Information Fusion 105 (2024) 102224

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Multimodal graph learning based on 3D Haar semi-tight framelet for student engagement prediction

Ming Li [a], Xiaosheng Zhuang [b], Lu Bai [c,d], Weiping Ding [e,*]

[a] Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, China
[b] Department of Mathematics, City University of Hong Kong, Hong Kong, China
[c] School of Artificial Intelligence, Beijing Normal University, Beijing, China
[d] Central University of Finance and Economics, Beijing, China
[e] School of Information Science and Technology, Nantong University, Nantong, China

ARTICLE INFO

Keywords:
Multimodal learning with graphs
Graph structure learning
3D haar semi-tight framelet
Student engagement prediction

ABSTRACT

With the increasing availability of multimodal educational data, there is a growing need to effectively integrate and exploit multiple data sources to enhance student engagement prediction accuracy. In this work, we propose a framework that combines multimodal data, including visual, textual and acoustic modalities that reflect the students' personalities, their demographic information, their learning behavior and attention, with graph learning techniques. Specifically, 3D Haar semi-tight framelet transforms are developed to capture the inter-modal relationships and model the complex interactions within the multimodal data. Subsequently, we introduce a novel module for adaptive graph structure learning based on the spectrum of multimodal data, which takes into consideration the distinct contributions of low-pass and high-pass framelet coefficients by adaptively weighing their impact. By addressing a standard semi-supervised node classification problem, we successfully achieve the objective of student engagement prediction. The experiment evaluations on a real-world educational dataset demonstrate the effectiveness of the proposed approach, achieving superior performance compared to state-of-the-art methods. Our experimental studies demonstrate the importance of multimodal graph learning in accurately predicting student engagement and its potential to enhance educational outcomes.

1. Introduction

Student engagement prediction has emerged as a crucial area of research within the field of educational data mining and learning analytics [1–3]. With the increasing availability of digital educational platforms and online learning environments, accurately predicting and understanding student engagement levels has become essential for improving educational outcomes and delivering personalized interventions [4–6]. By leveraging diverse data sources, including academic performance, online behavior, social interactions, and physiological signals, researchers have employed various machine learning and data mining techniques to develop predictive models [7,8]. To further enhance the accuracy and effectiveness of student engagement prediction, it is imperative to incorporate multimodal-based models that leverage multiple data sources, such as visual, textual, and acoustic modalities, reflecting their personality, demographic information, learning behavior and attention, to capture a holistic view of students' engagement patterns and facilitate a more comprehensive understanding of their learning experiences [9–11].

Multimodal graph learning is an emerging field that explores the fusion of multiple modalities in the context of graph data [12], receiving extensive attention in various domains such as disease prediction [13–17], recommender systems [18–20], time-series anomaly detection [21], sentiment analysis [22], computer vision (CV) [23–25] and natural language processing (NLP) [26–28]. It aims to leverage the rich information present in different modalities to enhance graph analysis, representation learning, and downstream tasks [29]. This interdisciplinary area combines techniques from graph mining, machine learning, computer vision, natural language processing, and signal processing to tackle the challenges associated with integrating and exploiting diverse data sources. By capturing inter-modal relationships, multimodal graph learning enables a more comprehensive

* Corresponding author.
E-mail addresses: mingli@zjnu.edu.cn (M. Li), xzhuang7@cityu.edu.hk (X. Zhuang), bailu@bnu.edu.cn (L. Bai), dwp9988@163.com (W. Ding).

https://doi.org/10.1016/j.inffus.2024.102224
Received 27 June 2023; Received in revised form 7 December 2023; Accepted 2 January 2024
Available online 5 January 2024

understanding of complex systems, such as social networks, multimedia data, biological networks, and sensor networks. It encompasses various research directions, including multimodal fusion, graph-based modeling, transfer learning, attention mechanisms, and deep learning architectures tailored for multimodal graph data. Despite the promising advancements in multimodal graph learning, the focus on its application in student engagement prediction remains limited. Therefore, there exists a significant opportunity for further exploration and research to harness the benefits of multimodal graph learning to improve student engagement prediction models, as one of the primary goals of this work.

In this study, within the context of student engagement prediction, we propose a novel multimodal graph learning framework by developing a 3D Haar semi-tight framelet (`3D-HaarFrame`) transform. The proposed `3D-HaarFrame` facilitates an efficient representation of multimodal data by decomposing its tensor form into a set of coefficients at different scales and directions. This decomposition allows for the extraction of relevant features and patterns from the unified multimodal data while achieving data compression and reducing redundancy. Additionally, `3D-HaarFrame` enables multiscale analysis that mines information at different frequency bands and effectively promotes the exploration of both global trends and fine-grained details in multimodal data. Moreover, based on `3D-HaarFrame`, we propose a spectrum-based graph structure learning module, in which the contribution of low-pass and high-pass framelet coefficients is adaptively adjusted, to learn the inter-modal relationships and complex interactions within the spectral-aware embeddings. By leveraging these components, our proposed framework enhances the accuracy of student engagement prediction, as verified convincingly based on extensive experimental studies on a real-world multimodal educational dataset, compared with several baselines.

In summary, as our main technical contributions, the proposed framework of the 3D Haar semi-tight framelet offers several advantages in the context of multimodal graph learning and student engagement prediction:

- **Effective representation**: `3D-HaarFrame` provides an efficient representation of multimodal data by decomposing it into a set of coefficients with different scales and orientations. This decomposition allows for the extraction of relevant features and patterns from the data while achieving data compression and reducing redundancy. It enables a compact representation of the multimodal data, facilitating efficient processing and analysis.
- **Multiscale analysis**: `3D-HaarFrame` supports the multiscale analysis of multimodal data by capturing information at different scales[1] Its associated transforms decompose the data into framelet coefficients at different scales. By analyzing the framelet coefficients at different scales, it is possible to identify and extract key information from the data. The multiscale analysis is particularly beneficial for understanding and modeling complex relationships and interactions within the data, enabling a comprehensive exploration of student engagement patterns.
- **Inter-modal relationship mining**: `3D-HaarFrame` allows for the capture of inter-modal relationships within multimodal data. By decomposing the data into different frequency bands, it enables the exploration of correlations and dependencies between modalities at different scales. This ability to analyze inter-modal relationships is crucial for understanding the complex interactions between different data sources and leveraging the complementary information provided by each modality.
- **Robustness to noise**: `3D-HaarFrame` has inherent noise-robust properties due to its sparsity-promoting nature. It effectively suppresses noise and irrelevant information by concentrating energy in a small number of coefficients, allowing for more reliable and robust analysis of the multimodal data. This robustness is particularly advantageous when dealing with noisy or incomplete data commonly encountered in real-world educational settings.

The remainder of this paper is organized as follows: Section 2 provides a concise overview of the relevant works pertaining to multimodal learning with graphs and student engagement prediction. In Section 3, we present a comprehensive exposition of our proposed methodology, commencing with a detailed explanation of the 3D Haar Semi-Tight Framelet employed in the study. Subsequently, we delve into the graph structure learning techniques leveraged to effectively capture inter-modal relationships and intricately model the complex interactions intrinsic to the data. In Section 4, we detail the experiments that were conducted to verify the effectiveness of our proposed method. Finally, we conclude this paper and discuss future work in Section 5.

2. Related works

2.1. Multimodal learning with graphs

Multimodal learning with graphs (MLG) has witnessed significant progress in recent years since researchers have increasingly recognized the power of incorporating multiple modalities (text/video/images/audio, etc.) with graph information into a unified framework that leverages the rich information present in each modality and their possible relationship. In relation to its application, MLG receives extensive attention in the area of recommender systems, disease prediction, NLP and CV. For example, MMGCN [18] introduces a class of multimodal graph neural network recommendation models that address these issues. Their model enhances the representation learning process by creating fine-grained bipartite graphs for each modality based on the original user-item bipartite graph. Subsequently, modality-specific graph representation learning is performed on each bipartite graph, followed by the fusion of structured information, self-information, and inter-modality information through a joint layer. Motivated by MMGCN, MGAT [19] further improves the model by introducing a recommendation model based on a multimodal graph attention network. MKGAT [20] develops a multimodal knowledge graph attention model-driven recommendation system, utilizing a multimodal knowledge graph. In the context of medical big data analysis, some scholars have explored the application of GNNs for modeling multimodal medical data and have made efforts to develop multimodal learning with graphs for disease prediction. Holzinger et al. [29] highlight the significant role of graph neural networks in facilitating multi-modal causability by enabling the direct definition of causal connections between features through graph structures. MGNN [31] constructs bipartite graphs connecting patients with different modal pathological data, such as gene expression and copy number variation. They employed independent graph neural networks for representation learning under each modality and predicted the survival rate of cancer patients by incorporating clinical data representations. In addition, MLG enables more comprehensive representations, improved feature extraction, semantic relationship modeling, cross-modal fusion, and interpretability, leading to enhanced performance in tasks within the CV and NLP domains [12,32]. For instance, Saqur et al. [23] proposed the use of neurosymbolic graphs to capture the close relationships between the hidden concepts of different modalities in visual question answering tasks. MM-GNN [33] introduces a visual question answering model based on multimodal graph neural networks. They utilized three different subgraphs, representing visual, semantic, and numerical information, to capture the diverse aspects of a given image. By employing three types of graph network aggregators, they facilitated the exchange of information between different modalities and

[1] Multiscale analysis refers to the decomposition of signals, images, 3D data, or even high-dimensional data into a set of components, each capturing different levels of detail or information. Wavelets/framelets are commonly used in multiscale analysis to achieve such a purpose [30].

updated vertex representations. Mafla et al. [34] propose a GCN-based multimodal reasoning graph model for fine-grained image classification and retrieval tasks. In the field of social media analysis, MGCN [24] constructs graph data from textual content and visual conveyance content in news reports.

Within the domain of machine learning, both multimodal learning and multiview learning stand as key concepts. Although they are defined differently, their underlying similarities are evident. Specifically, data from various "modalities" can be interpreted as different "views" of the same underlying phenomenon.[2] In a broader sense, multiview learning which focuses on the challenge of learning from data characterized by multiple, distinct feature sets, has received significant attention in recent years. For example, [35] introduces a multi-view graph learning method that uses adaptive label propagation tailored for semi-supervised classification. This method integrates techniques like latent factor extraction, graph sparsification, and label propagation into a unified framework. [36] introduces a new graph convolution framework for anomaly detection in multiview-attributed networks. On the other hand, there is a growing trend to utilize graph-based techniques for multiview clustering. For instance, [37] employs a sparse graph learning technique to derive a consistent, sparsely structured similarity matrix from numerous views, which subsequently aids in multiview spectral clustering. Further contributions in [38,39] outline graph learning-centered multiview clustering strategies, which not only are able to construct a pivotal similarity graph within a spectral embedding domain (as opposed to the conventional feature space), they also facilitate clustering through the concurrent learning of spectral embedding matrices and low-rank tensor representation.

2.2. Student engagement prediction

Student (or learner) engagement prediction is a challenging undertaking [4], compounded by the lack of unanimous consensus on the precise definition of student engagement [5]. Instead, there exist various interpretations of this concept. Nevertheless, amidst these divergent definitions, it is commonly accepted that engagement is a multifaceted overarching construct encompassing behavioral, emotional, cognitive, and agentic dimensions [1–3]. The availability of multimodal corpora for student engagement prediction remains limited [6], impeding progress in the development of automated engagement predictive models that could assist teachers and tutors in obtaining a more accurate assessment of student engagement during courses and/or tutorials. In particular, the authors of [40] use machine learning approaches to assess visible engagement during classroom instruction, with a specific focus on students' attentive behavior. Maimaiti et al. [41] employ an activity theory perspective to comprehensively examine student disengagement in web-based video-conferencing supported online learning environments. In [42], Bayesian networks are applied for modeling student engagement, incorporating contextual factors to refine predictive models. The authors of [43] examine empirically the influence of lecturer-student exchange on student engagement and their propensity to prematurely withdraw from university. Furthermore, Davies et al. [44] conduct a case study to investigate student engagement with simulations, elucidating the impact of interactive learning environments on student involvement. The authors of [45] uncover meaningful classifications of student types and previously unclear patterns of student engagement based on the understanding of students' learning behavior in online courses by exploring alternative learning analytic approaches and visual representations. It should be noted that, however, these existing works predominantly operate within unimodal frameworks, only focusing on specific data modality (visual or textual). By solely relying on single-modal information, these approaches may overlook valuable insights that could be derived from multimodal data sources, leading to potential limitations in the accuracy and robustness of engagement prediction models. Sümer et al. [46] conduct a multimodal engagement analysis using existing computer vision methods to classify engagement from facial videos (i.e., audio and visual recordings of secondary school classes). However, their framework does not consider the relationship between different modalities. Technically, student engagement prediction using multimodal learning with graphs offers distinct advantages: (i) it enables the integration of diverse data modalities, such as visual, textual, and acoustic modality, allowing for a comprehensive representation of student engagement; (ii) it leverages the interconnections and interdependencies among modalities through graph structures that help capture complex relationships and interactions, contributing to accuracy improvement in student engagement prediction.

