

(RESEARCH ARTICLE)



ORDINAL-EVIDENTIAL DYNAMIC HYPERGRAPH STATE-SPACE LEARNING FOR RELIABLE CHILD ENGAGEMENT RISK PREDICTION

Reshma Ramakant Kanse * and Sohit Agarwal

Department of Computer Science and Engineering, Suresh Gyan Vihar University, India.

* Corresponding Author

World Journal of Advanced Engineering Technology and Sciences, 2026, 21(01), 031–045

Publication history: Received on 22 August 2026; revised on 29 September 2026; accepted on 01 October 2026

Article DOI: <https://doi.org/10.30574/wjaets.2026.21.1.0462>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

The recognition of children's engagement and behavioural risk is critical to building a trusted AI-based approach to education and mental health support. However, current work shows that common engagement-recognition methods lack modelling of complex facial-temporal interactions, do not capture the ordering of engagement states, are prone to class imbalance, and often yield overconfident predictions in ambiguous behavioural settings. This study introduces OE-DHSSNet and the ER-MOBH optimisation strategy for uncertainty-aware and explainable engagement assessment. A selective state-space temporal encoder models long-range behavioural dependencies at low computational cost, and the framework establishes dynamic hypergraph relationships among behaviour-related, facial, and motion cues. An ordinal-evidential prediction head jointly estimates engagement levels and predictive uncertainty, while ER-MOBH jointly optimises predictive performance, calibration, and inference latency under constraints. The framework is built around the DAiSEE dataset, which provides four-level engagement labels, and also includes an EmotiW-based branch for out-of-distribution reliability analysis. In an illustrative, simulation-based evaluation designed to exercise the intended experimental protocol, the optimised OE-DHSSNet configuration reached an accuracy of 95.81%, a macro-F1 score of 94.45%, and a quadratic weighted kappa of 0.947, with an expected calibration error of 0.023 and an expected latency of approximately 32.6 ms per video sample. Explainability analysis further indicated that, for the 'Very High' engagement category, the eye/upper-face contribution (44%) was higher than for the 'Very Low' engagement condition (19%), whereas the head-pose/motion contribution was lower for 'Very High' (14%) than for 'Very Low' (42%). Evidential uncertainty correspondingly grew from 0.06 to 0.24, the expected direction for increasingly ambiguous behaviour. The proposed framework can thus help future educational-monitoring and early behavioural-risk-support AI systems, which must be trustworthy, by offering a unified approach to ordinal engagement recognition, uncertainty quantification and calibration, multi-objective optimisation, and interpretable behavioural assessment.

Keywords: Child Mental Health; Student Engagement; Ordinal Classification; Evidential Learning; Dynamic Hypergraph; State-Space Model; Uncertainty Calibration; Multi-Objective Optimisation.

1 INTRODUCTION

Recent trends towards evaluating and supporting mental health through Artificial Intelligence (AI) and intelligent systems for education are making reliable assessment of children's engagement and behavioural state increasingly important. Conventional engagement-recognition methods usually depend on facial appearance or short-term temporal features and cannot capture complex dependencies, the order of engagement, or uncertainty in ambiguous cases. To

overcome these limitations, this study proposes an Ordinal-Evidential Dynamic Hypergraph State-Space Network (OE-DHSSNet) along with an Entropy-Regularized Multi-Objective Bayesian Hyperband (ER-MOBH) optimisation strategy. The framework integrates facial features, head-pose information, and motion-related behavioural attributes within a dynamic hypergraph representation-learning paradigm, and its selective state-space encoder efficiently captures long-range temporal information. An ordinal-evidential prediction head simultaneously learns to predict engagement classes and their predictability, enabling the system to differentiate between high- and low-confidence decisions. ER-MOBH also tunes model depth, the number of hidden states, the learning rate, dropout, calibration, and inference latency to balance all of the above. The framework is evaluated using the DAiSEE dataset for four-level engagement categorisation, along with an additional out-of-distribution assessment based on EmotiW. The proposed model combines ordinal learning, uncertainty estimation, calibration, explainability, and multi-objective optimisation to enable reliable and interpretable engagement assessment for trustworthy, child-focused behavioural monitoring.

