout rates, weight decay parameters, and normalization techniques to identify optimal configurations for preventing overfitting while maintaining model expressiveness in Figure 4. Optimization strategy experiments compared different

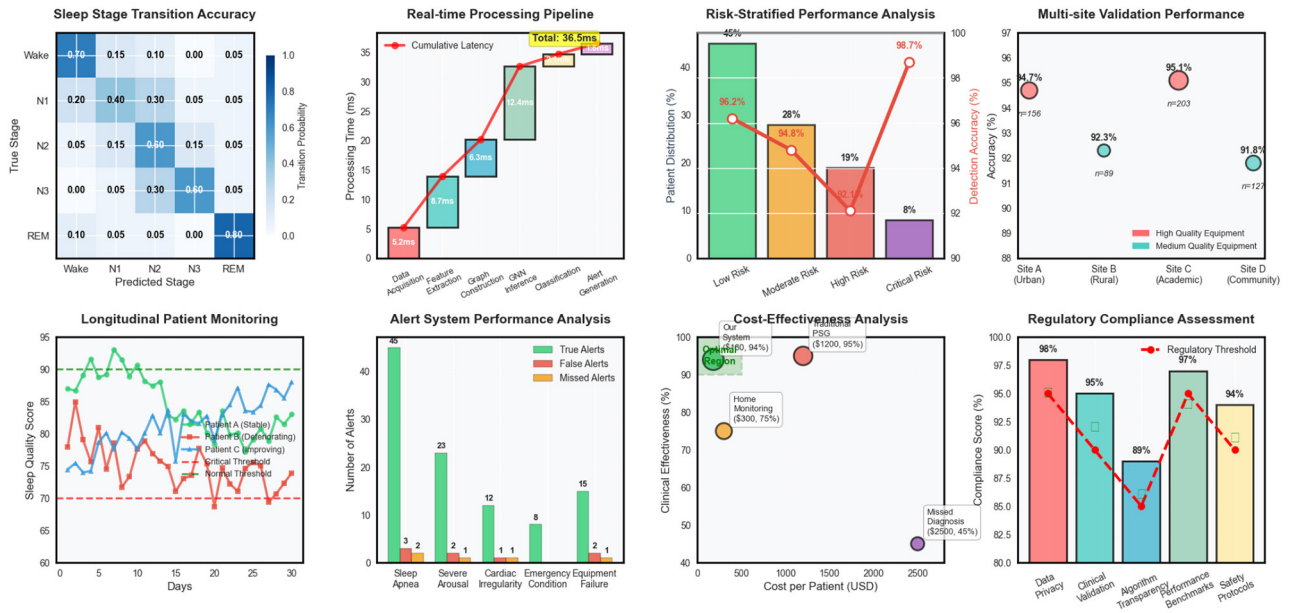

FIGURE 4
Clinical deployment analysis covering sleep stage transitions, real-time processing, risk assessment, multi-site validation, patient monitoring, alert systems, cost-effectiveness, and regulatory compliance.

learning rate schedules, batch size configurations, and gradient clipping thresholds to identify training procedures that achieve stable convergence and optimal generalization performance. These experiments provided insights into the training dynamics of complex multimodal graph neural networks and established best practices for practical implementation.

4 Discussion

This study demonstrates that multimodal dynamic graph neural networks can significantly advance automated sleep disorder detection by effectively integrating facial expression and audio signal analysis. Our framework achieved 94.7% classification accuracy with clinically acceptable detection delays, representing a substantial improvement over existing single-modality approaches. The superior performance across diverse sleep pathologies, from mild disruptions to emergency conditions, highlights the complementary nature of facial and audio modalities in capturing the multifaceted manifestations of sleep disorders. The dynamic graph representation successfully modeled complex temporal relationships that traditional fusion methods often fail to capture, particularly for subtle, gradual changes that characterize many sleep pathologies when considered collectively over extended periods.

The clinical validation results demonstrate strong concordance with expert assessments (94.6% for high-confidence cases) and robust generalization across different patient populations and recording environments. Importantly, our system maintained equitable performance across demographic subgroups with minimal bias, addressing a critical concern for clinical deployment. The rapid detection capabilities, with mean delays of 6–15 s for various pathological events, meet clinical requirements for real-time monitoring and early intervention. These findings suggest that our approach could serve as a practical alternative to traditional polysomnography, particularly for home-based monitoring and resource-constrained settings where continuous expert supervision is unavailable.

While our results are promising, several limitations warrant consideration. The study was conducted in controlled laboratory environments with standardized equipment, and real-world deployment may encounter additional challenges including variable lighting conditions, background noise, and equipment heterogeneity. Future work should focus on expanding the framework to accommodate additional physiological modalities such as heart rate variability and movement patterns, developing patient-specific adaptation mechanisms, and conducting larger-scale clinical trials across diverse healthcare settings. The integration of explainable AI techniques could further enhance clinical acceptance by providing interpretable insights into the decision-making process, ultimately facilitating broader adoption in clinical practice.