Summary. In our work, the use of graph-based methods is motivated by the desire to fully capture the underlying relationships among students as well as the complex correlation between different modalities. As aforementioned, numerous studies have delved into multimodal learning with graphs, albeit for varied tasks and contexts. These studies essentially share a similar rationale for adopting graph-centric approaches. To the best of our knowledge, our research is the first effort that employs a multimodal graph learning approach specifically for predicting student engagement. Within our framework, graph-based techniques are pivotal in understanding and leveraging the ties and interdependencies between modalities and in uncovering latent relationships/interactions among students. By integrating these elements, rather than solely focusing on a single modality without the insights offered by graphs, we achieve a better learning representation for students' learning behavior. This enhanced representation, as our experiments confirm (see Section 4), markedly improves the accuracy of student engagement predictions.

3. Methods

3.1. Notation

In this section, we provide a concise overview of the notations commonly used throughout the content. Let $\mathbb{R}$ represent the set of all real numbers, $\mathbb{Z}$ represent the set of integers, $\mathbb{C}$ represent the set of all complex numbers, respectively. For a number $a \in \mathbb{C}$, $\bar{a}$ denotes its complex conjugate. Let $\mathbf{X} = \left[x_1, x_2, \ldots, x_N\right]$ represent the raw multi-modal features of N students and $\mathbf{Y} = \left[y_1, y_2, \ldots, y_N\right]$ denote the corresponding labels. For each student i with M modalities, we define $x_i = Concat(x_i^1, x_i^2, \ldots, x_i^M)$ as the concatenation of $x_i^1, x_i^2, \ldots, x_i^M$, where $x_i^m \in \mathbb{R}^{d_m}$ represents the mth modality of the student i. Let $\mathbf{X}^m = \left[x_1^m, x_2^m, \ldots, x_N^m\right] \in \mathbb{R}^{d_m \times N}$ represent the features of the mth modality. Then, a student-based population graph $G = (V, E, \mathbf{X})$ can be constructed for (node-level) engagement prediction, where the node set $V = \{v_i\}_{i=1}^N$ represents the set of students and the edges in $E = \{e_{ij} = (v_i, v_j)_{i,j=1}^N\}$ stand for the connections between each pair of students (nodes). Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ denote the adjacency matrix, in which $A_{ij} \in \mathbf{A}$ represents the edge weight of $e_{ij} \in E$. Generally, we use $\boldsymbol{\mathcal{X}}$ to represent a 3D-tensor. Specifically, we define similar notations such as $\boldsymbol{\mathcal{X}}_{\cdot}$ and $\tilde{\boldsymbol{\mathcal{X}}}$ in subsequent sections where they are utilized with special concerns. In addition, $Sim(\cdot, \cdot)$ represents the similarity function between two vectors, while $Vec(\cdot)$ signifies the vectorization operation.

3.2. Overall framework

In this part, we overview the proposed multimodal graph learning framework based on `3D-HaarFrame` for student engagement prediction. As shown in Fig. 1, the framework comprises four key modules 1

[2] In this work, we narrow our focus to a specific facet of the concept 'multimodal'.

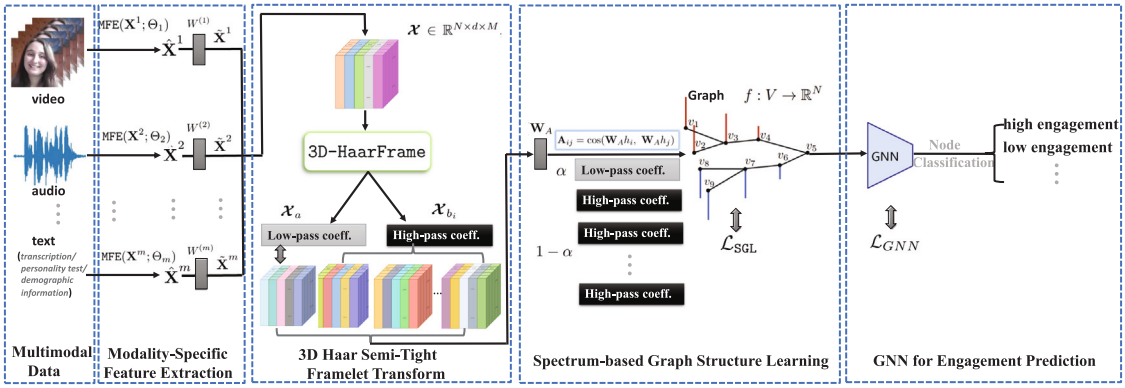

Fig. 1. Schematic of the proposed framework.

- Modality-specific feature extraction (MFE). Given $\mathbf{X}$ and $\mathbf{X}^m$ defined in Section 3.1, for each modality (visual/textual/acoustic etc.), a specified tool/model is used to conduct feature extraction, leading to $\hat{\mathbf{X}}^m \in \mathbb{R}^{\hat{d}_m \times N}$. To form a unified 3-D tensor, we transform each modal feature $\{\hat{\mathbf{X}}^m\}_{m=1}^M$ to $\{\tilde{\mathbf{X}}^m\}_{m=1}^M$ with the identical dimension d through the transform matrices $\{W^{(m)} \in \mathbb{R}^{d\times\hat{d}_m}\}_{m=1}^M$.
- 3D-HaarFrame construction. Given a 3D-tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N\times d\times M}$, of which the sth horizontal slice $\boldsymbol{\mathcal{X}}(s,:,:) = [\tilde{x}_s^1, \tilde{x}_s^2, \ldots, \tilde{x}_s^M] \in \mathbb{R}^{d\times M}$ $(s = 1, 2, \ldots, N)$ contains all the extracted multi-modal features for student s, we construct 3D-HaarFrame that efficiently converts $\boldsymbol{\mathcal{X}}$ to a frequency domain with low-pass and high-pass framelet coefficients.
- Spectrum-based graph structure learning (SGL). The objective of graph structure learning is to identify an appropriate adjacency matrix $\mathbf{A}$ that significantly aids in solving the problems of student engagement prediction. By treating the low-pass and high-pass framelet coefficients as signals in the spectral domain, we carefully consider their different contributions by incorporating adaptive weighting mechanisms to assess their impact on the learning graph structure. This enables us to effectively capture the structural characteristics of multimodal data, leading to $G = (V, E, \mathbf{X})$, to facilitate graph representation learning in downstream tasks.
- Model optimization. This module aims at solving a joint optimization problem by simultaneously optimizing the semi-supervised node classification loss and the regularization induced by SGL.

3.3. Modality-specific feature extraction

As an initial step, different feature extraction methods are employed for each modality to obtain initial feature representations, as the refined inputs for the following 3D-HaarFrame module. Technically, there are many choices for modality-specific feature extraction. For example, for visual modality (images/videos), one can use the Openface toolkit [47], the pre-trained MA-Net [48], the pre-trained VGG-Face [49], ResNet-50 [50] or their combination equipped with a complex deep network architecture. For acoustic modality, OpenSmile [51] or the pre-trained wav2vec [52] can be utilized for audio feature extraction. For textural modality, the pre-trained BERT [53] or its follow-up improvements like pre-trained RoBERTa [54] or pre-trained DeBERTa [55] can be used. We note that, if the given data source contains video, we select frame-wise images based on a sliding window, producing a sequence of image-based visual data inputs (see for example the experimental setup detailed in our experiments). Formally, for a given modality, i.e., $\mathbf{X}^m \in \mathbb{R}^{d_m\times N}$, in which m denotes visual/textural/acoustic modality, the goal of MFE lies in finding the initial feature representation $\hat{\mathbf{X}}^m \in \mathbb{R}^{\hat{d}_m\times N}$:

$$\hat{\mathbf{X}}^m := \mathtt{MFE}(\mathbf{X}^m; \boldsymbol{\Theta}_m), \tag{1}$$

where $\boldsymbol{\Theta}_m$ denotes the collection of weights to be fine-tuned for the specific feature extractor.

Then, we transform each modal feature $\{\hat{\mathbf{X}}^m\}_{m=1}^M$ to $\{\tilde{\mathbf{X}}^m\}_{m=1}^M$ with the identical dimension d through the transform matrices $\{W^{(m)} \in \mathbb{R}^{d\times\hat{d}_m}\}_{m=1}^M$, that is,

$$\tilde{\mathbf{X}}^m := W^{(m)}\hat{\mathbf{X}}^m \in \mathbb{R}^{d\times N}, \tag{2}$$

which finally produces a 3D-tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N\times d\times M}$ with the mth frontal slice expressed by $\boldsymbol{\mathcal{X}}(:,:,m) := (\tilde{\mathbf{X}}^m)^{\mathrm{T}}$ $(m = 1, 2, \ldots, M)$ and the sth frontal slice expressed by $\boldsymbol{\mathcal{X}}(s,:,:) := [\tilde{x}_s^1, \tilde{x}_s^2, \ldots, \tilde{x}_s^M] \in \mathbb{R}^{d\times M}$ $(s = 1, 2, \ldots, N)$. So far, the obtained 3D-tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N\times d\times M}$ contains all the necessary components for the construction of the 3D Haar semi-tight framelet, as detailed in the subsequent section.

3.4. Construction of 3D-HaarFrame

Given a 3D-tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N\times d\times M}$, we next detail the 3D-HaarFrame that efficiently converts $\boldsymbol{\mathcal{X}}$ to a frequency domain with low-pass and high-pass framelet coefficients, $\boldsymbol{\mathcal{X}}_a, \boldsymbol{\mathcal{X}}_{b_r}, r = 1, \ldots, 13$, each of which is also a 3D-tensor in $\mathbb{R}^{N\times d\times M}$, where the 3D-HaarFrame is determined by a filter bank $\mathtt{DHF}_3 = \{a, b_1, \ldots, b_{13}\}$ of 3D filters. The filter a is the Haar low-pass filter while the other 13 filters are high-pass filters.

A 3D filter (mask) $\mathcal{F} = \{\mathcal{F}(k)\}_{k\in\mathbb{Z}^3} : \mathbb{Z}^3 \to \mathbb{R}$ is a sequence of filter taps (real/complex numbers) on $\mathbb{Z}^3$. Using $\boldsymbol{\delta}$, we denote *the Dirac sequence* such that $\boldsymbol{\delta}(0) = 1$ and $\boldsymbol{\delta}(k) = 0$ for all $k = (k_1, k_2, k_3) \in \mathbb{Z}^3\backslash\{0\}$. For $\gamma = (\gamma_1, \gamma_2, \gamma_3) \in \mathbb{Z}^3$, we also use the notation $\boldsymbol{\delta}_\gamma$ to stand for the sequence $\boldsymbol{\delta}(\cdot - \gamma)$, i.e., $\boldsymbol{\delta}_\gamma(\gamma) = 1$ and $\boldsymbol{\delta}_\gamma(k) = 0$ for all $k \in \mathbb{Z}^3\backslash\{\gamma\}$. For filters $a, b_1, \ldots, b_L$, we say that a filter bank $\{a; b_1, \ldots, b_L\}$ is a (3-dimensional dyadic) *framelet filter bank* if

$$\sum_{k\in\mathbb{Z}^3} a(\gamma + 2k)\overline{a(n + \gamma + 2k)} + \sum_{r=1}^{L}\sum_{k\in\mathbb{Z}^3} b_i(\gamma + 2k)\overline{b_i(n + \gamma + 2k)} = \frac{1}{8}\boldsymbol{\delta}(n), \tag{3}$$

For all $\gamma \in \{0, 1\}^3$ and for all $n \in \mathbb{Z}^3$, note that

$$\{0, 1\}^3 = \{(0, 0, 0), (1, 0, 0), \ldots, (1, 1, 1)\} = [0, 1]^3 \cap \mathbb{Z}^3$$

is the set of 8 vertex points in the unit cube $[0, 1]^3$. The filter a is typically a *lowpass filter* satisfying $\sum_k a(k) = 1$ while b_i's are the *highpass filters* satisfying $\sum_k b_i(k) = 0$. Such a filter bank $\{a; b_1, \ldots, b_L\}$ corresponds to a *framelet system* $\{\varphi; \psi_1, \ldots, \psi_L\}$ through refinement relations. For more details, please refer to [30].

