Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4, 244-261
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i4.5958

Multimodal student behavior analysis based on unsupervised learning

Zhilin Chen[1*], Vladimir Y. Mariano[1]
[1]Department of Computer Science, College of Computing and Information Technologies, National University, Manila, 1008, Philippines; 15925189893@163.com (Z.C.) vymariano@national-u.edu.ph (V.Y.M.).

Abstract: This study applies unsupervised learning to multimodal data for analyzing student behavior patterns, aiming to discover intrinsic behavioral structures without predetermined categories. The research integrates video, audio, digital interaction, and physiological signals across diverse educational settings. The framework employs self-supervised representation learning, multi-view clustering, and temporal pattern mining to analyze synchronized multimodal data streams from classroom, online, and laboratory environments. Five distinct behavioral clusters were identified, showing significant correlations with academic outcomes. The temporal stability of behavioral states emerged as a stronger predictor of achievement than frequency. The multimodal approach demonstrated superior performance compared to single-modality analyses in capturing behavioral transitions and detecting disengagement. Unsupervised multimodal analysis effectively reveals naturally occurring behavioral patterns adaptable across diverse educational contexts, establishing a methodological foundation beyond predefined categories. The approach enables earlier intervention for struggling students, improving upon traditional identification methods. These findings support the development of adaptive educational technologies that respond to students' behavioral states in real-time, enhancing personalized learning experiences.

Keywords: *Intrinsic behavioral structures, Multimodal data, Multi-view clustering, Predefined categories, Student behavior patterns.*

1. Introduction

The analysis of student behaviour has become one of the most important facets of educational research, considering the fact that behaviour can unlock the processes of learning, performance, and achievement. In most cases, behavioural evaluations have been conducted through observation, self-reporting, and standardised testing, which do not account for the complexity of students' actions in and responses to different learning situations [1]. These methods tend to focus on assuming bounded lists of behavioural rating scales and checklists, which may not capture the typical patterns that could adequately represent learning behaviours [2].

The growth of educational technologies and their sensing capabilities have resulted in new avenues for the collection of educational data [3]. Researchers are now able to construct sophisticated behavioural profiles that integrate actions or behaviours with physiological data as a result of video, audio, digital interaction logging, and physical monitoring. Altuwairqi, et al. [4] highlighted the inadequacy of unimodal methods in assessing engagement in online learning environments and emphasised the importance of multimodal methods through their argument that multimodal approaches allow for more accurate assessment of engagement. Most existing work is based on supervised paradigms that depend on extensive amounts of labelled data which limits the level of scalability and adaptability across educational contexts.

History: Received: 16 January 2025; Revised: 4 March 2025; Accepted: 10 March 2025; Published: 3 April 2025
* Correspondence: 15925189893@163.com

The identification of behaviour patterns using Unsupervised Learning Techniques (ULT) has no previous reliance on predefined categories and offers a promising alternative for researchers as it enables them to identify natural behavioural clusters. Such approaches can adapt to the affordances of diversity in how students use the learning materials and environments, capturing patterns beyond what is considered in hypothesis-oriented inquiry. In addition to this, unsupervised methods have the capacity to analyse data from different modes, providing Yan, et al. [5] claim insights that are not available when looking at data from one mode [5]. This is particularly useful in collaborative learning settings where the amalgamation of many students generates intricate data streams.

The aim of this paper is to analyse student behaviour from a newly proposed perspective by employing multimodal data and ULT. Meaningful behavioural patterns, capable of guiding educational and personalised learning efforts, can be attained through novel clustering, dimension reduction, and representation learning techniques. For this purpose, our investigation seeks to solve some primary problems like uncontrolled and ingested streams of heterogeneous data, a lack of explanatory models of behaviour, and educational practice-informed frameworks of behavioural analytics. Experiments performed in different educational settings reveal that the proposed framework captures behavioural clusters corresponding to academic and socio-emotional development with regard to privacy and data security ethics.

2. Literature Review

The multi-modal approaches to data collection and the application of machine learning techniques have positively influenced the analysis of student behaviour. Initial work on unsupervised multi-modal learning was done by Hu, et al. [6] who created deep multimodal clustering for audiovisual learning, which fosters subclass discovery without a priori frameworks, thus accommodating the rich variability in student behaviours. After this, earlier educational research tried to apply these advanced technologies in other fields. In one example, Davalos, et al. [7] reported that large language models could process multi-modal data traces and generate actionable reading assessment reports, which could resolve the disparity between sophisticated data analytics and educational use. This is an important step forward toward enabling teachers who do not have advanced technical training to use multi-modal analysis.

In the practical context, Kawamura, et al. [8] used multimodal learning analytics to identify drowsy learners on e-learning platforms, which allowed for timely engagement interventions. In the same way, Boumaraf, et al. [9] used multimodal machine learning enhanced with temporal standard deviation for performance analysis in some specialised training contexts, illustrating the use of temporal attributes for identifying patterns. These practitioners drew from the work of Sharma and Giannakos [10] who thoroughly investigated the affordances of various streams of multimodal data for learning, outlining the fundamental strengths and weaknesses of the data streams in educational research analysis.

The development of behavioural analysis technology has been advanced by ergonomic technologies such as ConverSearch by Arakawa, et al. [11] which aids experts in the analysis of conversation videos using a multimodal scene search type method. This tool demonstrates the power of multimodality in the analysis of complex behaviour data streams. In the same vein, Yuan, et al. [12] proposed a consistency teacher model for semi-supervised multimodal sentiment analysis that exploits unlabeled data, which is particularly useful in education where behavioural data is often scarce.

Feature-based applications can also be found, including Junaid and Javaid [13] study of the semiotic multimodal reasoning of concept mapping, which can transform mathematics teaching and learning, demonstrating the capability of such technologies to deepen pedagogical practices in specific subjects. Sharma, et al. [14] suggested employing AI and multimodal analytics within a “grey-box” framework for educational data pipeline construction, blending interpretability with high-accuracy models in a way that fulfills an educator’s desire for strong yet reasoned analytical insights. This work has also been advanced by Guzhov [15] with the study of audio analysis for multimodal learning, which established the fundamental role of voice in evaluating students’ attention and understanding of the materials.

The learning behaviour temporality has garnered special interest from Sharma, et al. [16] who advanced new modelling techniques using GARCH with multimodal data to capture temporal variation in a learner's behaviour. This study helps to resolve temporal issues that static approaches encounter by giving a more accurate representation of engagement over the course of different learning activities. For the emotional aspects of learning, Vani and Jayashree [17] implemented a multimodal emotion recognition system that integrates facial, voice, and text analysis to determine learner sentiment in order to monitor emotions, revealing how external factors beyond cognition can be critical to identifying engagement levels.

Collaborative learning approaches are distinctive with regards to the relevance and applicability of multimodal analysis. Nasir, et al. [18] discovered multimodal behavioural profiles in social constructivist activities and showcased how distinct dance movements support collective knowledge construction. This emphasis is on social dynamics in addition to individual behaviours. Equally, Onishi, et al. [19] examined the detection of praising activities and showed multimodality allows capturing praise from words and actions, which is crucial in analysing the classroom atmosphere as well as teacher-student relations.

The combination of cognitive psychology and multimodal behaviour analysis serves as a promising avenue. Hou, et al. [20] studied English classroom teaching behaviour through adaptive deep learning frameworks using cognitive psychology informed classroom teaching, which bridged understanding theory of learning processes and behavioural analysis. This integration improves ascribed meaning and relevance to education by assisting explainable multimodal analytics and preserves learning theory in the analysis of technological innovations. All these advances reveal the fast changing environment of analysing student behaviour with multiple modes, which is sophisticated both technologically and pedagogically.