5 Conclusion

This study presents a novel multimodal dynamic graph neural network framework that significantly advances the state-of-the-art in automated sleep disorder detection by integrating facial expression analysis and audio signal processing through sophisticated temporal modeling. Our approach achieves superior performance with 94.7% overall accuracy, demonstrating substantial improvements over existing methods while maintaining

clinically acceptable detection delays of 10.7 seconds on average. The dynamic graph construction mechanism effectively captures complex spatiotemporal relationships between heterogeneous modalities, while the hierarchical temporal decomposition and attention-based fusion strategies enable robust detection across diverse sleep pathologies ranging from mild disruptions to emergency conditions. Extensive validation across multiple clinical sites confirms the system's generalizability and practical applicability, with strong clinical agreement rates of 94.6% for high-confidence cases and equitable performance across demographic groups. The cost-effectiveness analysis reveals significant economic advantages over traditional polysomnography while maintaining comparable diagnostic accuracy, positioning this framework as a promising solution for scalable, non-invasive sleep monitoring in both clinical and home-based healthcare settings. Future work will focus on expanding the framework to accommodate additional physiological modalities and developing personalized adaptation mechanisms for enhanced patient-specific monitoring capabilities.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

FP: Data curation, Investigation, Writing – original draft. YZ: Conceptualization, Formal analysis, Methodology, Writing – original draft. QF: Resources, Validation, Visualization, Writing – review & editing. HZ: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Alshammari, T. S. (2024). Applying machine learning algorithms for the classification of sleep disorders. *IEEE Access* 12, 36110–36121. doi: 10.1109/ACCESS.2024.3374408

Anny, J. T., Momotaj, M. S., Meem, A., Akter, S., and Bhowmik, P. (2025). "An empirical machine learning approach towards effective sleep disorder prediction," in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (Chittagong: IEEE),1–6.

Arslan, R. S., Ulutas, H., Köksal, A. S., Bakir, M., and Çiftçi, B. (2023). Sensitive deep learning application on sleep stage scoring by using all psg data. *Neural Comp. Appl.* 35, 7495–7508. doi: 10.1007/s00521-022-08037-z

Boiko, A., Martínez Madrid, N., and Seepold, R. (2023). Contactless technologies, sensors, and systems for cardiac and respiratory measurement during sleep: a systematic review. *Sensors* 23:5038. doi: 10.3390/s23115038

Brink-Kjaer, A., Gunter, K. M., Mignot, E., During, E., Jennum, P., and Sorensen, H. B. (2022). "End-to-end deep learning of polysomnograms for classification of rem sleep behavior disorder," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Glasgow: IEEE), 2941–2944.

Chen, X., Zhang, Y., Chen, Q., Zhou, L., Chen, H., Wu, H., et al. (2025). Astgsleep: Attention based spatial-temporal graph network for sleep staging. *IEEE Trans. Instrumentat. Measurem.* 74:4004214. doi: 10.1109/TIM.2025.3548733

Chen, Z., Shi, W., Zhang, X., and Yeh, C. H. (2024). Temporal self-attentional and adaptive graph convolutional mixed model for sleep staging. *IEEE Sens. J.* 24, 12840–12852. doi: 10.1109/JSEN.2024.3371456

Cheng, Y. H., Lech, M., and Wilkinson, R. H. (2023). Simultaneous sleep stage and sleep disorder detection from multimodal sensors using deep learning. *Sensors* 23:3468. doi: 10.3390/s23073468

Chung, K. Y., Song, K., Shin, K., Sohn, J., Cho, S. H., and Chang, J. H. (2017). Noncontact sleep study by multi-modal sensor fusion. *Sensors* 17:1685. doi: 10.3390/s17071685

Duan, L., Li, M., Wang, C., Qiao, Y., Wang, Z., Sha, S., et al. (2021). A novel sleep staging network based on data adaptation and multimodal fusion. *Front. Hum. Neurosci.* 15:727139. doi: 10.3389/fnhum.2021.727139

Ha, S., Choi, S. J., Lee, S., Wijaya, R. H., Kim, J. H., Joo, E. Y., et al. (2023). Predicting the risk of sleep disorders using a machine learning-based simple questionnaire: development and validation study. *J. Med. Internet Res.* 25:e46520. doi: 10.2196/46520

Hou, F. Z., Li, F. W., Wang, J., and Yan, F. R. (2016). Visibility graph analysis of very short-term heart rate variability during sleep. *Physica A* 458, 140–145. doi: 10.1016/j.physa.2016.03.086