Now we construct a 3D directional Haar filter bank $\mathtt{DHF}_3^0 = \{a, b_1, \ldots, b_L\}$ that satisfies Eq. (3). Consider

$$a^H := \frac{1}{8}\left(\boldsymbol{\delta}_{(0,0,0)} + \boldsymbol{\delta}_{(0,0,1)} + \boldsymbol{\delta}_{(0,1,0)} + \boldsymbol{\delta}_{(0,1,1)} + \boldsymbol{\delta}_{(1,0,0)} + \boldsymbol{\delta}_{(1,0,1)} + \boldsymbol{\delta}_{(1,1,0)} + \boldsymbol{\delta}_{(1,1,1)}\right)$$

to be the 3-dimensional Haar low-pass filter. Now, for any two different vertex points γ_1, γ_2 in the unit cube $[0, 1]^3$, we place $+\frac{1}{8}, -\frac{1}{8}$ at each of

$$
\begin{aligned}
b_x &= \tfrac{1}{4}(\boldsymbol{\delta}_{(1,0,0)} - \boldsymbol{\delta}_{(0,0,0)}), & b_y &= \tfrac{1}{4}(\boldsymbol{\delta}_{(0,1,0)} - \boldsymbol{\delta}_{(0,0,0)}), & b_z &= \tfrac{1}{4}(\boldsymbol{\delta}_{(0,0,1)} - \boldsymbol{\delta}_{(0,0,0)}),\\
b_{xy} &= \tfrac{\sqrt{2}}{8}(\boldsymbol{\delta}_{(1,1,0)} - \boldsymbol{\delta}_{(0,0,0)}), & b_{x,y} &= \tfrac{\sqrt{2}}{8}(\boldsymbol{\delta}_{(1,0,0)} - \boldsymbol{\delta}_{(0,1,0)}), & b_{xz} &= \tfrac{\sqrt{2}}{8}(\boldsymbol{\delta}_{(1,0,1)} - \boldsymbol{\delta}_{(0,0,0)}),\\
b_{x,z} &= \tfrac{\sqrt{2}}{8}(\boldsymbol{\delta}_{(1,0,0)} - \boldsymbol{\delta}_{(0,0,1)}), & b_{yz} &= \tfrac{\sqrt{2}}{8}(\boldsymbol{\delta}_{(0,1,1)} - \boldsymbol{\delta}_{(0,0,0)}), & b_{y,z} &= \tfrac{\sqrt{2}}{8}(\boldsymbol{\delta}_{(0,1,0)} - \boldsymbol{\delta}_{(0,0,1)}),\\
b_{xyz} &= \tfrac{1}{8}(\boldsymbol{\delta}_{(1,1,1)} - \boldsymbol{\delta}_{(0,0,0)}), & b_{xy,z} &= \tfrac{1}{8}(\boldsymbol{\delta}_{(1,1,0)} - \boldsymbol{\delta}_{(0,0,1)}), & b_{x,yz} &= \tfrac{1}{8}(\boldsymbol{\delta}_{(1,0,0)} - \boldsymbol{\delta}_{(0,1,1)}),\\
b_{xz,y} &= \tfrac{1}{8}(\boldsymbol{\delta}_{(1,0,1)} - \boldsymbol{\delta}_{(0,1,0)}).
\end{aligned}
$$

Box I.

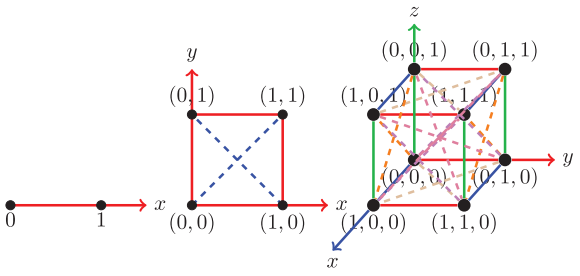

Fig. 2. Directional Haar tight framelet filter banks in $p = 1, 2, 3$ respectively, where each line connecting two vertices $\gamma_1, \gamma_2 \in \{0,1\}^p$ represents a high-pass filter $b_i := 2^{-p}(\boldsymbol{\delta}_{\gamma_1} - \boldsymbol{\delta}_{\gamma_2})$.

these two vertices, respectively, and the corresponding high-pass filter is given by $\frac{1}{8}(\boldsymbol{\delta}_{\gamma_1} - \boldsymbol{\delta}_{\gamma_2})$. After collecting all such filters, we have the set $\{b_1, \ldots, b_L\} := \{\frac{1}{8}(\boldsymbol{\delta}_{\gamma_1} - \boldsymbol{\delta}_{\gamma_2}) : \gamma_1, \gamma_2 \in \{0,1\}^3 \text{ and } \gamma_1 < \gamma_2\}$ of highpass filters. Here $\gamma_1 < \gamma_2$ is understood in the sense of lexicographical order. Then, we have in total $L = \binom{2^3}{2} = 28$ high-pass filters. It was shown in [56] (see also [57,58] for the generalization) that

$$\mathtt{DHF}_3^0 = \{a^H, b_1, \ldots, b_{28}\}$$

is a tight framelet filter bank such that all the highpass filters $b_1, \ldots, b_L$ have only two taps and exhibit 13 directions in dimension 3. We remark that such types of filter banks exist any dimension $p \geq 1$. In particular, for $p = 1$, the tight framelet filter bank is just the standard Haar orthogonal wavelet filter bank $\mathrm{DHF}_1 := \{a^H, b\}$ with $a^H = \frac{1}{2}(\boldsymbol{\delta}_0 + \boldsymbol{\delta}_1)$ and $b = \frac{1}{2}(\boldsymbol{\delta}_0 - \boldsymbol{\delta}_1)$. For $p = 2$, the corresponding tight framelet filter bank reduces to the directional Haar tight framelet filter bank $\mathrm{DHF}_2 := \{a^H, b_1, \ldots, b_6\}$ in [59]. See Fig. 2 for an illustration of the directional Haar tight framelet filter banks in dimension $p = 1, 2, 3$, respectively.

In practice, we employ the `UDFmT` (undecimated discrete framelet transforms) for the decomposition and reconstruction of a 3D tensor. By considering filters with the same direction, the 28 high-pass filters in $\mathtt{DHF}_3^0$ can be regrouped to 13 filters as a 3D Haar semi-tight filter bank as follows:

$$\mathtt{DHF}_3 = \{a^H; b_x, b_y, b_z, b_{xy}, b_{x,y}, b_{xz}, b_{x,z}, b_{yz}, b_{y,z}, b_{xyz}, b_{xy,z}, b_{x,yz}, b_{xz,y}\}$$

where (see the equation in Box I).

For simplicity, we use $\mathtt{DHF}_3 = \{a; b_1, \ldots, b_{13}\}$ to denote the above filter bank with 13 high-pass filters. We call the filter bank $\mathtt{DHF}_3$ as our `3D-HaarFrame`. Note that the filter bank $\mathtt{DHF}_3$ satisfies the partition of unity condition:

$$\sum_{k\in\mathbb{Z}^3} a(\gamma + 2k)\overline{a(\gamma + 2k)} + \sum_{r=1}^{13}\sum_{k\in\mathbb{Z}^3} b_r(\gamma + 2k)\overline{b_r(\gamma + 2k)} = \frac{1}{8}. \tag{4}$$

Now we discuss the decomposition and reconstruction of the 3D-tensor $\boldsymbol{\mathcal{X}}$ using our `3D-HaarFrame`. For a 3D filter $\mathcal{F}$, we denote $\boldsymbol{\mathcal{X}}_{\mathcal{F}}$ the (circular) convolution of $\boldsymbol{\mathcal{X}}$ with the 3D filter $\mathcal{F}$, i.e., $\boldsymbol{\mathcal{X}}_{\mathcal{F}} := \boldsymbol{\mathcal{X}} \star \mathcal{F}$ with

$$\boldsymbol{\mathcal{X}}_h(k) := \sum_{k'\in\mathbb{Z}^3} \widetilde{\boldsymbol{\mathcal{X}}}(k' - k)\cdot \mathcal{F}(k'), \quad k = (k_1, k_2, k_3), k' = (k_1', k_2', k_3') \in \mathbb{Z}^3,$$

where the above $\widetilde{\boldsymbol{\mathcal{X}}}$ is considered the periodic extension of $\boldsymbol{\mathcal{X}}$. Note that $\boldsymbol{\mathcal{X}}_h \in \mathbb{R}^{N\times d\times M}$ is a 3D tensor. Consequently, using the filter bank $\mathtt{DHF}_3$, we can decompose $\boldsymbol{\mathcal{X}}$ to 1 low-pass filter coefficient tensor $\boldsymbol{\mathcal{X}}_a$ and 13 high-pass framelet coefficient tensors $\boldsymbol{\mathcal{X}}_{b_r}, r = 1, \ldots, 13$. Thanks to Eq. (4), the decomposition set $\{\boldsymbol{\mathcal{X}}_a, \boldsymbol{\mathcal{X}}_{b_i}, i = 1, \ldots, 13\}$ of 3D tensors can be used to reconstruct $\boldsymbol{\mathcal{X}}$ perfectly through

$$\boldsymbol{\mathcal{X}}_a \star \bar{a} + \sum_{r=1}^{13} \boldsymbol{\mathcal{X}}_{b_r} \star \bar{b}_r = \boldsymbol{\mathcal{X}},$$

where for a filter $\mathcal{F}$, the filter $\bar{\mathcal{F}}$ is defined as $\bar{\mathcal{F}}(k) = \mathcal{F}(-k), k \in \mathbb{Z}^3$. $\star$ stands for the convolution operation. The set $\{\boldsymbol{\mathcal{X}}_a, \boldsymbol{\mathcal{X}}_{b_r}, r = 1, \ldots, 13\}$ is the one-level decomposition of $\boldsymbol{\mathcal{X}}$. For multi-level decomposition, the input $\boldsymbol{\mathcal{X}}$ is then replaced by $\boldsymbol{\mathcal{X}}_a$ and the filter bank is upsampled, iteratively. Please refer to [60] for the detailed implementation of the `UDFmT` based on the $\mathtt{DHF}_3$.