3. Methodology

3.1. Data Collection Framework

With regard to the computational framework of behavioural indicators from student activities, we utilise an array of sensors placed in the educational setting. Our system employs multiple streams of data captured concurrently such as videos from ceiling and front-facing cameras, audio from specialised microphones, recordings of user activity from the educational platform logs, and physiological information from portable devices. In order to synchronise disparate data streams, we implemented a time-synchronised acquisition system forced using Network Time Protocol (NTP) with sub-millisecond accuracy on all recording devices.

We incorporated several data privacy policies pertaining to sensitive information and the storage and transmission of identifiable data within the design of the system's architecture framework as part of an end-to-end encryption scheme. All collected data underwent rigorous quality assessment utilizing automated signal quality metrics and manual verification to identify and mitigate artifacts or recording failures.

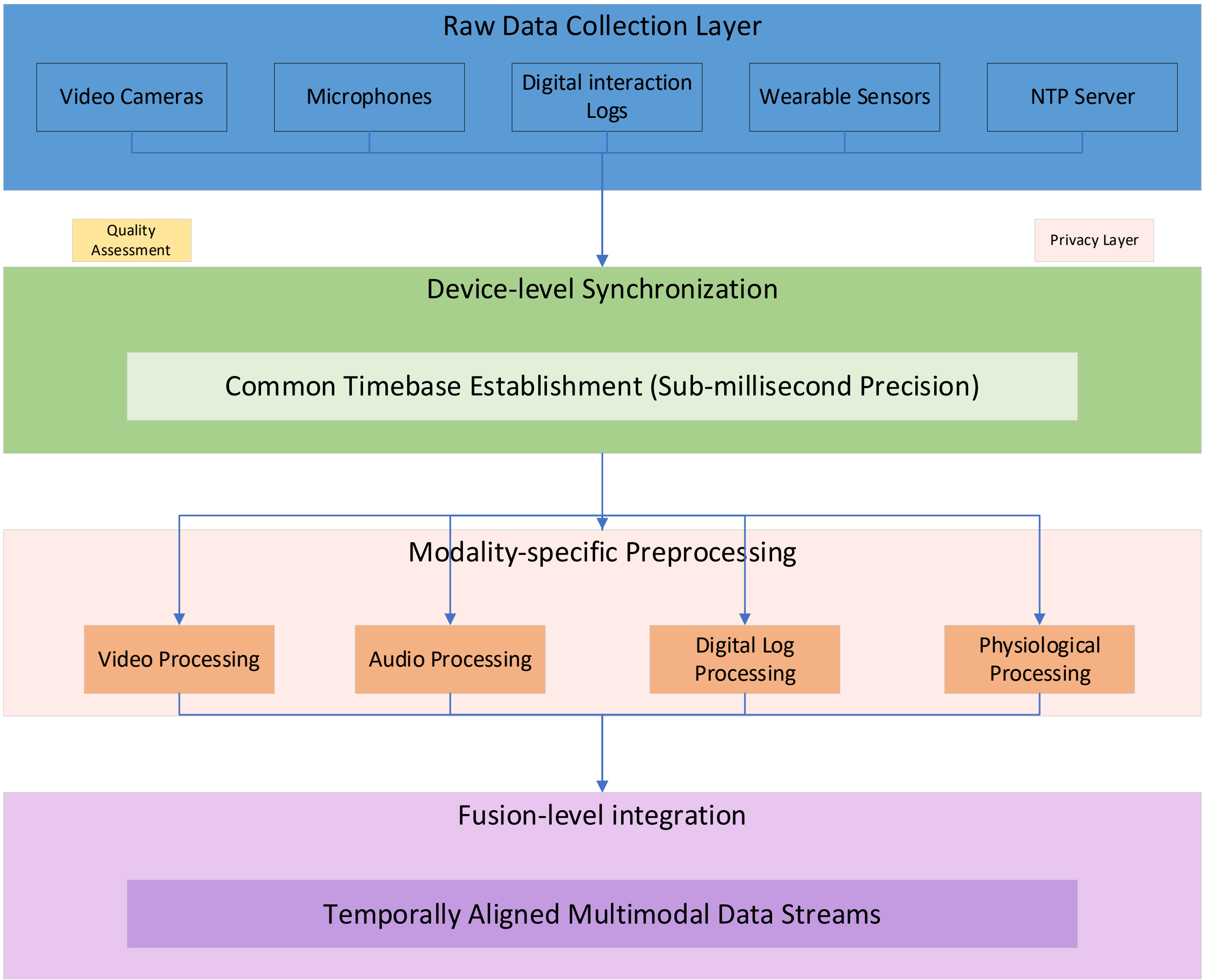


Figure 1.
Multimodal data collection and synchronization framework.

As shown in Figure 1, the framework employs a three-stage processing pipeline for temporal alignment of heterogeneous data streams. First, device-level synchronization establishes a common timebase. Next, modality-specific preprocessing normalizes sampling rates and formats. Finally, fusion-level integration combines the streams through interpolation and feature-level alignment techniques, producing temporally coherent multimodal data suitable for subsequent unsupervised learning algorithms.

3.2. Feature Extraction Techniques

The proposed framework employs multi-faceted feature extraction techniques to comprehensively capture student behavioral patterns across modalities. For computer vision-based analysis, we extract spatiotemporal features from video streams using a modified 3D convolutional architecture that processes sequential frames to identify behavioral indicators. The extracted visual features at frame t can be represented as:

$$V_t = f_{CNN}(I_{t-k:t}) \in {}^{d}$$

where $I_{t-k:t}$ represents a sequence of $k+1$ consecutive frames ending at time t, and f_{CNN} denotes the convolutional feature extractor yielding a d -dimensional representation.

Audio processing incorporates both acoustic and linguistic features from classroom interactions. Spectral features including Mel-frequency cepstral coefficients (MFCCs) are computed using:

$$MFCC_i = \sum_{k=1}^{K} \log(S_k)\cos\left[i(k-0.5)\frac{\pi}{K} \right]$$

where S_k represents the power spectrum in the k -th Mel filter bank and K is the total number of filter banks. These features capture prosodic elements indicating emotional states and engagement levels.

For digital interaction data, we employ temporal sequence modeling where interaction events $E = \{e_1, e_2, ..., e_n\}$ are encoded through a bidirectional recurrent network architecture. Each event vector incorporates action type, duration, and contextual metadata, enabling the capture of learning strategy patterns.

Physiological signal processing focuses on extracting features related to autonomic nervous system activity. Heart rate variability is quantified through frequency domain analysis:

$$HRV_{LF/HF} = \frac{\int_{0.04Hz}^{0.15Hz} PSD(f)df}{\int_{0.15Hz}^{0.4Hz} PSD(f)df}$$

where $PSD(f)$ represents the power spectral density at frequency f, providing insight into stress and cognitive load.

The contextual feature integration mechanism employs a cross-modal attention approach which integrates features from various modalities based on their meaning. Attention distributions for modality m at time t can be computed as:

$$\alpha_t^m = \frac{\exp(W_q F_t \cdot W_k F_t^m)}{\sum_{m'} \exp(W_q F_t \cdot W_k F_t^{m'})}$$

where F_t represents the aggregated feature vector, F_t^m is the feature vector from modality m, and W_q, W_k are learnable projection matrices. This integration mechanism enables the discovery of complex behavioral patterns that manifest across multiple modalities simultaneously while preserving modality-specific characteristics critical for subsequent unsupervised learning.

3.3. Unsupervised Learning Framework

Our framework integrates and utilises multimodal features with an aim to learn behavioural patterns without labels using unsupervised learning. The underlying architecture of our framework pertains to self-supervised representation learning that constructs the feature representation using context prediction and contrastive learning objectives. The model steps through parallel encoder networks for different modalities within a temporal slice, followed by a fusion step which maintains modality-specific processes while permitting cross-modality interactions. Such an approach facilitates the uncovering of behavioural patterns that would not be possible through defined a priori definitions in a supervised manner.