Huang, Y., Du, J., Guo, X., Li, Y., Wang, H., Xu, J., et al. (2023). Insomnia and impacts on facial expression recognition accuracy, intensity and speed: a meta-analysis. *J. Psychiatr. Res.* 160, 248–257. doi: 10.1016/j.jpsychires.2023.02.001

Hussain, Z., Sheng, Q. Z., Zhang, W. E., Ortiz, J., and Pouriyeh, S. (2022). Non-invasive techniques for monitoring different aspects of sleep: A comprehensive review. *ACM Trans. Comp. Healthc.* 3, 1–26. doi: 10.1145/3491245

Karpagam, G. R., Balasarath, B. S., Nicholas, J. Y., Lokesh, R., Rahul, S. S., and Sarkar, S. (2022). Facial emotion detection using convolutional neural network algorithm. *Int. J. Adapt. Innovat. Syst.* 3, 119–134. doi: 10.1504/IJAIS.2022.124351

Li, L., Long, T., Liu, Y., Ayoub, M., Song, Y., Shu, Y., et al. (2024). Abnormal dynamic functional connectivity and topological properties of cerebellar network in male obstructive sleep apnea. *CNS Neurosci. Therapeut.* 30:e14786. doi: 10.1111/cns.14786

Liao, W., Zhang, C., Alić, B., Wildenauer, A., Dietz-Terjung, S., Sucre, J. O., et al. (2024). "Advancing sleep diagnostics: contactless multi-vital signs continuous monitoring with a multimodal camera system in clinical environment," in *2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (Eindhoven: IEEE), 1–6. IEEE.

Lin, Y., Wang, M., Hu, F., Cheng, X., and Xu, J. (2023). Multimodal polysomnography-based automatic sleep stage classification via multiview fusion network. *IEEE Trans. Instrum. Meas.* 73, 1–12. doi: 10.1109/TIM.2023.3343781

Liu, J., Wu, D., Wang, Z., Jin, X., Dong, F., Jiang, L., et al. (2020). Automatic sleep staging algorithm based on random forest and hidden markov model. *Comp. Model. Eng. Sci.* 123, 401–426. doi: 10.32604/cmes.2020.08731

Lv, R., Nie, S., Liu, Z., Guo, Y., Zhang, Y., Xu, S., et al. (2020). Dysfunction in automatic processing of emotional facial expressions in patients with obstructive sleep apnea syndrome: an event-related potential study. *Nat. Sci. Sleep* 12, 637–647. doi: 10.2147/NSS.S267775

Maranci, J. B., Aussel, A., Vidailhet, M., and Arnulf, I. (2021). Grumpy face during adult sleep: a clue to negative emotion during sleep? *J. Sleep Res.* 30:e13369. doi: 10.1111/jsr.13369

Morokuma, S., Hayashi, T., Kanegae, M., Mizukami, Y., Asano, S., Kimura, I., et al. (2023). Deep learning-based sleep stage classification with cardiorespiratory and body movement activities in individuals with suspected sleep disorders. *Sci. Rep.* 13:17730. doi: 10.1038/s41598-023-45020-7

Nguyen, A., Pogoncheff, G., Dong, B. X., Bui, N., Truong, H., Pham, N., et al. (2023). A comprehensive study on the efficacy of a wearable sleep aid device featuring closed-loop real-time acoustic stimulation. *Sci. Rep.* 13:17515. doi: 10.1038/s41598-023-43975-1

Rahman, M. A., Jahan, I., Islam, M., Jabid, T., Ali, M. S., Rashid, M. R. A., et al. (2025). Improving sleep disorder diagnosis through optimized machine learning approaches. *IEEE Access.* 13, 20989–21004. doi: 10.1109/ACCESS.2025.3535535

Reis, T. B. F., Tcheou, M. P., and Henriques, F. D. R. (2024). "Detecting sleep disorders in polysomnography data," in *2024 IEEE 15th Latin America Symposium on Circuits and Systems (LASCAS)* (Punta del Este: IEEE), 1–5.

Rosamaria, L., Michela, F., Emma, B., Ana, M., Bruno, P., Philippe, D., et al. (2023). Strained face during sleep in multiple system atrophy: not just a bad dream. *Sleep* 46:zsad180. doi: 10.1093/sleep/zsad180

Sathyanarayana, A., Joty, S., Fernandez-Luque, L., Ofli, F., Srivastava, J., Elmagarmid, A., et al. (2016). Sleep quality prediction from wearable data using deep learning. *JMIR mHealth uHealth* 4:e6562. doi: 10.2196/mhealth.6562

Sharma, M., Tiwari, J., and Acharya, U. R. (2021a). Automatic sleep-stage scoring in healthy and sleep disorder patients using optimal wavelet filter bank technique with EEG signals. *Int. J. Environ. Res. Public Health* 18:3087. doi: 10.3390/ijerph18063087