Robustness property of `3D-HaarFrame`. Theoretically, the `3D-HaarFrame` is a tight framelet system that can be represented as an operator $\boldsymbol{\mathcal{H}}$ satisfying $\boldsymbol{\mathcal{H}}^{\mathrm{T}}\boldsymbol{\mathcal{H}} = \mathbf{I}$. When the 3D tensor $\boldsymbol{\mathcal{X}}$ is perturbed by a noise tensor $\boldsymbol{\mathcal{M}}$, the decomposition of $\boldsymbol{\mathcal{X}}+\boldsymbol{\mathcal{M}}$ results in the framelet coefficient $\boldsymbol{\mathcal{H}}(\boldsymbol{\mathcal{X}}+\boldsymbol{\mathcal{M}})$. Compared to the original framelet coefficient $\boldsymbol{\mathcal{H}}(\boldsymbol{\mathcal{X}})$, it is evident that $\|\boldsymbol{\mathcal{H}}(\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{M}}) - \boldsymbol{\mathcal{H}}(\boldsymbol{\mathcal{X}})\|^2 = \|\boldsymbol{\mathcal{H}}(\boldsymbol{\mathcal{M}})\|^2 = \langle \boldsymbol{\mathcal{H}}^{\mathrm{T}}\boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{M}}\rangle = \langle \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{M}}\rangle = \|\boldsymbol{\mathcal{M}}\|^2$, in view of the tightness of system $\boldsymbol{\mathcal{H}}$. Hence, if noise $\boldsymbol{\mathcal{M}}$ is minimal, the coefficient alteration remains correspondingly slight, underscoring the inherent robustness of the framelet representation. For a deeper theoretical exploration concerning the stability and robustness of the framelet system, we direct readers to [30] for more in-depth theoretical analysis on the stability/robustness of framelet system. To further validate the robustness of the whole framework we developed in this work, we provide empirical verification in Section 4.7.

3.5. Spectrum-based graph structure learning

Based on the obtained framelet coefficients $\boldsymbol{\mathcal{X}}_a, \boldsymbol{\mathcal{X}}_{b_r}, r = 1, \ldots, 13$, each of which is also a 3D-tensor in $\mathbb{R}^{N\times d\times M}$, reflecting the spectrum information of the multimodal data tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N\times d\times M}$, we propose a method for graph structure learning, termed `SGL`. Technically, one of the key points for graph structure learning lies in the design of an appropriate metric function characterizing the similarity between different subjects/instances. In [13], a straightforward yet efficient metric function is proposed that can be learned jointly with the graph neural networks (GNN) used for downstream node classification objective:

$$\mathbf{A}_{ij} = Sim(h_i, h_j) = \cos(\mathbf{W}_A h_i,\ \mathbf{W}_A h_j) \tag{5}$$

where $\mathbf{W}_A$ is a learnable weight matrix and $\mathbf{A}_{ij}$ is computed as the weighted cosine similarity between student i and j, h_i and h_j represent the embedding vector for student (subject) i and student (subject) j, respectively.

However, the metric function defined in Eq. (5) does not consider the different role of the low-pass coefficient tensor $\mathcal{X}_a$ and the collection of low-pass coefficient tensors $\mathcal{X}_{b_i}, i = 1, \ldots, 13$. Under this consideration, we respectively concretize the embedding h_i according to the low-pass components and high-pass components. Formally, we formulate h^{low} and h^{high_r} $(r = 1, \ldots, 13)$ by vectorizing $\mathcal{X}_a$ and $\mathcal{X}_{b_q}, i = q, \ldots, 13$, respectively, as follows:

$$h_i^{\text{low}} := Vec(\mathcal{X}_a(i, :, :)) \in \mathbb{R}^{1\times dM}, \quad i = 1, 2, \ldots, N, \tag{6}$$

$$h_i^{\text{high}_r} := Vec(\mathcal{X}_b(i, :, :)) \in \mathbb{R}^{1\times dM}, \quad i = 1, 2, \ldots, N, \tag{7}$$

which yields $h_i \in \mathbb{R}^{1\times dM}$ as follows:

$$h_i := \varphi(h_i^{\text{low}}, h_i^{\text{high}_1}, \ldots, h_i^{\text{high}_{13}}),$$

where φ is a user-defined aggregation function such as mean, max/min, concatenate, or multi-layer perception (MLP).

In practice, it is common for a realistic adjacency matrix to exhibit characteristics of sparsity and non-negativity. In our case, the adjacency matrix $\mathbf{A}$ represents a dense graph with elements $\mathbf{A}_{ij}$ bounded within the range of [−1,1], which can be computationally demanding. Therefore, similar to [13], we employ a graph sparsification trick to transform $\mathbf{A}$ into a non-negative sparse graph, utilizing a threshold parameter θ. This process involves several steps. Firstly, we normalize the range of $\mathbf{A}$ into [0,1]. Subsequently, we assign a value of zero to the elements in $\mathbf{A}$ that are smaller than the threshold θ. Lastly, we scale the non-zero elements in $\mathbf{A}$ from the interval $[\theta, 1]$ to $[0, 1]$. This simple trick effectively selects neighboring nodes that possess link weights greater than θ for each student. In our practical implementation (see the experiments in Section 4), we set θ to 0.5. The node classification performance (i.e., student engagement prediction accuracy) replies heavily on graph structure $\mathbf{A}$. It is verified in [61] that the constraint on the sparsity, connectivity, and smoothness of the learned graph is also important for adaptive graph learning. Therefore, in this work, we also consider these constraints to regularize the graph learning process. As for the smoothness constraint, we define a framelet-based Dirichlet energy using the obtained low-pass and high-pass coefficients, referring to Eqs. (6) and (7):

$$\mathcal{L}_{smh}(\mathbf{A}, \mathbf{H}) := \frac{\alpha}{N^2}\sum_{i,j=1}^{N}\mathbf{A}_{ij}\left\|h_i^{\text{low}} - h_j^{\text{low}}\right\|_2^2 + \frac{(1-\alpha)}{13N^2}\sum_{r=1}^{13}\sum_{i,j=1}^{N}\mathbf{A}_{ij}\left\|h_i^{\text{high}_r} - h_j^{\text{high}_r}\right\|_2^2, \tag{8}$$

where α is the hyper-parameter used to balance the weighting of low-pass and high-pass framelet coefficients.

Furthermore, to avoid a trivial solution (i.e., $\mathbf{A} = \mathbf{0}$), additional regularization terms are imposed on $\mathbf{A}$ [61], i.e.,

$$\mathcal{L}_{con}(\mathbf{A}) := -\frac{\beta}{N}\mathbf{1}^\top \log(\mathbf{A}\cdot\mathbf{1}) + \frac{\gamma}{N^2}\|\mathbf{A}\|_F^2, \tag{9}$$

where β and γ are two hyper-parameters to balance the regularization terms.

Finally, the objective of SGL is defined as:

$$\mathcal{L}_{\text{SGL}}(\mathbf{A}, \mathbf{H}) := \mathcal{L}_{smh}(\mathbf{A}, \mathbf{H}) + \mathcal{L}_{con}(\mathbf{A}). \tag{10}$$

The primary objective in graph structure learning for node-level predictive tasks is to train a graph structure learner based on $\mathcal{L}_{\text{SGL}}(\mathbf{A}, \mathbf{H})$, which yields an appropriate graph structure. This refined structure is then utilized by the GNN classifier to perform message passing and generate node representations and predictions. The aim of SGL is to produce optimal graph structures that lead to satisfactory classification performance by the GNN classifier. In essence, the training of the GNN classifier with SGL can be viewed as a nested optimization problem, which is discussed in detail in the subsequent section.

3.6. Optimization

Based on the framelet coefficients $\mathbf{H}$, which represent the given graph signal, and the learned sparse graph structure $\mathbf{A}$, graph neural networks can be employed to generate predictive results $\hat{Y}_{GNN} = GNN(\mathbf{A}, \mathbf{H})$. It should be noted that all students, both in the training and testing sets, are considered in generating the student-population graph. Therefore, we formulate the prediction of student engagement as a semi-supervised node classification task, which can be expressed in a general form as follows:

$$\mathcal{L}_{GNN}(Y, \hat{Y}_{GNN}) = \sum_{\mathcal{Y}_{train}} \text{Cross-entropy}(\hat{Y}_{GNN}, Y).$$

Subsequently, the model is trained using a joint optimization objective that integrates the primary (task-aware) loss with the regularization constraint of SGL.:

$$\mathcal{L} = \lambda_1\mathcal{L}_{GNN}(Y, \hat{Y}_{GNN}) + \lambda_2\mathcal{L}_{\text{SGL}}(\mathbf{A}, \mathbf{H}). \tag{11}$$

where λ_1 and λ_2 are two regularization factors weighting the contribution of $\mathcal{L}_{GNN}$ and $\mathcal{L}_{\text{SGL}}$, respectively.

4. Experiments

In this section, we provide an experimental study to verify the effectiveness of our proposed method for student engagement prediction. To begin with, we introduce a benchmark educational dataset with multimodal data collected in the "in-the-wild" online environment of Zoom. Furthermore, several GNN-based existing methods are introduced, followed by a demonstration of the experiment results and an extensive discussion. Then, details of our experimental setup, including data preprocessing, the setting of the hyper-parameters, etc., are presented. In addition, we conduct ablation studies to further validate the role of the key modules of our proposed framework. We implement all experiments in Python 3.8.13 with PyTorch on one NVIDIA® Tesla A100 GPU with 6912 CUDA cores and 80 GB HBM2 mounted on an HPC cluster.

4.1. Dataset

To evaluate the effectiveness of our proposed method, we employ the RoomReader[3] dataset [10] as a benchmark including over 8 h of video and audio recordings, capturing the interactions of 118 participants across 30 sessions that take place in the online environment of Zoom. The RoomReader dataset consists of multimodal, multiparty conversational interactions that simulate a collaborative online student-tutor scenario, where audios, videos, as well as transcriptions are recorded accordingly. The dataset focuses on measuring off-task/on-task engagement, with the instructor leading the given task. Additionally, it contains extensive engagement annotations, group cohesion measures, and supplementary information about the participants, such as personality test results. Notably, the student participants in the corpus have been continuously annotated for engagement using a unique continuous scale, allowing for the detailed examination of engagement dynamics. This rich collection of data allows for a comprehensive analysis and exploration of various aspects of the tutorial sessions and participant characteristics, fitting well with the task of student engagement prediction. Furthermore, this multimodal corpus provides questionnaires that measure engagement and group cohesion, collected from annotators, tutors, and the participants themselves, and also offers a variety of supplementary data, such as personality tests and behavioral assessments. To date, the RoomReader corpus stands out as a valuable openly-accessible dataset that examines the multimodal indicators of conversational engagement and the behavioral facets of

[3] https://sigmedia.tcd.ie/.

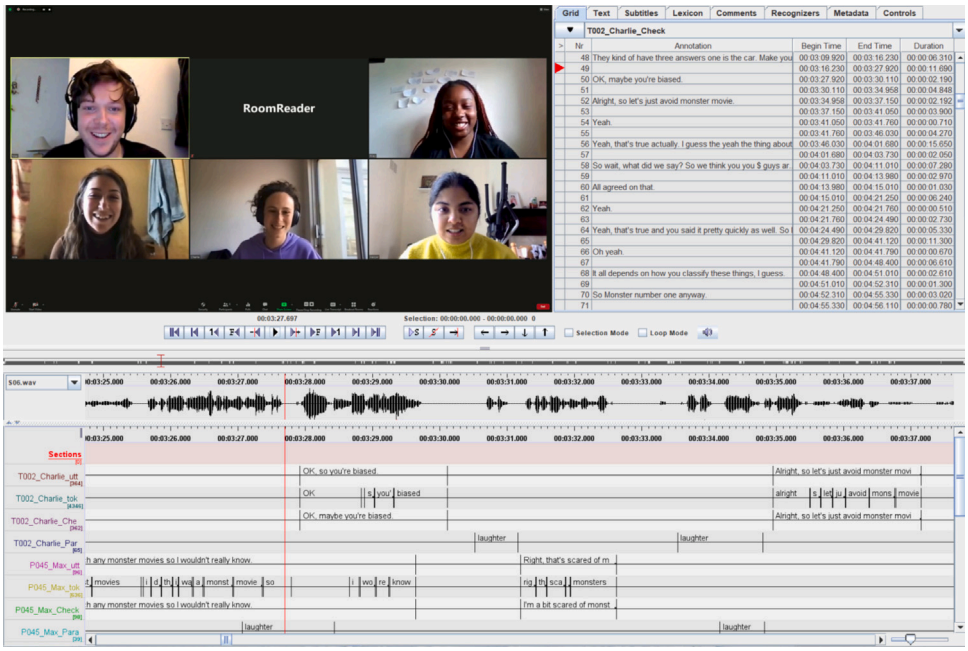

Fig. 3. A screenshot from [10]: Example of a session collected in the RoomReader corpus, where the cropped videos of participants, participant-level and session-level audios, ASR-generated and rich manually-corrected transcriptions are recorded.

collaborative interaction in online learning environments. It is expected to serve as a significant resource that enables researchers to explore numerous potential areas, including multimodal, multiparty conversations in online settings, intelligent education techniques and applications, and multimodal human–robot interactions (HRI), etc. [10] (see Fig. 3).