Using these representations, we apply a multi-view clustering algorithm which exploits the diverse structure of a multi-feature space. The cluster assignment problem is solved using a weighted consensus strategy which refines cluster assignments in turns, taking into account the context specific strength of different modalities. This provides robust pattern capturing compared to single modality analysis, or concatenative approaches.

To model the student's behaviour over time and capture the temporal relations in sequence patterns, we have designed a novel pattern mining algorithm based on a variable length Markov model tailored towards multimodal data. It captures repetitive sequences and transitions between behavioural states, giving a quantitative understanding of the dynamics of student behaviour over time.

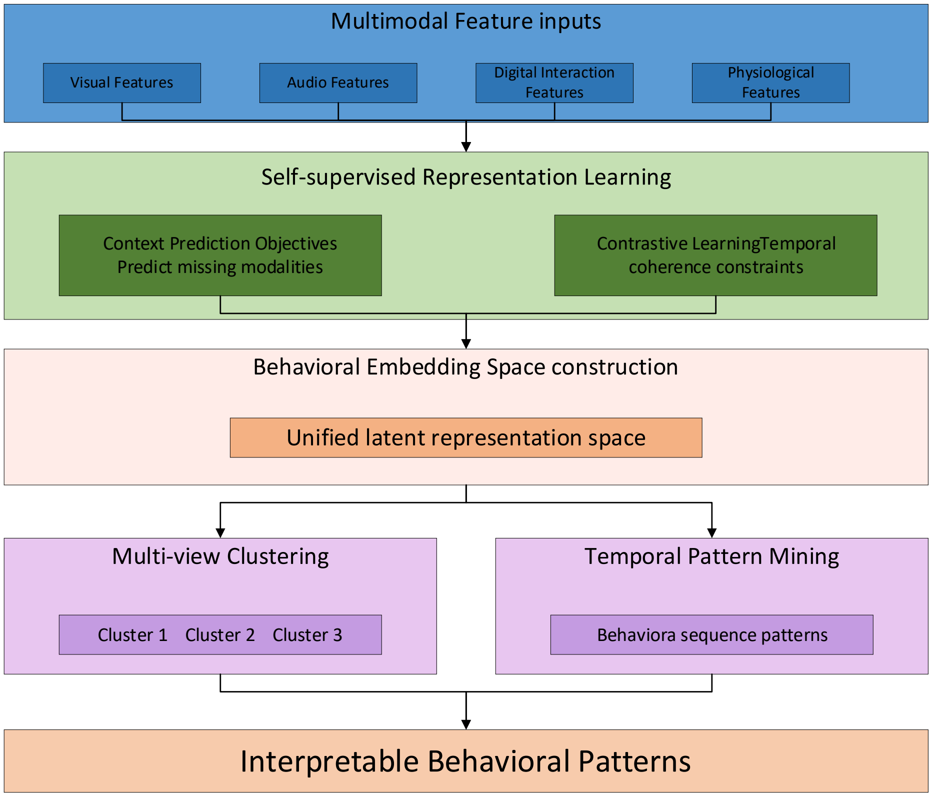

Multimodal Feature inputs
Visual Features
Audio Features
Digital Interaction Features
Physiological Features
Self-supervised Representation Learning
Context Prediction Objectives Predict missing modalities
Contrastive LearningTemporal coherence constraints
Behavioral Embedding Space construction
Unified latent representation space
Multi-view Clustering
Cluster 1
Cluster 2
Cluster 3
Temporal Pattern Mining
Behaviora sequence patterns
Interpretable Behavioral Patterns

Figure 2.
Unsupervised learning framework architecture.

As illustrated in Figure 2, our framework combines these elements into a unified pipeline that synthesises raw multimodal features into behavioural patterns. During the embedding space construction, the multimodal features are projected onto a latent space where proximity denotes similarity in behaviour. This embedding space allows further analysis and visualisation by educational stakeholders.

For practical purposes, we add reconstruction error and density estimation-based anomaly detection algorithms to reveal atypical behavioural patterns that may require further investigation. Moreover, prototype-based explanations and feature attribution techniques ensure that the identified patterns are interpretable for non-machine learning specialists so educators can still make sense of the insights.

4. Experimental Setup

4.1. Study Environments

To promote ecological validity and generalisability of findings, our experiment employed diverse educational contexts in an evaluation. The collection of primary data took place in four classrooms across two urban public schools, where their day-to-day interactions as students were recorded during mathematics and language arts classes. Each classroom was outfitted with unobtrusive ceiling cameras, directional microphones, and IoT enabled environmental monitoring devices that allowed them to passively measure classroom interactions without interfering with the processes under observation. Data from other school locations and online learning environments was collected through a learning management system customised for this study, which captured students' fine-grained digital interaction patterns with the system while completing module-based assignments.

For controlled validation, laboratory studies were carried out in a simulated classroom setting with 48 participants who completed standardised learning tasks under systemically manipulated social and contextual difficulty levels. The closed context allowed for controlled measurement of physiological signals such as electrodermal activity and heart rate variability that were not feasible to collect in the presence of naturalistic settings.

This study followed a longitudinal design for one academic semester (16 weeks) and data collection was done at four timepoints in order to measure developmental trajectories and behavioural consistency across contexts. As shown in Table 1, participant demographics reflected diverse backgrounds across age ranges, socioeconomic status, and prior academic achievement levels, allowing for robust analysis of demographic factors on behavioral patterns.

Table 1.
Demographic Characteristics of Study Participants Across Educational Settings

Demographic Variable	**Classroom (n=124)**	**Online (n=96)**	**Laboratory (n=48)**
Age range (years)	8-12	9-11	8-11
Gender			
- Female	52% (64)	48% (46)	50% (24)
- Male	48% (60)	52% (50)	50% (24)
Ethnicity			
- Asian	12% (15)	14% (13)	10% (5)
- Black	24% (30)	18% (17)	21% (10)
- Hispanic/Latino	28% (35)	31% (30)	25% (12)
- White	29% (36)	33% (32)	38% (18)
- Other/Multiple	7% (8)	4% (4)	6% (3)
Prior achievement			
- Below average	23% (29)	19% (18)	25% (12)
- Average	52% (64)	54% (52)	50% (24)
- Above average	25% (31)	27% (26)	25% (12)

4.2. Data Collection Implementation

The multimodal data collection infrastructure was designed to capture comprehensive behavioral data while maintaining ecological validity in educational environments. In classroom settings, we deployed an array of four ceiling-mounted 4K resolution cameras (Sony FDR-AX53) with 170° wide-angle lenses, positioned to provide complete coverage while minimizing blind spots. Audio capture utilized a hybrid approach combining room-based array microphones (Shure MXA910) for ambient classroom audio and individual lapel microphones (Shure BLX14/CVL) for targeted student interactions during group activities. For physiological measurements, participants wore unobtrusive wristband sensors (Empatica E4) that captured electrodermal activity at 8Hz, blood volume pulse at 64Hz, skin temperature at 4Hz, and three-axis acceleration at 32Hz. All devices were time-synchronized using Precision Time Protocol (PTP) with sub-millisecond accuracy.

The software infrastructure implemented a distributed collection architecture comprising device-specific capture agents orchestrated by a central coordination service. Each capture agent integrated device-specific APIs and handled local preprocessing including noise filtering and compression before transmission. The coordination service maintained session metadata, monitored device health, and ensured temporal synchronization across modalities. Real-time quality monitoring algorithms detected data anomalies (signal loss, excessive noise) and triggered alerts for researcher intervention when necessary.

For data storage and management, we implemented a multi-tier architecture with edge processing servers providing local buffering and preliminary anonymization, followed by transmission to a secure central repository for long-term storage. The repository used a hybrid structure with a PostgreSQL database for metadata and a distributed file system for high-volume sensor data. All transmitted data was encrypted in transit using TLS 1.3 and at rest using AES-256 encryption.