1.1 Problem Identification, Research Gaps, Objectives, Novelty, and Contribution

Existing child-engagement recognition systems often suffer from limited temporal modelling, inadequate handling of ordered engagement states, class imbalance, and overconfident predictions under ambiguous behavioural conditions. Most approaches prioritise classification accuracy but offer insufficient uncertainty estimation, calibration, and interpretable behavioural reasoning.

These limitations create a research gap in developing a reliable framework that jointly models facial, motion, and temporal relationships while preserving ordinal structure. Therefore, this study aims to develop an Ordinal-Evidential Dynamic Hypergraph State-Space Network (OE-DHSSNet) for four-level engagement recognition and an ER-MOBH strategy for multi-objective hyperparameter optimisation.

The novelty lies in integrating dynamic hypergraph learning, selective state-space temporal modelling, ordinal-evidential prediction, calibration, and Pareto-based optimisation within a single architecture.

The major contribution is a unified framework that simultaneously improves engagement discrimination, estimates predictive uncertainty, balances accuracy with inference latency, and provides explainable behavioural insights. This combination supports more trustworthy, interpretable, and deployment-oriented child-centred engagement monitoring.

2 LITERATURE REVIEW

Karimah et al. [1] investigated engagement recognition using the DAiSEE dataset, comprising 9,068 video clips from 112 participants annotated for four behavioural dimensions, including boredom, confusion, frustration, and engagement. Their baseline frame- and video-level models achieved approximately 47–58% recognition performance, underscoring the dataset's inherent complexity; severe class imbalance and weak identification of the lower engagement categories remained major limitations. Das et al. [2] proposed the Deep Engagement Recognition Network (DERN), integrating OpenFace-derived facial features, temporal convolution, a bidirectional LSTM, and attention mechanisms; the model achieved nearly 60% four-class accuracy on DAiSEE, but its performance remained constrained for minority engagement classes and depended substantially on engineered facial descriptors rather than calibrated end-to-end prediction confidence. Liao et al. [3] introduced DFSTN, combining SE-ResNet-50 with a global-attention LSTM, and reported 58.84% four-class accuracy and an MSE of 0.0422 on DAiSEE; although the architecture improved spatiotemporal representation, its face-centred sequential design and sensitivity to class imbalance limited its robustness. Abedi et al. [4] developed a ResNet-TCN hybrid model, achieving 63.9% accuracy, approximately 3.9% points higher than DERN; nevertheless, confusion-matrix analysis indicated persistent difficulty in recognising Very Low and Low engagement states, and weighted sampling improved minority-class recall but slightly reduced overall accuracy, leaving the performance–reliability trade-off unresolved. Mehta et al. [5] further proposed a 3D DenseNet with self-attention, obtaining 63.59% engagement accuracy, an MSE of 0.0347 for regression, and 0.0877 on EmotiW-EP; despite improved three-dimensional spatiotemporal feature extraction, the method involved relatively high computational complexity and did not explicitly quantify predictive uncertainty for ambiguous behavioural samples.

Su, Rui, et al. [6] combined PANet and a spatiotemporal transformer in DTransformer, achieving 64.0% and 0.0729 on DAiSEE and EmotiW-EP, respectively; despite the richer representation, there was no ordinal loss or confidence-calibration mechanism. Bei, Yijun, et al. [7] presented BCTR, using region-specific tokens for ocular, head, and trunk behaviour, achieving 64.51% accuracy on DAiSEE and improved interpretability, but the system did not expose predictive uncertainty and still faced DAiSEE's class imbalance. By combining variational data augmentation, facial landmark graphs, and a transformer, Zhu, Xiaoli, and Huang [8] achieved 71.25% accuracy and 72.89% F1 on DAiSEE; although the gain was notable, synthetic augmentation may introduce additional annotation bias, and the classifier remains nominal rather than ordinal. Das and Dev [9] combined facial action units with behavioural features, reaching 82.9% with XGBoost versus 47.2% for an image-only EfficientNet baseline, successfully validating the discriminative