4.2. Baselines

To undertake a comprehensive performance comparison, we compare the proposed method with various existing works from different families or with different application purposes. We select four baselines: ConvLSTM [62], TEMMA [63], EnsModel [64], Bootstrap [65] which have already been employed for engagement prediction. However, these four methods do not consider graph-based learning and only consider single modality in algorithm implementation. On the other hand, there is a lack of existing literature that investigates the task of student engagement prediction specifically utilizing the RoomReader dataset. To evaluate extensively the effectiveness of our proposed method, we conduct a comparative analysis against several specific baseline models that have exhibited exceptional performance in disease prediction tasks, such as PopGCN [66], EV-GCN [67], MMGL [13]. It should be noted that we re-implement all these baselines to fit the basic requirements of the RoomReader dataset. Basic information and their source code repositories are introduced as follows:

- ConvLSTM [62]: ConvLSTM uses CNN and LSTM networks to empower robots in calculating a singular engagement value during their interactions with humans. These networks leverage standard video streams, which are captured from the perspective of the interacting robot. We obtain the source code which we utilize for implementation and customization from the repository.[4]
- TEMMA [63]: In TEMMA, a convolutional neural network-transformer encoder is proposed, which incorporates a trans former-encoder with a self-attention mechanism. Its primary objective is to model the temporal dependency in the context of single modal affect recognition. We obtain the source code that we utilize for implementation and customization from the repository.[5]
- EnsModel [64]: The proposed methodology consists of three essential (and general) stages: feature extraction, regression, and model ensemble. Specifically, a combination of long short-term memory (LSTM) and fully connected layers is employed to capture the temporal information and predict the engagement intensity based on the extracted features, followed by a fusion strategy applied to enhance the overall performance of the model. We obtain the source code that we utilize for the implementation and customization from the repository.[6]
- Bootstrap [65]: This method is proposed to predict the engagement intensity value of a student when he or she is watching an online MOOCs video in various conditions. For problem-solving, it maintains the framework of multi-instance learning with the LSTM network and use the classical bootstrap aggregation method to perform the model ensemble. We obtain the source code which we utilize for implementation and customization from the repository[7]
- PopGCN [66]. PopGCN utilizes demographic information to construct a population graph manually and then applies the GCN model [68] to aggregate the imaging features of subjects for classification purposes. We obtain the source code that we utilize for implementation and customization from the repository.[8]
- EV-GCN [67]: In EV-GCN, the connections within the population graph are computed using a learnable function that takes into account non-imaging measurements. We obtain the source code that we utilize for implementation and customization from the repository.[9]

[4] https://github.com/LCAS/engagement_detector.

[5] https://github.com/Sunner4nwpu/TEMMA.

[6] https://github.com/AnshulSood11/Engagement-Level-Prediction.

[7] https://github.com/kaiwang960112/EmotiW_2019_engagement_regression.

[8] https://github.com/parisots/population-gcn.

[9] https://github.com/SamitHuang/EV_GCN.

- MMGL [13]: Compared with the aforementioned models, MMGL has been recognized as a state-of-the-art (SOTA) multi-modal graph learning framework (for disease prediction task). MMGL uses a modal-aware representation learning (MARL) module for mining modality-specific and modality-shared embeddings. Also, MMGL incorporates an adaptive graph structure learning (AGL) block that unveils the inherent relationships among subjects, facilitating the construction of an optimized graph structure tailored for downstream tasks. Our proposed method is also motivated from the general schematic of MMGL. The key differences lie in the fact that we use the proposed `3D-HaarFrame` and `SGL` to replace the MARL and AGL modules, respectively (see Section 3 for details). We obtain the source code that we utilize for implementation and customization from the repository.[10]
- MM-DFN [69]: presents a graph-based multimodal fusion strategy tailored for emotion recognition in multimodal conversation (MERC). The graph structure is expertly designed to capture the nuances of the intra-speaker context and the interdependencies between modalities. Such a configuration ensures a seamless incorporation of multi-modal data, while also synthesizing comprehensive contextual insights, even from long-distance sources. To align with our specific problem formulation to predict student engagement, we modify the original source code[11] by feeding new inputs derived from specialized pre-processing stages for students' multimodal data, revising the training objective, and tailoring the outputs accordingly.
- M^3Net [70]: As the latest state-of-the-art (SOTA) model for the MERC task, M^3Net employs a multivariate multi-frequency multimodal GNN for problem-solving. It delves deep into the intricate relationships between various modalities and their contexts. Moreover, it efficiently harnesses frequency data to discern and highlight the contrasts and overlaps in emotional expressions. In line with the modifications we make to MM-DFN [69], we specifically tailor the original source code of M^3Net[12] to aptly serve the task of student engagement prediction.

We note that ConvLSTM [62], TEMMA [63], EnsModel [64], Bootstrap [65], PopGCN [66] perform on single modal data (i.e., visual modality) and only PopGCN [66] uses the graph-based method by manually constructing static graphs. In comparison, EV-GCN [67] and MMGL [13] make use of multi-modal data, and they both consider graph structure learning during the model's training process. For the methods which only use single modality, we select the visual modality (i.e., image fragments obtained from the video data) from RoomReader.

4.3. Experimental setup

Data Preparation. In the RoomReader dataset, continuous annotations for engagement are provided, where the engagement labels range from [−2,2]. As demonstrated in [71], approximately 80.2% of the samples correspond to highly engaged instances, of which the annotated engagement values range from (1, 2], 18.3% to low engagement with the annotated engagement values ranging from (0,1], 1.3% to low disengagement with the annotated engagement values ranging from (−1,0], and 0.2% to high disengagement with the annotated engagement values ranging from [−2,−1]. This exhibits severe class imbalance, which technically is out of the main scope and focus of this research. To show the primary merits of our proposed method for modality learning with graphs, without the potential influence of the class imbalance challenge, which usually involves effective oversampling techniques and specific-designed (advanced) class-sensitive loss functions, we study a simple and balanced two-class classification task in our experiments. Specifically, from the 30 sessions, we only randomly select 16 000 highly engaged instances (Class-I) and 16 000 low engagement samples (Class-II), and average the engagement scores over 8 s, leading to 2000 highly engaged subjects and 2000 poorly engaged subjects, respectively. This means that a student-population graph with 4000 nodes will be constructed in the `SGL` module of our proposed framework, by which a two-class semi-supervised node classification task is to be solved. We note that the number of nodes here does not represent the number of participants in the RoomReader project because we have selected multiple samples for each participant. Based on the averaged annotated labels and the time window size, we select the associated visual (i.e., the facial image segment sequences during the time window), textual (i.e., the transcripts), and acoustic (i.e., the audio) modalities as multimodal inputs, consistent with the formulation presented in Section 3.

Data Preprocessing. As aforementioned, ConvLSTM [62], TEMMA [63], EnsModel [64], and Bootstrap [65] use visual modality for problem-solving. In our experiments, similar to [71], we utilize the normalized eye gaze direction, location of the head, location of 3D landmarks, and facial action units extracted via Open-Face [72] as the input features. Building upon the work presented in [10], which provides all the Open-Face features across all sessions in conjunction with multimodal data sources, we conduct experiments on ConvLSTM [62], TEMMA [63], EnsModel [64], and Bootstrap [65] using these features as inputs. Specifically, for the PopGCN [66] approach, which focuses solely on visual modality, we employ the selected image data to align with their algorithm implementation. Furthermore, for the remaining baselines that involve graph construction or graph learning components, we incorporate all the necessary modalities and re-implemented their models to predict student engagement.

Modality-specific Feature Extraction. The specification of the pre-trained model used for modality-specific feature extraction is essential, as generally discussed in Section 3.3. In our practical algorithm implementation, we employ advanced pre-trained models to perform feature extraction for each modality, realizing Eq. (1). Specifically, visual modality is processed using the pre-trained MA-Net [48], which yields 1024-dimensional frame-level facial features. The textural modality utilizes the pre-trained DeBERTa [55], resulting in 1024-dimensional features. Lastly, the acoustic modality employs the pre-trained wav2vec [52], generating 512-dimensional acoustic features.

Hyperparameter setting. During training, we use the Adam optimization scheme with a learning rate of 0.001 and a weight decay of 1e−6. To alleviate the over-fitting problem, Dropout [73] is also utilized with a rate of $p = 0.5$. We set the unified dimension d to 100 (see Eq. (2)). For the hyper-parameter tuning, both β and γ (see Eq. (9)) are tuned through the Hyperopt library [74]. Based on our empirical experience, both λ_1 and λ_2 (see Eq. (11)) are set to 1. As for the key hyperparameter α, determining the contribution of low-pass and high-pass coefficients to the graph structure learning (see Eq. (8)), we select $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and study in-depth the robustness of the model in terms of the setting of α, as detailed later.

4.4. Results and discussion

Table 1 presents the results of the performance comparison, which shows clearly that our proposed method outperforms all the baselines in classification accuracy. The mean and standard deviation are obtained based on 5 independent trials. It is also evident that approaches employing multi-modal techniques consistently outperform those relying solely on a single modality. Moreover, the methods utilizing the strategy of adaptive graph structure learning (i.e., Graph Type-'Dynamic'), namely EV-GCN, MMGL, MM-DFN and M^3Net, demonstrate a significant improvement in performance compared to those simply using a heuristic method of graph construction (i.e., Graph Type-'Static') . This observation confirms the advantage of learning graph structure over

[10] https://github.com/SsGood/MMGL.
[11] https://github.com/zerohd4869/MM-DFN.
[12] https://github.com/feiyuchen7/M3NET.

Table 1
The performance comparison of student engagement prediction accuracy. N/A: graph-based learning approach/strategy is not involved.

Method	ACC. (%)	Modal type	Graph type
ConvLSTM [62]	76.50 ± 1.85	Single	N/A
TEMMA [63]	80.90 ± 2.47	Single	N/A
EnsModel [64]	75.30 ± 3.50	Single	N/A
Bootstrap [65]	73.80 ± 3.35	Single	N/A
PopGCN [66]	84.30 ± 2.15	Single	Static
EV-GCN [67]	87.80 ± 3.32	Multiple	Dynamic
MMGL [13]	88.86 ± 1.18	Multiple	Dynamic
MM-DFN [69]	88.12 ± 2.24	Multiple	Dynamic
M^{3}Net [70]	89.36 ± 1.32	Multiple	Dynamic
Haar-MGL (Ours)	**90.18 ± 1.34**	Multiple	Dynamic

Table 2
Ablation study for Haar-MGL. N/A: graph-based learning approach/strategy is not involved.

Candidates	ACC. (%)	Graph Type
Haar-MGL (full)	**90.18 ± 1.34**	Dynamic
MLP+SGL$_{refined}$	82.12 ± 2.57	Dynamic
Concat.+SGL$_{refined}$	81.56 ± 2.67	Dynamic
Haar-MGL$_{w/o\ \mathtt{SGL}}$	86.37 ± 1.32	N/A
3D-HaarFrame+G_{popGCN}	84.50 ± 1.68	Static
3D-HaarFrame+G_{kNN}	83.28 ± 2.35	Static

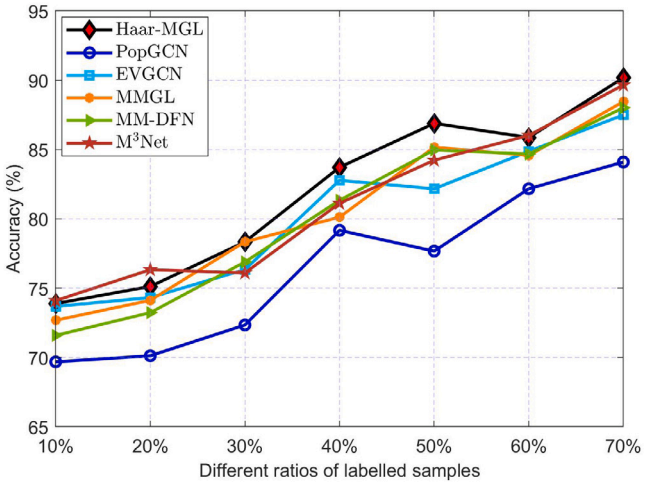

Fig. 4. Visualization of the performance comparison of PopGCN, EV-GCN, MMGL, MM-DFN, M^{3}Net, and Haar-MGL with labeled samples of different ratios.

using static ones, as they offer more appropriate graph topology and benefits the down-stream graph representation learning task.