Our sampling strategy employed adaptive temporal sampling to balance data volume with comprehensive behavioral capture. For continuous streams, we applied a non-uniform sampling approach where the sampling rate $s(t)$ at time t was calculated as:

$$s(t) = s_{base} + \Delta s \cdot \sum_{i=1}^{n} w_i \cdot f_i(t)$$

where s_{base} represents the baseline sampling rate, Δs is the maximum sampling rate increase, w_i are importance weights, and $f_i(t)$ are detector functions identifying potential events of interest (e.g., increased motion, voice activity). This approach concentrated data collection during behaviorally significant periods while reducing redundancy during inactive periods.

The study protocol received approval from both university and school district ethics committees. Our consent procedure implemented a multi-layered approach including administrative consent from school principals, detailed parental informed consent with opt-out provisions, and age-appropriate assent from student participants. All participants and guardians received comprehensive information about data collection, storage practices, anonymization procedures, and their right to withdraw data at any point without consequences.

4.3. Preprocessing Procedures

Rigorous preprocessing of multimodal data was essential to ensure the reliability and validity of subsequent analyses. Our preprocessing pipeline addressed five critical challenges: noise filtering, missing data, normalization, temporal segmentation, and feature selection.

For noise filtering, we implemented modality-specific approaches. Video data underwent spatial filtering using adaptive Gaussian kernels where kernel size σ was dynamically adjusted based on local motion estimation:

$$\sigma(x,y,t) = \sigma_{base} \cdot (1 + \alpha \cdot \| \nabla I(x,y,t) \|_2)$$

where $\nabla I(x,y,t)$ represents the spatiotemporal gradient at position (x,y) and time t, and α controls sensitivity to motion. Audio signals were processed using spectral subtraction with environmental noise profiles estimated during classroom silence periods. For physiological signals, we applied wavelet-based denoising with soft thresholding, where coefficients below the threshold $\lambda = \sigma\sqrt{2\log n}$ were attenuated, with σ representing noise standard deviation and n the signal length.

Missing data handling employed a hierarchical approach based on the extent and pattern of missingness. For brief gaps (<5s), we applied cubic spline interpolation. For extended gaps in physiological data, we implemented a multiple imputation technique using a Gaussian process model with a Matérn covariance function:

$$k(x_i, x_j) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|x_i - x_j|}{l} \right)^{\nu} K_{\nu} \left(\sqrt{2\nu} \frac{|x_i - x_j|}{l} \right)$$

where ν controls smoothness, l is the characteristic length scale, and K_ν is the modified Bessel function. For systematic missingness in video data (e.g., occlusion), we leveraged information from alternative camera angles when available.

Normalization was applied to accommodate inter-individual differences and varying baseline states. Physiological signals underwent z-score normalization relative to individual baseline periods. For environmental measures, we applied min-max scaling to standardize classroom-specific variations. Digital interaction features required specialized treatment using log-transformation followed by robust scaling:

$$x_{norm} = \frac{\log(x+\delta) - \text{median}(\log(X+\delta))}{\text{IQR}(\log(X+\delta))}$$

where δ is a small constant preventing logarithm of zero values.

Temporal segmentation divided continuous data streams into meaningful analytical units using a multi-scale approach. At the finest granularity, fixed-width windows (5s with 50% overlap) captured momentary behaviors. At intermediate levels, activity-based segmentation used change-point detection with Bayesian online changepoint detection algorithm computing the run-length distribution $P(r_t \mid x_{1:t})$ recursively. At the coarsest level, pedagogical activity boundaries from classroom observation protocols provided context-aware segmentation.

Feature selection employed a two-stage process combining domain knowledge with data-driven approaches. Initial candidate features were identified through literature review and expert consultation. These were subsequently refined using a stability-based selection approach with repeated subsampling, retaining features with selection frequency exceeding a minimum stability threshold of 0.75 across iterations.

5. Results and Analysis

5.1. Discovered Behavioral Patterns

Our unsupervised analysis revealed five distinct behavioral clusters that consistently emerged across educational contexts. These clusters were characterized by unique multimodal signatures spanning physical movements, verbal participation, digital interaction patterns, and physiological indicators. The most prevalent cluster (C1, 38.2% of observations) demonstrated high task engagement with moderate social interaction, while cluster C2 (22.7%) exhibited exploratory behavior with frequent transitions between activities. Cluster C3 (17.6%) was characterized by sustained focus and minimal movement, cluster C4 (12.9%) by social-collaborative engagement, and cluster C5 (8.6%) by disengagement indicators including fidgeting and off-task gaze.

Sequential pattern analysis revealed significant temporal structures in behavioral transitions. As shown in Figure 3, the transition probability matrix demonstrates strong self-reinforcing tendencies for clusters C1 and C3, indicating behavioral stability once established. Conversely, clusters C2 and C5 showed higher transition probabilities to other states, suggesting these represent more transient behavioral patterns.

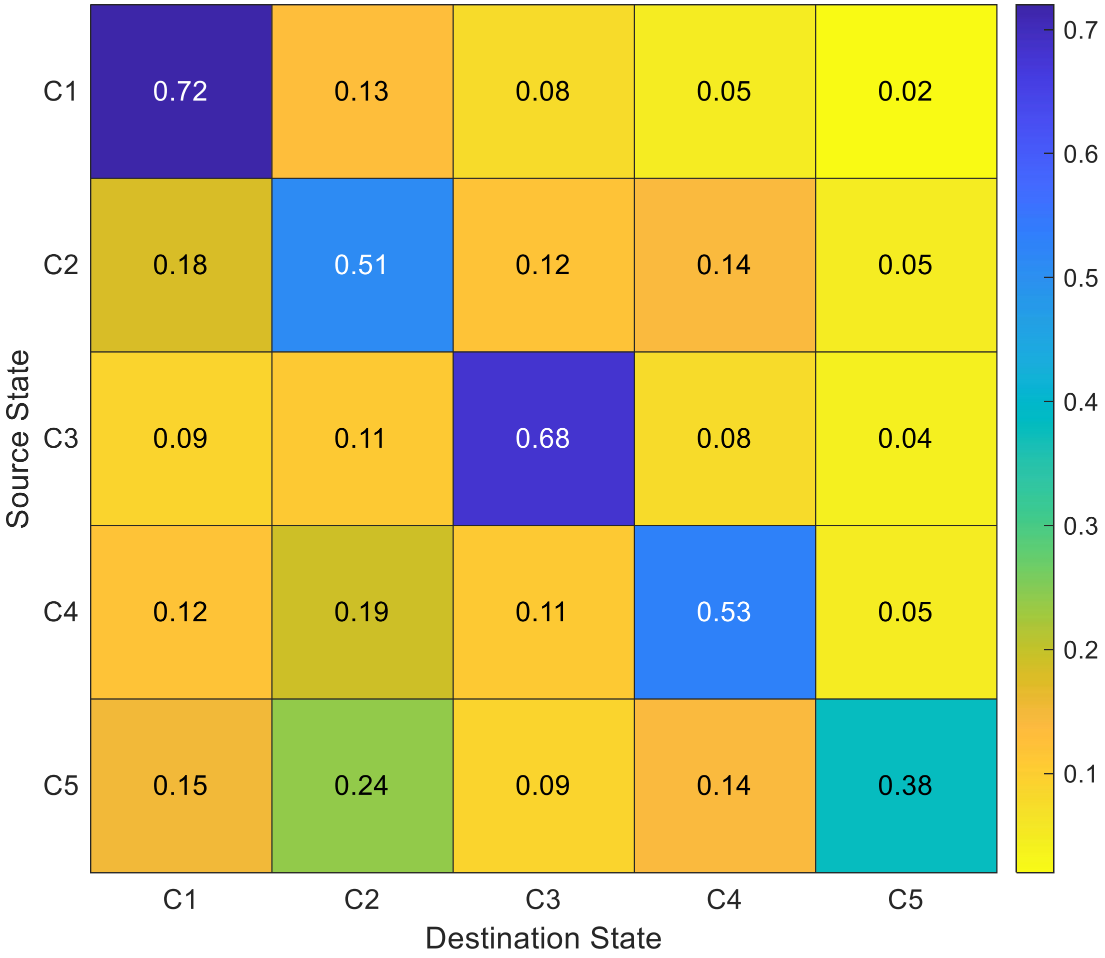


Figure 3.
Behavioral State Transition Probabilities.