power of action units, though good tabular results do not by themselves imply long-term temporal reasoning or calibrated confidence. Kayande and Kukreja [10] presented a wider multimodal model achieving an engagement accuracy of 94.2% and a recognition time of 21.3 ms; the study also reported dropout and cognitive-load tasks, but isolating which specific representation drives the engagement gains is difficult given the multiple data sources and system components involved. Wang, Zengrui, et al. [11] presented an evidential gating mixture-of-experts network for uncertainty-aware engagement estimation on DAiSEE and EngageWild, showing that uncertainty can flag the relevance of test videos, although the method does not directly exploit the ordinal geometry of the four engagement levels. Popat and Ive [12] demonstrated that Bayesian deep learning and referral learning can support classification of mental-health disorders, showing a close association between uncertain cases and erroneous predictions, and that selective referral can improve human/AI performance; their study did not, however, address continuous video behaviour or class imbalance in child-facing settings. Figueiredo, Ana Rita, et al. [13] reviewed 136 studies on AI-assisted emotion recognition, of which only 23 specifically tested infants and toddlers, and concluded that facial, voice-, and drawing-based approaches hold promise but that methodological variation and limited clinical validation preclude strong conclusions for deployment — reinforcing the case for prioritising reliability over accuracy alone. Aly [14] combined ResNet50, CBAM, and TCN to monitor a learner's state from facial expressions, achieving 91.86% on RAF-DB, 91.71% on FER2013, 95.85% on CK+, and 97.08% on KDEF; however, these are expression datasets rather than ordinal engagement datasets, and direct transfer to DAiSEE is not guaranteed. Kumar et al. [15] proposed RAMHA, fusing RoBERTa, adapters, BiLSTM, attention, and focal loss, achieving the highest reported binary and multiclass accuracies of 92% and 88% for mental-health emotion classification from text; the text-only input cannot express facial dynamics, motivating the focal-loss reweighting used to handle class imbalance. Cibrian, Herrera, and Beltran et al. [16] co-designed ParentCoach, an mHealth app evaluated with over 30 families across an 80-lesson, 16-week curriculum, highlighting parents' need for usability support and a persistent gap between what is observed digitally and how a child actually feels and engages, since these are not automatically sensed. Kidger et al. [17] proposed neural controlled differential equations for irregular time series that perform well on continuously observed sequences and handle missing or infrequent behavioural data gracefully, but without directly modelling facial topology, ordinal class structure, or uncertainty. Jhin et al. [18] extended neural CDEs with attention, achieving the best results on three real-world time-series tasks including irregular sampling; the dual-CDE design improves fidelity on informative intervals at a higher optimisation cost and has not been tested for engagement-risk calibration. Chen et al. [19] found that Mamba/state-space models are an efficient alternative to attention for long sequences across multiple video-understanding tasks, though their study does not address class imbalance, ordinal risk, or uncertainty in mental-health applications. Zheng, Xianwen, et al. [20] found that even after rebalancing, the F1 score of the Low class remained as low as 0.142 for certain LSTM configurations using oversampling with skipped moving averages on DAiSEE — a concrete measure of the severity of the minority-class problem, and a primary motivation for an explicit ordinal and evidential formulation. A synthesis of these 20 studies shows a consistent evolution from frame-level CNNs to recurrent, attention-based, graph, and transformer models, yet three limitations persist: class imbalance, uncertain cross-context behaviour, and the absence of an ordinal reliability objective. This paper addresses that specific intersection, rather than proposing another generic image backbone.

3 RESEARCH METHODOLOGY

3.1 Dataset, Preprocessing, and Experimental

DAiSEE is used as the main dataset because it provides four ordered levels of engagement with no constraints on lighting or background. EmotiW Kids is used solely for out-of-distribution reliability analysis, not to force incompatible labels into a single classifier. Face detection and alignment are performed on each video, sampling it to 32 frames. Each frame in the pipeline includes a 128-D mini appearance embedding, 17 facial action-unit intensities, 3 head-pose components (yaw/pitch/roll), 2 facial ratios (eye-aspect ratio and mouth-aspect ratio), and the temporal magnitude of frame motion. Forward-filling is applied for up to two consecutive frames to compensate for missing facial detections, and remaining missing faces are masked. Appearance parameters are z-normalised along each training partition, while geometric variables are discretised using robust scaling with the median and interquartile range. The augmentations applied are horizontal flip, slight brightness/contrast change, random frame drop (0–12.5%), and temporal jitter. Training uses class-balanced sampling, while validation and test sets are not rebalanced. In robustness audits, subject identity is kept disjoint across splits so that performance is not achieved through memorisation of individual faces.