It is worth noting that the results presented in Table 1 were obtained using a training/testing partition ratio of 70%/30%, with an additional 10% of samples from the training set reserved for validation purposes. With a specific focus on GNN-based methods, namely PopGCN, EV-GCN, MMGL, MM-DFN, M^{3}Net, we also conduct a performance comparison under varying training set sizes. Specifically, the label ratio of samples was varied from 10% to 70%, and for each case, we compared the test accuracy of PopGCN, EV-GCN, MMGL, MM-DFN, M^{3}Net, and Haar-MGL. As depicted in Fig. 4, our proposed Haar-MGL consistently outperforms the other three baselines, even when the number of training samples is limited. This observation suggests that Haar-MGL achieves favorable performance in the context of semi-supervised node classification.

4.5. Ablation study

We conduct ablation studies to assess the effectiveness of the 3D Haar semi-tight framelet transform (i.e., `3D-HaarFrame`) for multi-modal information fusion, and spectrum-based graph structure learning (i.e., `SGL`). Specifically, the following candidates are considered in our ablation experiments:

- MLP+SGL$_{refined}$: We replace `3D-HaarFrame` with multi-layer perceptron (MLP). In such a case, we need to modify `SGL` since low-pass and high-pass framelet coefficients are not available if the `3D-HaarFrame` is removed. In practice, we simply use the outputs from MLP to redefine the Dirichlet energy defined in (8), where the framelet coefficient signal is changed by the feature vector obtained by MLP;
- Concat.+SGL$_{refined}$: Similar to the above operation, we replace `3D-HaarFrame` with a direct concatenation trick. As such, we also need to convert `SGL` to the aforementioned SGL$_{refined}$;
- Haar-MGL$_{w/o\ \mathtt{SGL}}$: We remove the `SGL` module from Haar-MGL. In such a case, instead of proceeding to the graph learning module, we directly input the low-pass and high-pass framelet coefficients produced by `3D-HaarFrame` to a fully-connected classifier for engagement prediction.
- `3D-HaarFrame`+G_{popGCN}: We replace `SGL` with the graph construction method adopted in popGCN [66]. For this purpose, both low-pass and high-pass framelet coefficients are treated as signals (in the frequency domain) with an equal contribution to the graph construction G_{popGCN};
- `3D-HaarFrame`+G_{kNN}: We substitute `SGL` with a standard kNN graph G_{kNN} using the RBF kernel. Similar to the above, all the framelet coefficients produced by `3D-HaarFrame` are viewed as regular inputs for building G_{kNN}.

Table 2 summarizes the ablation study results for the aforementioned variants of Haar-MGL. It is clear that all the modified candidates have lower accuracy than the full model, verifying convincingly the contribution of each module in Haar-MGL. Specifically, "MLP+SGL$_{refined}$" and "Concat.+SGL$_{refined}$" lead to poorer accuracy than the other variants, which validates the effectiveness and contribution of `3D-HaarFrame` for multimodal feature fusion. Interestingly, "Haar-MGL$_{w/o\ \mathtt{SGL}}$" achieves favorable performance despite the absence of `SGL`, which again verifies the strength of `3D-HaarFrame`, as the primary contribution of this work. Moreover, both "`3D-HaarFrame`+G_{popGCN}" and "`3D-HaarFrame`+G_{kNN}" exhibit a decline in performance compared to "Haar-MGL$_{w/o\ \mathtt{SGL}}$". This observation highlights the significance of selecting an adaptive way to update the graph structure along with training the whole model, since an inappropriate (fixed) graph structure can have detrimental effects on the overall performance.

As previously discussed, ConvLSTM [62], TEMMA [63], EnsModel [64], and Bootstrap [65] employ visual modality for problem-solving. In our experiments, following the approach of [71], we use the OpenFace features (i.e., normalized eye gaze direction, the location of the head, the location of 3D landmarks, and facial action units) as input features for ConvLSTM [62], TEMMA [63], EnsModel [64], and Bootstrap [65]. To be consistent with this experimental setup and to further verify the effectiveness of the ability of `3D-HaarFrame` when using only visual features, we consider another variant of Haar-MGL, which removes the block of using pre-trained models (see the discussion in Section 4.3) and directly uses the aforementioned Open-Face features provided in the RoomReader corpus, i.e., Haar-MGL$_{single}$. Table 3 compares the performance of Haar-MGL$_{single}$ and ConvLSTM [62], TEMMA [63], EnsModel [64], Bootstrap [65]. It is clear that Haar-MGL$_{single}$ outperforms the other four baselines. This observation implies that the superior performance of Haar-MGL can be primarily attributed to the utilization of `3D-HaarFrame` and `SGL`. Overall, by showcasing

Table 3
Performance comparison in the case of using only visual modality. N/A: graph-based learning approach/strategy is not involved.

Method	ACC. (%)	Modal type	Graph type
ConvLSTM [62]	76.50 ± 1.85	Single	N/A
TEMMA [63]	80.90 ± 2.47	Single	N/A
EnsModel [64]	75.30 ± 3.50	Single	N/A
Bootstrap [65]	73.80 ± 3.35	Single	N/A
Haar-MGL$_{single}$	**85.40 ± 2.15**	Single	Dynamic

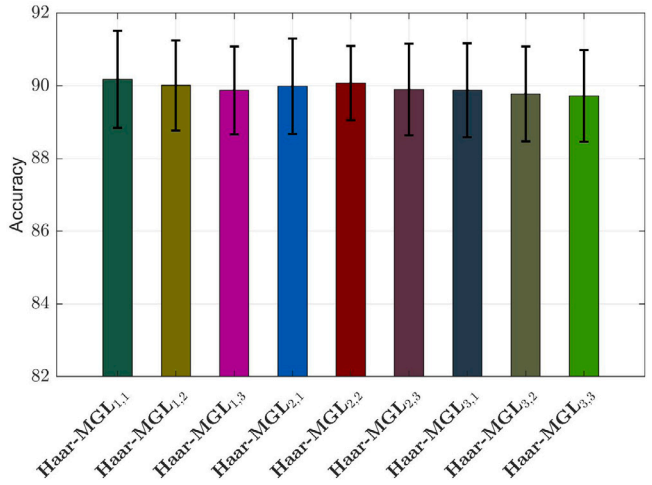

Fig. 5. Ablation study on Haar-MGL with different configurations of feature extractors for visual and textual features.

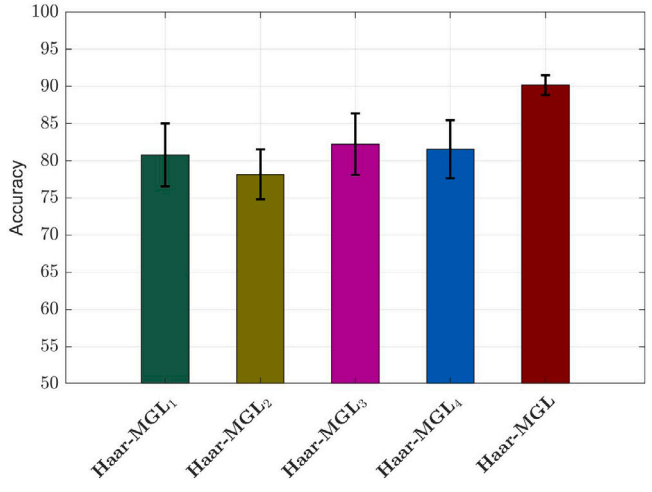

Fig. 6. Performance comparison for evaluating the role of each loss.

the superior results compared to the other baselines and variants, the findings of the ablating studies highlights convincingly the technical contribution and effectiveness of our proposed framework.

It is worth noting that the **Haar-MGL** model employs the pre-trained MA-Net [48], DeBERTa [55], and wav2vec [52] as feature extractors for visual, textual, and acoustic modalities, respectively. Technically, as highlighted in Section 3.3, there are multiple choices for the configuration of the feature extractor for each modality. We carry out ablation studies to assess how the employment of different feature extractors influence the ultimate prediction accuracy. In the literature, numerous pre-trained models exist for visual, textual, and acoustic feature extraction, however our analysis focuses solely on the widely used models detailed in Section 3.3. Regarding the acoustic feature extractor, we are only able to deploy smoothly the original source code for the pre-trained wav2vec [52]. Unfortunately, the code for OpenSmile [51] failed to compile successfully in our experiments due to unresolved bugs. Hence, this ablation study focuses solely on various configurations combining visual and textual feature extractors. We evaluate a total of nine configurations, yielding 9 distinct models: Haar-MGL$_{i,j}$, where $i \in 1,2,3$ corresponds to the ordered selection from {MA-Net [48], VGG-Face [49], ResNet-50 [50]}, and $j \in 1,2,3$ represents the choice among {DeBERTa [55], RoBERTa [54], BERT [53]} accordingly. It should be noted that in Haar-MGL1, 1 matches the model '**Haar-MGL (Ours)**' presented in Table 2. The accuracy of each model, including their mean and standard deviation values, are plotted in Fig. 5. Observably, the performance of Haar-MGL$_{1,1}$ slightly surpasses that of the other models. The least effective combination (using ResNet-50 + BERT) achieves 89.72 ± 1.26, yet still outperforms the second-best model, M^{3}Net, as shown in Table 2. In summary, these ablation studies consistently validate the superior efficacy and benefits of our proposed framework.

4.6. Assessing the role of each loss

In Section 4.5, we have assessed the significance of graph structure learning by analyzing the performance of the variant model: Haar-MGL$_{w/o\ \mathtt{SGL}}$, which discards the loss term $\mathcal{L}\mathtt{SGL}$ from the training objective of Haar-MGL. However, understanding the individual impact of each component within $\mathcal{L}\mathtt{SGL}$ remains crucial. As a reminder, we define $\mathcal{L}_{\mathtt{SGL}}(\mathbf{A},\mathbf{H}) := \mathcal{L}_{smh}(\mathbf{A},\mathbf{H}) + \mathcal{L}_{con}(\mathbf{A}) := \mathcal{L}_{smh}(\mathbf{A},\mathbf{H}) - \frac{\beta}{N}\mathbf{1}^\top \log(\mathbf{A}\cdot\mathbf{1}) + \frac{\gamma}{N^2}\|\mathbf{A}\|_F^2$ (see Section 3.5). In pursuit of this deeper understanding, we conduct further experiments concentrating on:

- **Haar-MGL$_1$**: This model variant is trained based on a refined loss by excluding $\mathcal{L}_{smh}$ from $\mathcal{L}_{\mathtt{SGL}}$. In such a case, $\mathbf{A}$ is updated, bypassing the similarity metric learning phase;
- **Haar-MGL$_2$**: This refined model is trained by omitting $\mathcal{L}_{con}$ (i.e., the last two terms) from $\mathcal{L}_{\mathtt{SGL}}$;
- **Haar-MGL$_3$**: This is a variant emphasizing the role of $\beta = 0$ by setting its value to zero in $\mathcal{L}_{con}$. Similar to the previous experiments, γ is tuned through the Hyperopt library [74];
- **Haar-MGL$_4$**: Conversely, this variant considers $\gamma = 0$ in $\mathcal{L}_{con}$, while β is tuned through the Hyperopt library [74].