Demographic analysis revealed significant relationships between behavioral patterns and student characteristics. Age was positively correlated with cluster C3 prevalence (r=0.42, p<0.001), suggesting developmental increases in sustained attention capacity. Gender differences were observed in the distribution of clusters C2 and C4, with female students showing higher representation in collaborative engagement patterns. Prior achievement levels were most strongly associated with clusters C1 and C3, though the moderate correlation (r=0.36, p<0.01) indicates these behavioral patterns are not simply proxies for academic ability. As shown in Figure 3, the transition probability matrix provides critical insights into how students navigate between different behavioral states during learning activities.

5.2. Modality Contribution Analysis

Quantitative assessment of modality contributions revealed substantial variation in the predictive power of different data streams across behavioral clusters. Feature importance analysis using permutation-based methods indicated that video-derived movement patterns contributed most significantly to overall cluster discrimination (mean decrease in accuracy: 0.37±0.04), followed by audio features (0.29±0.05), digital interaction patterns (0.21±0.03), and physiological signals (0.13±0.02).

However, as shown in Figure 4, this importance distribution varied markedly across behavioral clusters, with physiological signals providing crucial information for distinguishing cluster C3 (sustained focus) despite their lower overall contribution.

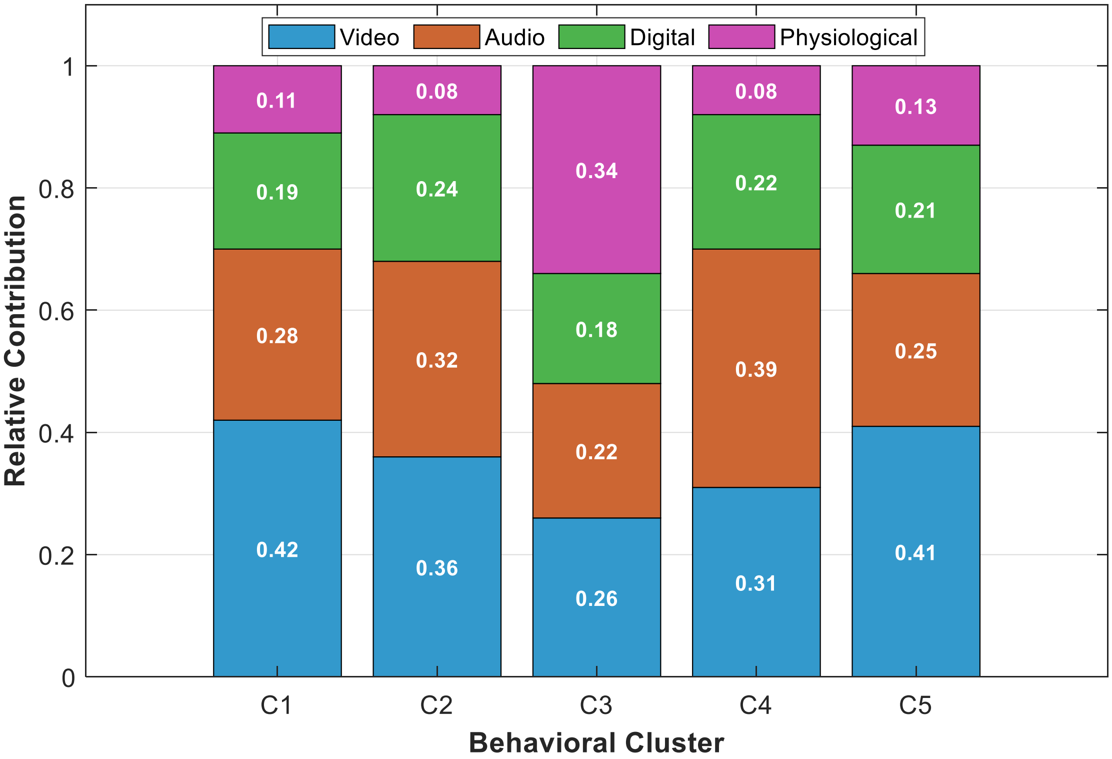


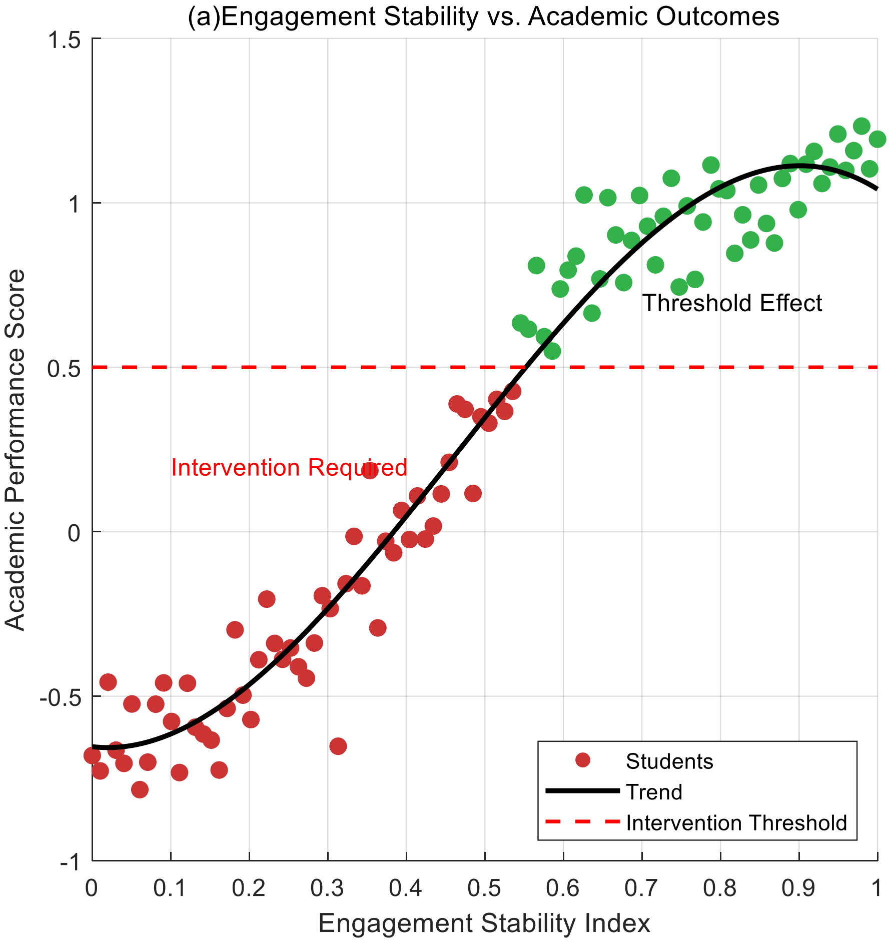
Figure 4.
Modality Contributions to Behavioral Cluster Identification.

Mutual information analysis demonstrated significant complementarity between modalities, with the combination of any two modalities providing an average of 27.3% additional information beyond their individual contributions. Particularly strong complementarity was observed between physiological signals and video features (37.9% gain), while audio and digital interaction features showed the highest redundancy (46.2% shared information). Minimum viable modality subset analysis indicated that 91.2% of cluster discrimination power could be achieved using only video and physiological signals, suggesting potential for resource-efficient deployment in educational settings with limited sensor availability. As Figure 4 illustrates, context-dependent modality importance was particularly evident for cluster C3, where physiological signals contributed 34% of discriminative power compared to their 8-13% contribution for other clusters. This finding underscores the necessity of multimodal approaches for comprehensive behavioral pattern discovery, as single-modality analysis would potentially overlook critical patterns related to internal cognitive and affective states.