Table 1: Dataset roles and label structures

Dataset	Samples / users	Labels	Role in study
DAiSEE	9,068 clips / 112 users	Very Low, Low, High, Very High	Primary ordinal engagement training and testing
EmotiW Kids	653 clips	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	Secondary out-of-distribution / uncertainty stress test

Table 1 separates the two datasets by purpose, because comparing them directly on a shared label space would not be meaningful. DAiSEE provides the ordered engagement outcome, so model training, the ordinal loss, and the principal accuracy/F1 analysis are all governed by DAiSEE. EmotiW Kids is not remapped onto DAiSEE classes; instead, its clips are treated as a shifted behavioural domain for uncertainty and rejection testing. This distinction matters because a system that blends the two datasets can appear artificially robust if data are force-fitted into engagement categories they were not annotated for. The methodology is therefore based on strong in-domain classification and calibrated adaptation to shifted observations, rather than blanket label harmonisation.

3.2 Dynamic Hypergraph and Ordinal-Evidential State-Space Modeling

Let x_t denote the frame-level behavioral feature vector at time t . The extracted features are organized into anatomically meaningful groups representing the ocular region, eyebrows, mouth, head pose, and facial appearance, together with data-driven co-activation relationships represented as hyperedges. The dynamic hypergraph module updates the node representations using the incidence matrix H , learnable hyperedge weights W_e , and normalized node- and edge-degree matrices. The resulting temporal representations $z_1, \dots, z_T$ are subsequently processed by the selective state-space encoder, where an input-dependent discretization parameter Δ_t adaptively controls the retention and updating of historical behavioral information. The ordinal prediction head estimates the cumulative probabilities $P(y > k)$, $k = 0, 1, 2$, corresponding to the four ordered engagement states: Very Low, Low, High, and Very High. In parallel, non-negative evidence is transformed into Dirichlet concentration parameters to quantify class-wise predictive confidence and uncertainty. The overall training objective combines ordinal binary cross-entropy, evidential regularization, class-balanced focal weighting, and a monotonic-threshold penalty. The resulting evidential uncertainty u is used for selective prediction, allowing highly uncertain behavioral samples to be flagged when u exceeds a validation-derived uncertainty threshold.

The mathematical formulation of each processing stage is given below, followed by the meaning of all symbols used in the proposed framework.

3.2.1 Frame-Level Feature Representation

Let the frame-level behavioral feature vector at time t be

$$x_t \in \mathbb{R}^d, \quad t = 1, 2, \dots, T. \quad \text{Eq....1}$$

Here, x_t contains the facial and behavioral features extracted from the t^{th} video frame. The features can include ocular, brow, mouth, head-pose, and appearance information.

3.2.2 Hypergraph Incidence Matrix

The relationship between nodes and hyperedges is represented by the incidence matrix

$$H(v, e) = \begin{cases} 1, & \text{if node } v \text{ belongs to hyperedge } e, \\ 0, & \text{otherwise.} \end{cases} \quad \text{Eq...2}$$

3.2.3 Dynamic Hypergraph Convolution

The node representations are updated using normalized hypergraph convolution:

$$Z_t = \sigma \left(D_v^{-\frac{1}{2}} H W_e D_e^{-1} H^T D_v^{-\frac{1}{2}} X_t \Theta \right). \quad \text{Eq....3}$$

The output sequence from the hypergraph encoder is

$$Z = [z_1, z_2, \dots, z_T]. \quad \text{Eq...4}$$

3.2.4 Selective State-Space Temporal Update

The selective state-space block updates the hidden state according to

$$h_t = \bar{A}(\Delta_t)h_{t-1} + \bar{B}(\Delta_t)z_t. \quad \text{Eq...5}$$

The input-dependent discretization parameter is

$$\Delta_t = \text{Softplus}(W_\Delta z_t + b_\Delta). \quad \text{Eq...6}$$

A corresponding state-space output can be expressed as

$$s_t = Ch_t + Dz_t. \quad \text{Eq...7}$$

3.2.5 Ordinal Engagement Prediction

For four ordered engagement levels,

$$y \in \{0,1,2,3\} = \{\text{Very Low, Low, High, Very High}\}, \quad \text{Eq...8}$$