As shown clearly in Fig. 6, **Haar-MGL** achieves superior performance compared to the variant models tailored with specific loss functions. Similar to the previous experiment setup, the mean and standard deviation are obtained based on five independent trials. This empirical evidence underscores the integral roles played by each component of $\mathcal{L}_{\mathtt{SGL}}$. Coupled with insights from our prior ablation study on Haar-MGL$_{w/o\ \mathtt{SGL}}$ (see Table 2), it is evident that while graph structure learning proves valuable, the design of the loss function necessitates meticulous attention, especially considering constraints on the sparsity, connectivity, and smoothness of the learned graph.

4.7. Robustness analysis

To empirically assess the robustness of the $\mathtt{3D-HaarFrame}$ system (as highlighted in Section 3.4 from a theoretical viewpoint), we carry out additional experiments to evaluate the robustness of $\mathtt{3D-HaarFrame}$ against varying levels of noise. Specifically, before the decomposition phase of $\mathtt{3D-HaarFrame}$ (refer to Fig. 1), we introduce additive white Gaussian noise $\mathcal{M} \sim \mathcal{N}(0,\sigma^2) \in \mathbb{R}^{N\times d\times M}$ directly to the 3D-tensor $\mathcal{X} \in \mathbb{R}^{N\times d\times M}$. Here, $\sigma = p(\max(\mathcal{X}) - \min(\mathcal{X}))$, with p selected from the set $0.01, 0.03, 0.05, 0.08, 0.1$, which indicates the extent of noise introduced to $\mathcal{X}$. In the experiments, we execute this experimental design for both our proposed **Haar-MGL** model and the baseline MMGL, which also employs a 3D-tensor representation in their framework. In

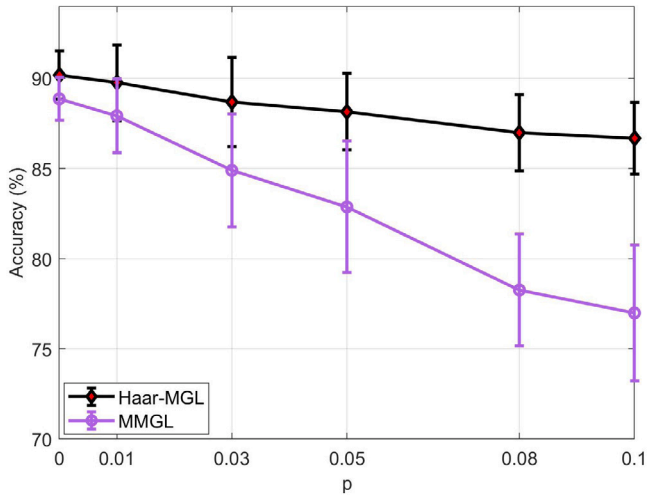

Fig. 7. Performance demonstration for robustness analysis.

Fig. 7, we plot the accuracy of **Haar-MGL** and MMGL under varying noise degrees. It is clear that **Haar-MGL** achieves superior robustness to MMGL across these noise levels.

5. Conclusion and future work

In summary, in the context of the student engagement prediction task, this paper develops a multimodal graph learning method based on a 3D Haar semi-tight framelet for problem-solving. The proposed 3D Haar semi-tight framelet transform is able to capture intermodal relationships and model complex interactions within multimodal data. Additionally, we introduce an adaptive graph structure learning module that considers the different contributions of low-pass and high-pass framelet coefficients by adaptively weighing their impact. Through extensive experimental evaluations on a real-world educational dataset, we demonstrate that our approach achieves superior performance compared to state-of-the-art methods, highlighting the effectiveness of multimodal graph learning in accurately predicting student engagement.

Limitations. We note that, even when executed on an HPC cluster, the average training time of our method is roughly three hours, a significant portion of which is consumed by hyperparameter tuning using the Hyperopt library. Furthermore, the inference (testing) time approaches 15 s. This suggests that adapting our method for real-time systems, which require efficient operation with streaming data in genuine educational environments, poses a substantial challenge. As educational systems become more interconnected, and as lessons increasingly migrate online, processing real-time, streaming data becomes vital. Another pertinent limitation is the method's current incapability to handle missing modality issues seamlessly. This is pivotal as in real-world scenarios, not all modalities may be present at all times. On the other hand, privacy concerns are another frontier we must address, especially when working with video sources. Video data can inadvertently lead to privacy breaches, notably the unintended release of student portraits or other personally identifiable information. Ensuring that our framework adopts stringent data handling and protection protocols is crucial to mitigate these risks. Due to the scarcity of benchmark datasets for multimodal-based student engagement prediction, assessing the generalizability of our framework across a wider range of datasets and benchmarks presents a challenge. While the robustness of the `3D – HaarFrame` system has been evaluated in Section 4.7, its performance in actual educational environments with more complex uncertainties like missing frames and audio noise remains untested. Additionally, the relative scarcity of audio/text data compared to image data presents a further challenge for implementing our framework in real-world educational settings. Finally, for practical application, deploying our proposed framework in an actual classroom setting necessitates additional work in the development of a comprehensive system and platform.

Future Work. Several paths for future research and exploration exist: Firstly, an extension of our proposed framework to accommodate multimodal data with incompleteness is essential. This incompleteness may arise from diverse sources such as sensor damage, data corruption, or human errors during recording. Incorporating robust methods to handle and integrate incomplete multimodal data would be valuable. Second, enhancing the explainability/interpretability of engagement prediction models, and gaining insights into the specific contributions of each modality, are crucial for understanding the underlying factors that contribute to student engagement. This understanding has far-reaching implications on personalized learning, optimal resource allocation, early intervention and targeted learning support. Moreover, it is meaningful to develop advanced techniques for domain adaptation and transfer learning could enable the transfer of knowledge learned from one setting to another, enhancing the generalizability and scalability of student engagement prediction models. Additionally, applying our proposed framework, particularly the components of `3D–HaarFrame` and `SGL`, to challenging multimodal data modeling tasks in natural sciences and medicine would be highly desirable. Last, in our subsequent project, we plan to offer pre-trained models for each modality and release crucial checkpoints that include these models at specific stages. This will enable the creation of a demo that operates independently of the original RoomReader data source, thereby alleviating the need for extensive storage. Thereafter, we will enhance our coding by dedicating additional engineering resources to the development of both front-end and back-end systems and platforms.

CRediT authorship contribution statement

Ming Li: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Xiaosheng Zhuang:** Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Lu Bai:** Visualization, Writing – review & editing. **Weiping Ding:** Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Ming Li acknowledges the supports from the Key R&D Program of Zhejiang Province, China (No. 2024C03262), the National Natural Science Foundation of China (No. 62172370, No. U21A20473), the Key R&D Program of Zhejiang Province, China (No. 2022C03106), and Zhejiang Provincial Natural Science Foundation, China (No. LY22F020004). The work of Xiaosheng Zhuang was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project CityU 11309122 and Project CityU 11302023. Lu Bai acknowledges the support from the National Natural Science Foundation of China (No. T2122020, No. 61976235, No. 61602535). Weiping Ding acknowledges the supports from Natural Science Foundation of Jiangsu Province (No. BK20231337), the Natural Science Key Foundation of

Jiangsu Education Department (No. 21KJA510004), and the National Natural Science Foundation of China (No. 61976120). We also would like to thank the authors in [10] for providing the RoomReader dataset (https://sigmedia.github.io/datasets/room_reader/)(We note that the entire database was downloaded by us in October 2022). The present study's experimental implementation utilizes the underlying source code of MMGL (https://github.com/SsGood/MMGL) as a fundamental reference.