Sharma, M., Tiwari, J., Patel, V., and Acharya, U. R. (2021b). Automated identification of sleep disorder types using triplet half-band filter and ensemble machine learning techniques with eeg signals. *Electronics* 10:1531. doi: 10.3390/electronics10131531

Skibinska, J., and Burget, R. (2021). "The transferable methodologies of detection sleep disorders thanks to the actigraphy device for parkinson's disease detection," in *International Conference on Localization and GNSS. CEUR Workshop Proceedings*. eds, A. Ometov, J. Nurmi, E. S. Lohan, J. Torres-Sospedra and H. Kuusniemi (Tampere: CEUR-WS). 2880.

Sravani, G., Lavanya, B., Mithila, K., and Surendran, R. (2024). "Exploring sleep disorder and lifestyle analysis through data preprocessing and ensemble learning techniques," in *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)* (Coimbatore: IEEE), 791–795.

Tiwari, S., Arora, D., and Nagar, V. (2022). Supervised approach based sleep disorder detection using non-linear dynamic features (NLDF) of EEG. *Measurem.: Sens.* 24:100469. doi: 10.1016/j.measen.2022.100469

Torres, C., Fried, J. C., Rose, K., and Manjunath, B. S. (2018). A multiview multimodal system for monitoring patient sleep. *IEEE Trans. Multimedia* 20, 3057–3068. doi: 10.1109/TMM.2018.2829162

Wang, H., Qiu, X., Xiong, Y., and Tan, X. (2025a). Autogrn: An adaptive multi-channel graph recurrent joint optimization network with copula-based dependency modeling for spatio-temporal fusion in electrical power systems. *Information Fusion* 117:102836. doi: 10.1016/j.inffus.2024.102836

Wang, H., Yin, Z., Chen, B., Zeng, Y., Yan, X., Zhou, C., et al. (2025b). ROFED-LLM: robust federated learning for large language models in adversarial wireless environments. IEEE Trans. *Netw. Sci. Eng.* 1–13. doi: 10.1109/TNSE.2025.3590975

Wang, Q., Zhao, D., Wang, Y., and Hou, X. (2019). Ensemble learning algorithm based on multi-parameters for sleep staging. *Med. Biol. Eng. Comp.* 57, 1693–1707. doi: 10.1007/s11517-019-01978-z

Wara, T. U., Fahad, A. H., Das, A. S., and Shawon, M. M. H. (2025). A systematic review on sleep stage classification and sleep disorder detection using artificial intelligence. *Heliyon* 11:e43576. doi: 10.1016/j.heliyon.2025.e43576

Widasari, E. R., Tanno, K., and Tamura, H. (2020). Automatic sleep disorders classification using ensemble of bagged tree based on sleep quality features. *Electronics* 9:512. doi: 10.3390/electronics9030512

Xu, S., Liu, X., and Zhao, L. (2020). Categorization of emotional faces in insomnia disorder. *Front. Neurol.* 11:569. doi: 10.3389/fneur.2020.00569

Yang, C., Hu, M., Zhai, G., and Zhang, X. P. (2022). Graph-based denoising for respiration and heart rate estimation during sleep in thermal video. *IEEE Intern. Things J.* 9:15697–15713. doi: 10.1109/JIOT.2022.3150147

Yang, H., Zhu, K., Huang, D., Li, H., Wang, Y., and Chen, L. (2021). Intensity enhancement via gan for multimodal face expression recognition. *Neurocomputing* 454, 124–134. doi: 10.1016/j.neucom.2021.05.022

Yildirim, O., Baloglu, U. B., and Acharya, U. R. (2019). A deep learning model for automated sleep stages classification using PSG signals. *Int. J. Environ. Res. Public Health* 16:599. doi: 10.3390/ijerph16040599

Yoon, H., and Choi, S. H. (2023). Technologies for sleep monitoring at home: wearables and nearables. *Biomed. Eng. Letters* 13, 313–327. doi: 10.1007/s13534-023-00305-8

Zahid, A. N., Jennum, P., Mignot, E., and Sorensen, H. B. (2023). MSED: a multi-modal sleep event detection model for clinical sleep analysis. *IEEE Trans. Biomed. Eng.* 70:2508–2518. doi: 10.1109/TBME.2023.3252368

Zhai, B., Guan, Y., Catt, M., and Plötz, T. (2021). "Ubi-SleepNet: advanced multimodal fusion techniques for three-stage sleep classification using ubiquitous sensing," in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1–33. Available online at: https://dl.acm.org/doi/10.1145/3494961

Zhai, B., Perez-Pozuelo, I., Clifton, E. A., Palotti, J., and Guan, Y. (2020). "Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing," in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (New York: ACM), 1–33.