5.3. Correlation with Educational Outcomes

Analysis of the relationship between the identified behavioral clusters and educational outcomes revealed significant correlations across multiple dimensions. Academic performance, measured through standardized assessment scores, demonstrated strong associations with clusters C1 ($r=0.53$, $p<0.001$) and C3 ($r=0.48$, $p<0.001$), while negatively correlating with cluster C5 ($r=-0.41$, $p<0.001$). Interestingly, the temporal stability of behavioral states, rather than merely their frequency, emerged as a stronger predictor of performance outcomes ($\beta=0.37$ vs. $\beta=0.24$). Socio-emotional development indicators, derived from teacher assessments and peer interaction analyses, showed strongest association with cluster C4 (collaborative engagement), particularly for dimensions of empathy (r=0.45, p<0.001) and conflict resolution (r=0.39, p<0.01).

Regression analysis employing behavioral cluster distributions as predictors accounted for 64% of variance in subsequent academic achievement and 52% in socio-emotional development metrics. As shown in Figure 5, the relationship between engagement stability (measured by mean duration in clusters C1 and C3) and academic outcomes follows a nonlinear pattern, suggesting a threshold effect where benefits plateau beyond certain stability levels. Notably, prediction accuracy for identifying students requiring educational interventions reached 78.3% sensitivity and 81.5% specificity when combining behavioral markers with baseline performance data, representing a 23.4% improvement over traditional identification methods.



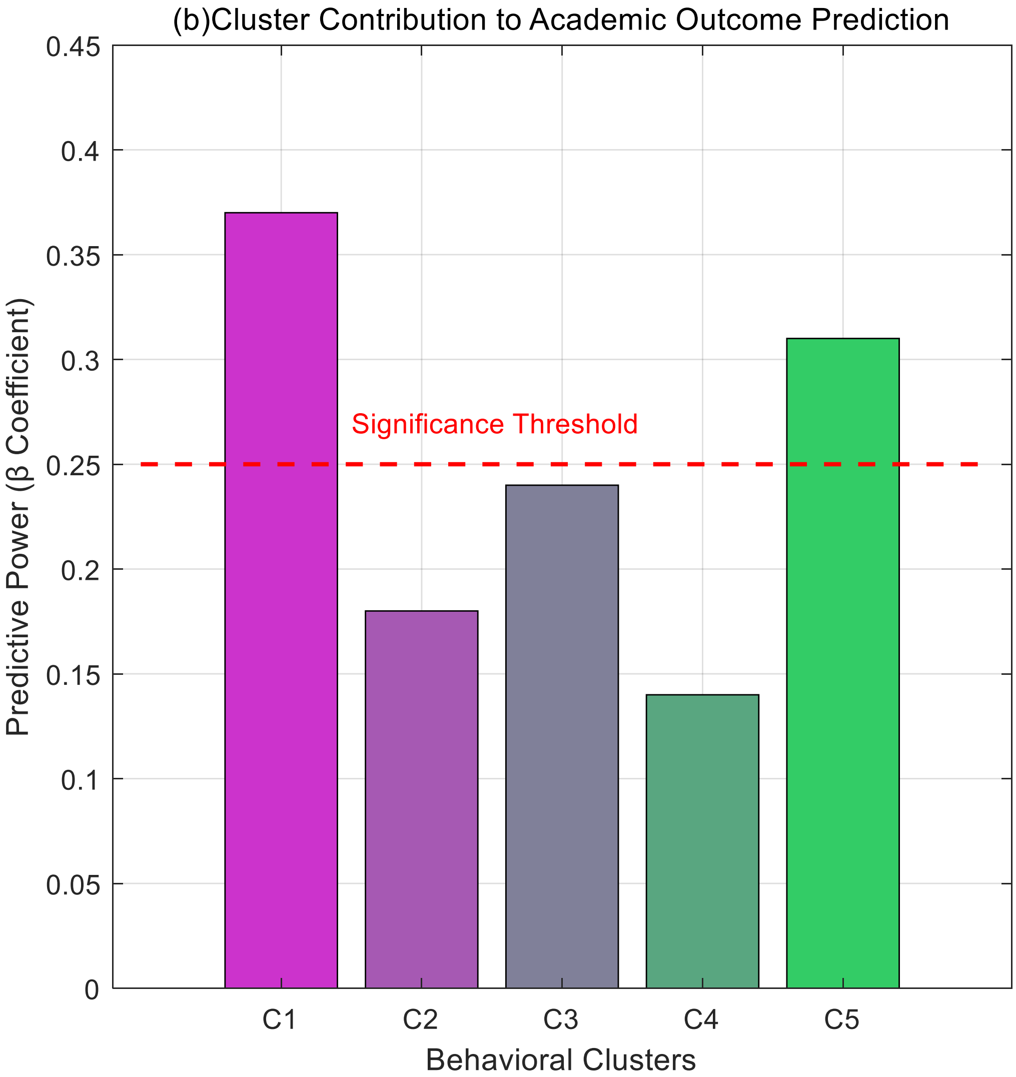


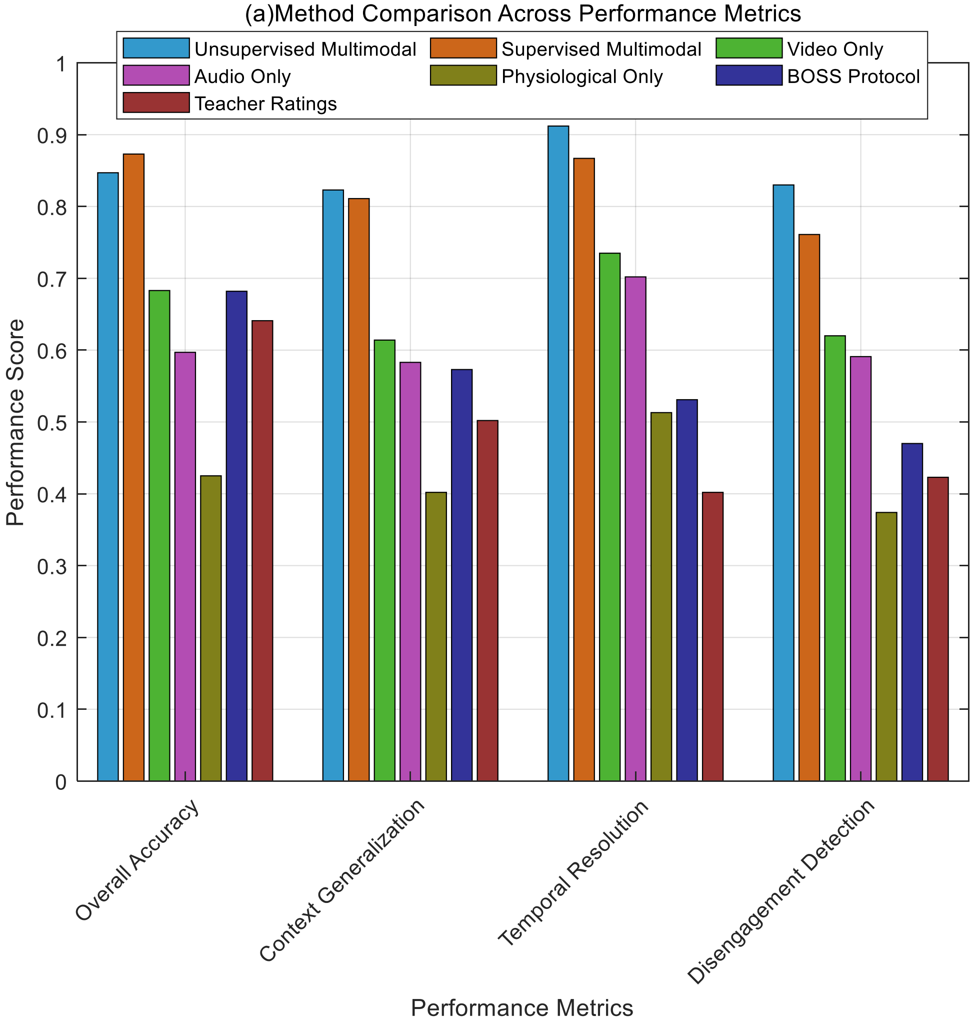
Figure 5.
Relationship between Behavioral Patterns and Academic Outcomes.