References

[1] J.A. Fredricks, P.C. Blumenfeld, A.H. Paris, School engagement: Potential of the concept, state of the evidence, Rev. Educ. Res. 74 (1) (2004) 59–109.
[2] J.A. Fredricks, M. Filsecker, M.A. Lawson, Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues, Learn. Instr. 43 (2016) 1–4.
[3] G.M. Sinatra, B.C. Heddy, D. Lombardi, The challenges of defining and measuring student engagement in science, Educ. Psychol. 50 (1) (2015) 1–13.
[4] S. D'Mello, E. Dieterle, A. Duckworth, Advanced, analytic, automated (AAA) measurement of engagement during learning, Educ. Psychol. 52 (2) (2017) 104–123.
[5] K. Doherty, G. Doherty, Engagement in HCI: conception, theory and measurement, ACM Comput. Surv. 51 (5) (2018) 1–39.
[6] S.K. D'Mello, Improving student engagement in and with digital learning technologies, in: OECD Digital Education Outlook 2021 Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots, 2021, pp. 79–104.
[7] L. Geng, M. Xu, Z. Wei, X. Zhou, Learning deep spatiotemporal feature for engagement recognition of online courses, in: 2019 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2019, pp. 442–447.
[8] J. Liao, Y. Liang, J. Pan, Deep facial spatiotemporal network for engagement prediction in online learning, Appl. Intell. 51 (2021) 6609–6621.
[9] F. Xu, L. Wu, K. Thai, C. Hsu, W. Wang, R. Tong, MUTLA: A large-scale dataset for multimodal teaching and learning analytics, 2019, arXiv preprint arXiv:1910.06078.
[10] J. Reverdy, S.O. Russell, L. Duquenne, D. Garaialde, B.R. Cowan, N. Harte, RoomReader: A multimodal corpus of online multiparty conversational interactions, in: Proceedings of the 13th Language Resources and Evaluation Conference, 2022, pp. 2517–2527.
[11] A. Sabuncuoglu, T.M. Sezgin, Developing a multimodal classroom engagement analysis dashboard for higher-education, Proc. ACM Hum.-Comput. Interact. 7 (EICS) (2023) 1–23.
[12] Y. Ektefaie, G. Dasoulas, A. Noori, M. Farhat, M. Zitnik, Multimodal learning with graphs, Nat. Mach. Intell. 5 (2023) 340–350.
[13] S. Zheng, Z. Zhu, Z. Liu, Z. Guo, Y. Liu, Y. Yang, Y. Zhao, Multi-modal graph learning for disease prediction, IEEE Trans. Med. Imaging 41 (9) (2022) 2207–2216.
[14] J. Mao, J. Liu, H. Lin, H. Kuang, Y. Pan, Multi-modal multi-kernel graph learning for Autism prediction and biomarker discovery, 2023, arXiv preprint arXiv:2303.03388.
[15] F. Yang, H. Wang, S. Wei, G. Sun, Y. Chen, L. Tao, Multi-model adaptive fusion-based graph network for Alzheimer's disease prediction, Comput. Biol. Med. 153 (2023) 106518.
[16] Y. Lin, K. Lu, S. Yu, T. Cai, M. Zitnik, Multimodal learning on graphs for disease relation extraction, J. Biomed. Inform. 143 (2023) 104415.
[17] W. Zheng, L. Yan, C. Gou, Z.-C. Zhang, J.J. Zhang, M. Hu, F.-Y. Wang, Pay attention to doctor–patient dialogues: multi-modal knowledge graph attention image-text embedding for COVID-19 diagnosis, Inf. Fusion 75 (2021) 168–185.
[18] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, T.-S. Chua, MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1437–1445.
[19] Z. Tao, Y. Wei, X. Wang, X. He, X. Huang, T.-S. Chua, MGAT: Multimodal graph attention network for recommendation, Inf. Process. Manage. 57 (5) (2020) 102277.
[20] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, K. Zheng, Multi-modal knowledge graphs for recommender systems, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 1405–1414.
[21] C. Ding, S. Sun, J. Zhao, MST-GAT: A multimodal spatial–temporal graph attention network for time series anomaly detection, Inf. Fusion 89 (2023) 527–536.
[22] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal sentiment analysis based on fusion methods: A survey, Inf. Fusion 95 (2023) 306–325.
[23] R. Saqur, K. Narasimhan, Multimodal graph networks for compositional generalization in visual question answering, Adv. Neural Inf. Process. Syst. (2020) 3070–3081.
[24] J. Wang, J. Hu, S. Qian, Q. Fang, C. Xu, Multimodal graph convolutional networks for high quality content recognition, Neurocomputing 412 (2020) 42–51.
[25] W. Zhang, J. Yu, W. Zhao, C. Ran, DMRFNet: deep multimodal reasoning and fusion for visual question answering and explanation generation, Inf. Fusion 72 (2021) 70–79.
[26] S. Uppal, S. Bhagat, D. Hazarika, N. Majumder, S. Poria, R. Zimmermann, A. Zadeh, Multimodal research in vision and language: A review of current and emerging trends, Inf. Fusion 77 (2022) 149–171.
[27] S. Mai, S. Xing, J. He, Y. Zeng, H. Hu, Multimodal graph for unaligned multimodal sequence analysis via graph convolution and graph pooling, ACM Trans. Multimed. Comput. Commun. Appl. 19 (2) (2023) 1–24.
[28] L.A. Passos, J.P. Papa, J. Del Ser, A. Hussain, A. Adeel, Multimodal audio-visual information fusion using canonical-correlated Graph Neural Network for energy-efficient speech enhancement, Inf. Fusion 90 (2023) 1–11.
[29] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI, Inf. Fusion 71 (2021) 28–37.
[30] B. Han, Framelets and wavelets, in: Algorithms, Analysis, and Applications, in: Applied and Numerical Harmonic Analysis, Birkhäuser, Cham, 2017, xxxiii.
[31] J. Gao, T. Lyu, F. Xiong, J. Wang, W. Ke, Z. Li, MGNN: A multimodal graph neural network for predicting the survival of cancer patients, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1697–1700.
[32] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, Graph structured network for image-text matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10921–10930.
[33] D. Gao, K. Li, R. Wang, S. Shan, X. Chen, Multi-modal graph neural network for joint reasoning on vision and scene text, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12746–12756.
[34] A. Mafla, S. Dey, A.F. Biten, L. Gomez, D. Karatzas, Multi-modal reasoning graph for scene-text based fine-grained image classification and retrieval, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 4023–4033.
[35] S. Li, H. Liu, Z. Tao, Y. Fu, Multi-view graph learning with adaptive label propagation, in: Proceedings of the IEEE International Conference on Big Data, IEEE, 2017, pp. 110–115.
[36] L.-H. Chen, H. Li, W. Zhang, J. Huang, X. Ma, J. Cui, N. Li, J. Yoo, AnomMAN: Detect anomalies on multi-view attributed networks, Inform. Sci. 628 (2023) 1–21.
[37] Z. Hu, F. Nie, W. Chang, S. Hao, R. Wang, X. Li, Multi-view spectral clustering via sparse graph learning, Neurocomputing 384 (2020) 1–10.
[38] Z. Li, C. Tang, X. Liu, X. Zheng, W. Zhang, E. Zhu, Consensus graph learning for multi-view clustering, IEEE Trans. Multimed. 24 (2021) 2461–2472.
[39] Z. Li, C. Tang, J. Chen, C. Wan, W. Yan, X. Liu, Diversity and consistency learning guided spectral embedding for multi-view clustering, Neurocomputing 370 (2019) 128–139.
[40] P. Goldberg, Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, U. Trautwein, Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction, Educ. Psychol. Rev. 33 (2021) 27–49.
[41] G. Maimaiti, C. Jia, K.F. Hew, Student disengagement in web-based videoconferencing supported online learning: an activity theory perspective, Interact. Learn. Environ. (2021) http://dx.doi.org/10.1080/10494820.2021.1984949.
[42] C.-Y. Ting, W.-N. Cheah, C.C. Ho, Student engagement modeling using bayesian networks, in: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, IEEE, 2013, pp. 2939–2944.
[43] B. Farr-Wharton, M.B. Charles, R. Keast, G. Woolcott, D. Chamberlain, Why lecturers still matter: the impact of lecturer-student exchange on student engagement and intention to leave university prematurely, High. Educ. 75 (2018) 167–185.
[44] C.H. Davies, Student engagement with simulations: a case study, Comput. Educ. 39 (3) (2002) 271–282.
[45] C. Coffrin, L. Corrin, P. de Barba, G. Kennedy, Visualizing patterns of student engagement and performance in MOOCs, in: Proceedings of the 4th International Conference on Learning Analytics and Knowledge, 2014, pp. 83–92.
[46] Ö. Sümer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, E. Kasneci, Multimodal engagement analysis from facial videos in the classroom, IEEE Trans. Affect. Comput. 14 (2) (2023) 1012–1027.
[47] L. Cosmo, A. Kazi, S.-A. Ahmadi, N. Navab, M. Bronstein, Latent-graph learning for disease prediction, in: Proceedings of the 23th International Conference on Medical Image Computing and Computer Assisted Intervention, Springer, 2020, pp. 643–653.
[48] Z. Zhao, Q. Liu, S. Wang, Learning deep global multi-scale and local attention features for facial expression recognition in the wild, IEEE Trans. Image Process. 30 (2021) 6544–6556.
[49] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: Proceedings of the British Machine Vision Conference, 2015.
[50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[51] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM International Conference on Multimedia, 2010, pp. 1459–1462.
[52] S. Schneider, A. Baevski, R. Collobert, M. Auli, wav2vec: Unsupervised pre-training for speech recognition, in: Proceedings of the Interspeech, 2019, pp. 1459–1462.
[53] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186.
[54] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
[55] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2020.
[56] B. Han, T. Li, X. Zhuang, Directional compactly supported box spline tight framelets with simple geometric structure, Appl. Math. Lett. 91 (2019) 213–219.
[57] Y. Xiao, X. Zhuang, Adaptive directional Haar tight framelets on bounded domains for digraph signal representations, J. Fourier Anal. Appl. 27 (2021) 1–26.
[58] J. Li, H. Feng, X. Zhuang, Convolutional neural networks for spherical signal processing via area-regular spherical haar tight framelets, IEEE Trans. Neural Netw. Learn. Syst. (2022) http://dx.doi.org/10.1109/TNNLS.2022.3160169.
[59] Y.-R. Li, R.H. Chan, L. Shen, Y.-C. Hsu, W.-Y. Isaac Tseng, An adaptive directional Haar framelet-based reconstruction algorithm for parallel magnetic resonance imaging, SIAM J. Imaging Sci. 9 (2) (2016) 794–821.
[60] Y.-R. Li, L. Shen, X. Zhuang, A tailor-made 3-dimensional directional Haar semi-tight framelet for pMRI reconstruction, Appl. Comput. Harmon. Anal. 60 (2022) 446–470.
[61] Y. Chen, L. Wu, M. Zaki, Iterative deep graph learning for graph neural networks: Better and robust node embeddings, Adv. Neural Inf. Process. Syst. (2020) 19314–19326.
[62] F. Del Duchetto, P. Baxter, M. Hanheide, Are you still with me? Continuous engagement assessment from a robot's point of view, Front. Robot. AI 7 (2020) 116.
[63] H. Chen, D. Jiang, H. Sahli, Transformer encoder with multi-modal multi-head attention for continuous affect recognition, IEEE Trans. Multimed. 23 (2020) 4171–4183.
[64] V. Thong Huynh, S.-H. Kim, G.-S. Lee, H.-J. Yang, Engagement intensity prediction withfacial behavior features, in: Proceedings of the International Conference on Multimodal Interaction, 2019, pp. 567–571.
[65] K. Wang, J. Yang, D. Guo, K. Zhang, X. Peng, Y. Qiao, Bootstrap model ensemble and rank loss for engagement intensity regression, in: Proceedings of the International Conference on Multimodal Interaction, 2019, pp. 551–556.
[66] S. Parisot, S.I. Ktena, E. Ferrante, M. Lee, R.G. Moreno, B. Glocker, D. Rueckert, Spectral graph convolutions for population-based disease prediction, in: Proceedings of the 20th International Conference on Medical Image Computing and Computer Assisted Intervention, 2017, pp. 177–185.
[67] Y. Huang, A.C. Chung, Edge-variational graph convolutional networks for uncertainty-aware disease prediction, in: Proceedings of the 23th International Conference on Medical Image Computing and Computer Assisted Intervention, 2020, pp. 562–572.
[68] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, 2017.
[69] D. Hu, X. Hou, L. Wei, L. Jiang, Y. Mo, MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2022, pp. 7037–7041.
[70] F. Chen, J. Shao, S. Zhu, H.T. Shen, Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10761–10770.
[71] J. Ma, X. Jiang, S. Xu, X. Qin, Hierarchical temporal multi-instance learning for video-based student learning engagement assessment, in: Proceedings of the 13th International Joint Conference on Artificial Intelligence, 2021, pp. 2782–2789.
[72] T. Baltrusaitis, A. Zadeh, Y.C. Lim, L.-P. Morency, Openface 2.0: Facial behavior analysis toolkit, in: Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, 2018, pp. 59–66.
[73] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[74] J. Bergstra, D. Yamins, D. Cox, Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures, in: Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 115–123.

Ming Li is currently a "Shuang Long Scholar" Distinguished Professor at the Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, China. He received his Ph.D. degree from the Department of Computer Science and IT at La Trobe University, Australia, in 2017. He completed two Postdoctoral Fellowship positions with the Department of Mathematics and Statistics, La Trobe University, Australia, and the Department of Information Technology in Education, South China Normal University, China, respectively. He has published in top-tier journals and conferences, including IEEE TPAMI, Artificial Intelligence, IEEE TKDE, NeurIPS, ICML, IJCAI, etc. He, as a leading guest editor, organized a special issue "*Deep Neural Networks for Graphs: Theory, Models, Algorithms and Applications*" in IEEE Transactions on Neural Networks and Learning Systems. He is an Associate Editor of Neural Networks, Applied Intelligence, Alexandria Engineering Journal, Soft Computing, Neural Processing Letters.

Xiaosheng Zhuang received the bachelor's and master's degrees in mathematics from Sun Yat-sen (Zhongshan) University, Guangzhou, China, in 2003 and 2005, respectively, and the Ph.D. degree in applied mathematics from the University of Alberta, Edmonton, AB, Canada, in 2010. He was a Post-Doctoral Fellow with the Universität Osnabrück, Osnabrück, Germany, in 2011, and Technische Universität Berlin, Berlin, Germany, in 2012. He is currently an Associate Professor with the Department of Mathematics, City University of Hong Kong, Hong Kong. His research interests include applied and computational harmonic analysis, sparse approximation and directional multiscale representation systems, deep and machine learning, and signal/image processing.

Lu Bai received PhD degree from the University of York, U.K. He was a recipient of the National Award for Outstanding Self-Financed Chinese Students Study Aboard by the China Scholarship Council, in 2015, and the Best Student Paper Award of the International Conferences ICIAP 2015, the Best Scientific Paper Award of the International Conference ICPR 2018, and the Outstanding Paper Award of the International Conference IEEE IEEM 2019. He is now supported by the National Excellent Young Scientist Fund of NSFC. He was selected as one of the 2022 Baidu Global Top Chinese Young Scholars in Artificial Intelligence. He is now a Full Professor at the School of Artificial Intelligence, Beijing Normal University, Beijing, China, as well as Central University of Finance and Economics, Beijing, China. He has published more than 100 journal and conference papers, including TPAMI, TKDE, ICML, IJCAI, ICDE, etc. He is currently a member of the editorial board of the journals Pattern Recognition and IEEE Transactions on Neural Networks and Learning Systems.

Weiping Ding received the Ph.D. degree in Computer Science, Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013. From 2014 to 2015, he is a Postdoctoral Researcher at the Brain Research Center, National Chiao Tung University, Hsinchu, Taiwan. In 2016, He was a Visiting Scholar at National University of Singapore, Singapore. From 2017 to 2018, he was a Visiting Professor at University of Technology Sydney, Ultimo, NSW, Australia. He is a professor of School of Information Science and Technology, Nantong University, China. His research interests include deep neural networks, machine learning, and granular data mining. He has published over 200 articles in international journals, such as IEEE TFS, IEEE TNNLS, IEEE TCYB, IEEE Transactions on Evolutionary Computation. His fifteen authored/co-authored papers have been selected as ESI Highly Cited Papers. He serves on the Editorial Board of Information Fusion, Engineering Applications of Artificial Intelligence and Applied Soft Computing. He serves as an Associate Editor of IEEE TNNLS, IEEE TFS, IEEE/CAA Journal of Automatica Sinica, IEEE Transactions on TITS, IEEE TETCI, IEEE Transactions on AI, Information Sciences, Neurocomputing, and so on.