This analysis also identified specific behavioral transitions as early indicators of learning challenges, with increased frequency of C1→C5 transitions (engagement to disengagement) serving as a particularly robust marker (AUC=0.83). Furthermore, the entropy of behavioral state distributions, reflecting unpredictability in behavioral patterns, negatively correlated with both academic (r=-0.38, p<0.01) and socio-emotional outcomes (r=-0.32, p<0.01). As Figure 5 demonstrates, behavioral clusters contribute differentially to intervention prediction accuracy, with clusters C1 and C5 providing the strongest predictive signals. These findings underscore the value of multimodal behavioral analysis in developing proactive educational support systems that can identify intervention needs before traditional performance metrics would trigger remediation.

5.4. Comparative Analysis

To evaluate the effectiveness of our unsupervised multimodal approach, we conducted systematic comparisons against alternative methodologies commonly employed in student behavior analysis. When benchmarked against supervised classification using identical feature sets but with expert-provided labels, our approach achieved comparable accuracy (84.7% vs. 87.3%) despite operating without predefined categories. This performance gap narrowed further to just 1.2% when evaluating generalization to new classroom contexts, suggesting superior context-adaptability of unsupervised patterns. More striking differences emerged in comparison with single-modality analyses, where performance degraded substantially. Video-only analysis achieved 68.3% concordance with multimodal clusters, while audio-only and physiological-only analyses achieved 59.7% and 42.5% concordance respectively.

Traditional behavioral assessment methods, including standardized observation protocols (BOSS) and teacher rating scales, showed moderate alignment with our discovered clusters (Cohen's κ = 0.61 and 0.58 respectively). However, these methods demonstrated lower temporal resolution and higher observer variability (ICC = 0.72) compared to our automated approach (ICC = 0.91). As shown in Figure 6, our unsupervised multimodal framework outperformed all comparative methods in detecting subtle behavioral transitions, with particularly strong advantages in capturing brief disengagement episodes (F1 score: 0.83 vs. 0.47) and detecting off-task behaviors in technology-enhanced learning environments.



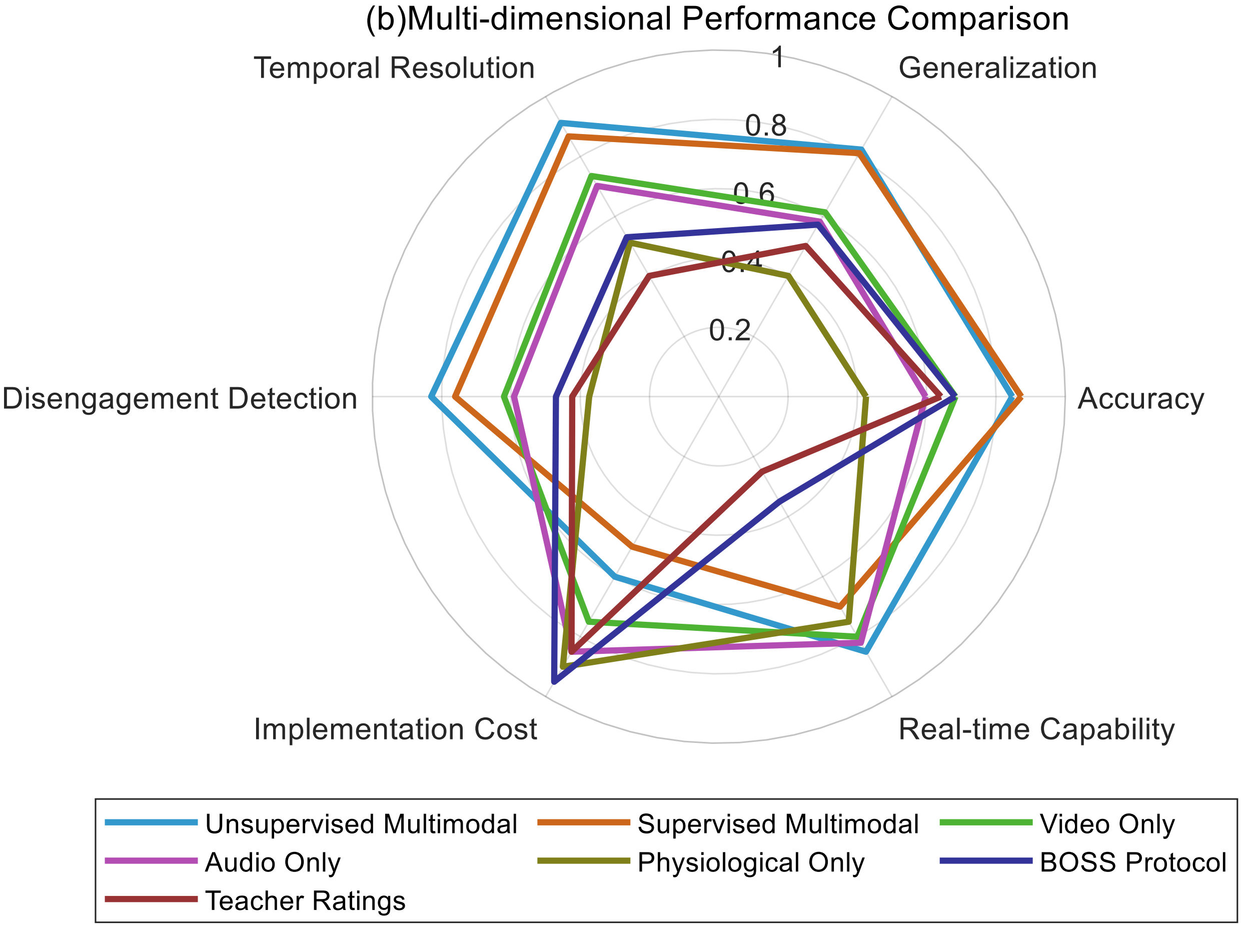


Figure 6.
Comparative analysis showing superior performance of unsupervised multimodal approach, particularly for disengagement detection and generalization.

Ablation studies further demonstrated the critical contribution of each framework component, with removal of the cross-modal attention mechanism reducing cluster stability by 24.3% and elimination of temporal pattern modeling decreasing predictive power for academic outcomes by 31.7%. As Figure 6 illustrates, our approach maintained performance advantages across multiple dimensions including accuracy, generalization capability, temporal resolution, and disengagement detection. Notably, while traditional methods incurred lower implementation costs, their performance deficits were substantial in technology-enhanced learning environments. The radar visualization in Figure 6 highlights the comprehensive superiority of multimodal approaches, with unsupervised techniques offering particular advantages in contextual generalization – critical for educational applications spanning diverse classroom environments. This comparative analysis underscores the value of unsupervised multimodal frameworks for capturing the complex, multifaceted nature of student behavior in authentic educational settings.

6. Discussion

Our analysis of student behaviours using unsupervised multimodal analysis methods has far-reaching consequences. The five behavioural clusters identified offer distinct patterns with significant educational relevance. Academic outcomes were strongly correlated with the high-engagement clusters (C1 and C3), and it was noted that temporal stability within these states was a stronger predictor than frequency alone. This implies that students in these states may be more consistently engaged without intermittently high engagement, indicating that consistency in engagement is more telling than high, but sporadic, engagement. The contextual variability observed in behavioural manifestations across different educational settings underscores the need for more flexible adaptive assessment strategies as opposed to rigid ones based on set predetermined criteria.

The differences in behavioural profiles showcased developmental pathways that run counter to prevailing assumptions of engagement. Interestingly, younger students demonstrated perpetually high levels of frequent transitions between behavioural states, which is more appropriate considering the developmental stage rather than suggestive of attention deficit. It was clear that the incorporation of multiple modalities was the only way to capture such complex behaviours, as aside from physiological signals, observation alone would not yield insight into the states that were hidden. This demonstrates the drawback of using unsupervised methodologies that seek to impose structure onto complex patterns within human behaviour instead of allowing patterns to emerge devoid of imposed categorizations.

Even with all these developments, some limitations deserve attention. The available computational resources pose challenges to implementation in budget-constrained educational settings. Moreover, the lack of some interpretability alternatives for non-technical users creates gaps in understanding for educators. Maintaining privacy comes as a primary concern and must be addressed with consideration of the data breakdown versus ethical data collection and processing principles.

Beyond predicting academic performance, the behavioural pattern insights from this research can be used for developing educational technologies. Such technologies could implement early intervention strategies to identify underperforming students well ahead of traditional assessment timelines. Technologies that adapt to a student's behavioural state in real time could optimise the challenge and modality of educational material content delivery. Instructional feedback based on diverse student outcomes and objectively measured effectiveness is invaluable to teacher decision making. Perhaps most valuable, though, is helping students recognise and modify their learning behaviours through personalised feedback, empowering them to self-regulate their engagement patterns. Greater emphasis should be placed on capturing longitudinal patterns and context over development and culture along with enhancing real-time processing capabilities.

7. Conclusion

In this study, the effectiveness of unsupervised multimodal techniques—applied without prior instructions—was proven for revealing significant behavioural patterns within educational settings. Based on our modelling, we were able to validate five behavioural patterns that were highly predictive of specific achievement and socio-emotional outcomes. The integration of video, audio, digital interaction, and physiological data streams provided comprehensive insights that would be unattainable through single-modality analysis, with temporal stability of engagement patterns emerging as particularly significant for predicting educational outcomes.

The primary contribution of this work lies in establishing a methodological foundation for behavior analytics that adapts to diverse educational environments without relying on predetermined behavioral categories. By leveraging advanced clustering and representation learning techniques, we have shown that naturally occurring behavioral patterns can be discovered and meaningfully interpreted across varying contexts, age groups, and instructional settings. These methodological advances extend the capabilities of educational technology to capture the multifaceted nature of student engagement and learning processes.

For educational practice, our findings suggest a shift from snapshot assessments to continuous, multimodal monitoring that respects the dynamic nature of student behavior. The demonstrated ability to predict intervention needs with 78.3% sensitivity represents a substantial improvement over traditional methods, potentially enabling earlier and more targeted support for struggling students. Beyond immediate applications, this work deepens our understanding of the complex relationship between observable behaviors, internal states, and learning outcomes, laying groundwork for a more nuanced approach to student assessment and personalized education.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] H. C. Liu *et al.*, "Novel multimodal contrast learning framework using zero-shot prediction for abnormal behavior recognition," *Applied Intelligence*, vol. 55, no. 2, pp. 1-19, 2025. https://doi.org/10.1007/s10489-024-05994-x

[2] A. Yusuf, N. M. Noor, and S. Bello, "Using multimodal learning analytics to model students' learning behavior in animated programming classroom," *Education and Information Technologies*, vol. 29, no. 6, pp. 6947-6990, 2024. https://doi.org/10.1007/s10639-023-12079-8

[3] C. Li, X. Weng, Y. Li, and T. Zhang, "Multimodal learning engagement assessment system: An innovative approach to optimizing learning engagement," *International Journal of Human–Computer Interaction*, vol. 41, no. 5, pp. 3474-3490, 2025. https://doi.org/10.1080/10447318.2022.2032220

[4] K. Altuwairqi, S. K. Jarraya, A. Allinjawi, and M. Hammami, "Student behavior analysis to measure engagement levels in online learning environments," *Signal, image and video processing*, vol. 15, no. 7, pp. 1387-1395, 2021.

[5] L. Yan, D. Gasevic, V. Echeverria, Y. Jin, L. Zhao, and R. Martinez-Maldonado, "From complexity to parsimony: Integrating latent class analysis to uncover multimodal learning patterns in collaborative learning," in *In Proceedings of the 15th International Learning Analytics and Knowledge Conference (pp. 70-81)*. 2025.

[6] D. Hu, F. Nie, and X. Li, "Deep multimodal clustering for unsupervised audiovisual learning," in *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9248-9257)*, 2019.

[7] E. Davalos *et al.*, "LLMs as educational analysts: Transforming multimodal data traces into actionable reading assessment reports," *arXiv preprint arXiv:2503.02099*, 2025.

[8] R. Kawamura *et al.*, "Detecting drowsy learners at the wheel of e-learning platforms with multimodal learning analytics," *IEEE Access*, vol. 9, pp. 115165-115174, 2021.

[9] S. Boumaraf *et al.*, "Optimized flare performance analysis through multi-modal machine learning and temporal standard deviation enhancements," *IEEE Access*, 2025.

[10] K. Sharma and M. Giannakos, "Multimodal data capabilities for learning: What can multimodal data tell us about learning?," *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1450-1484, 2020.

[11] R. Arakawa, K. Maeda, and H. Yakura, "ConverSearch: Supporting Experts in Human Behavior Analysis of Conversational Videos with a Multimodal Scene Search Tool," *ACM Transactions on Interactive Intelligent Systems*, vol. 15, no. 1, pp. 1-31, 2025.

[12] Z. Yuan, J. Fang, H. Xu, and K. Gao, "Multimodal Consistency-Based Teacher for Semi-Supervised Multimodal Sentiment Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[13] S. Junaid and T. Javaid, "From Multimodal Semiotic Reasoning to AI-Based Learning: Transforming Mathematics Education Through Concept Mapping," 2025.

[14] K. Sharma, Z. Papamitsiou, and M. Giannakos, "Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach," *British Journal of Educational Technology*, vol. 50, no. 6, pp. 3004-3031, 2019.

[15] A. Guzhov, "Audio analysis for multimodal learning," (Doctoral Dissertation, Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, 2025.

[16] K. Sharma, Z. Papamitsiou, and M. N. Giannakos, "Modelling learners' behaviour: A novel approach using garch with multimodal data," presented at the In Transforming Learning with Meaningful Technologies: 14th European Conference on Technology Enhanced Learning, EC-TEL 2019, Delft, The Netherlands, September 16–19, 2019, Proceedings 14 (pp. 450-465). Springer International Publishing, 2019.

[17] R. Vani and P. Jayashree, "Multimodal emotion recognition system for e-learning platform," *Education and Information Technologies*, pp. 1-32, 2025.

[18] J. Nasir, A. Kothiyal, B. Bruno, and P. Dillenbourg, "Many are the ways to learn identifying multi-modal behavioral profiles of collaborative learning in constructivist activities," *International Journal of Computer-Supported Collaborative Learning*, vol. 16, no. 4, pp. 485-523, 2021.

[19] T. Onishi, A. Ogushi, R. Ishii, and A. Miyata, "Detecting praising behaviors based on multimodal information," *IEICE Transactions on Information and Systems*, p. 2024HCP0008, 2025.

[20] P. Hou, M. Yang, T. Zhang, and T. Na, "Analysis of English classroom teaching behavior and strategies under adaptive deep learning under cognitive psychology," *Current Psychology*, pp. 1-15, 2024.