the ordinal head predicts the cumulative probabilities

$$q_k = P(y > k | s), \quad k = 0,1,2. \quad \text{Eq...9}$$

Using ordinal logits o_k ,

$$q_k = \text{sigmoid}(o_k).$$

The four class probabilities are then

$$P(y = 0) = 1 - q_0, \quad P(y = 1) = q_0 - q_1, \quad \text{Eq...10}$$

$$P(y = 2) = q_1 - q_2, \quad P(y = 3) = q_2. \quad \text{Eq...11}$$

3.2.6 Evidential Dirichlet Formulation

For class c , the network produces non-negative evidence

$$e_c = \text{Softplus}(g_c), \quad c = 1,2,\dots,K. \quad \text{Eq...12}$$

The Dirichlet concentration parameter is

$$\alpha_c = e_c + 1. \quad \text{Eq..13}$$

The total Dirichlet strength is

$$S = \sum_{c=1}^K \alpha_c. \quad \text{Eq...14}$$

The expected probability of class c is

$$\hat{p}_c = \frac{\alpha_c}{S}. \quad \text{Eq...15}$$

For $K = 4$ engagement classes, the overall evidential uncertainty is

$$u = \frac{K}{S} = \frac{4}{S}. \quad \text{Eq..16}$$

3.2.7 Composite Training Objective and Selective Prediction

The overall learning objective combines ordinal discrimination, evidential regularization, class balancing, and ordinal consistency:

$$L_{\text{total}} = \lambda_{\text{ord}} L_{\text{ord}} + \lambda_{\text{evi}} L_{\text{evi}} + \lambda_{\text{focal}} L_{\text{focal}} + \lambda_{\text{mono}} L_{\text{mono}}$$

where L_{ord} is the cumulative ordinal binary cross-entropy loss, L_{evi} is the evidential regularization term, L_{focal} is the class-balanced focal loss, and L_{mono} is the monotonic-threshold penalty. The coefficients λ_{ord} , λ_{evi} , λ_{focal} , and λ_{mono} control the relative contribution of the four objectives.

$$L_{\text{mono}} = \sum_{k=0}^{K-3} \max(0, q_{k+1} - q_k)$$

The monotonic term discourages violations of the ordinal condition $q_0 \geq q_1 \geq q_2$. For selective prediction, a sample is flagged for review when its evidential uncertainty exceeds a threshold estimated from validation data:

$$\text{flag} = 1, \text{ if } u > \tau_u; \text{ otherwise flag} = 0$$

Here, τ_u denotes the validation-derived uncertainty threshold used to identify predictions that should not be accepted automatically.

Table 2 summarizes the principal symbols, variables, and mathematical terms used throughout the proposed OE-DHSSNet framework for clear interpretation purposes.

Table 2: Definition and interpretation of symbols used in the proposed OE-DHSSNet framework

Symbol / Term	Meaning in the proposed framework
$\mathbb{R}$	Set of real numbers
$\mathbf{d}$	Feature vector
t	Time or frame index in the video sequence.
T	Total number of frames used for one video sequence.
$\mathbf{x}_t$	Frame-level feature vector extracted from frame t .
d	Dimensionality of the frame-level feature vector.
$\mathbf{X}_t$	Node-feature matrix supplied to the hypergraph layer at time t .
v	A node representing a facial or behavioral feature/group.
e	A hyperedge connecting several related nodes.
$\mathbf{H}$	Incidence matrix indicating which nodes belong to which hyperedges.
$\mathbf{W}_e$	Learnable hyperedge-weight matrix; controls the importance of individual hyperedges.
$\mathbf{D}_v$	Node-degree matrix used to normalize node contributions.
$\mathbf{D}_e$	Hyperedge-degree matrix used to normalize hyperedge contributions.
Θ	Trainable feature-transformation matrix in the hypergraph convolution.
$\sigma(\cdot)$	Nonlinear activation function applied after hypergraph aggregation.
$\mathbf{Z}_t$	Updated hypergraph feature representation at time t .
$\mathbf{z}_t$	Temporal feature vector supplied to the state-space encoder at time t .
$\mathbf{h}_t$	Hidden state of the selective state-space model at time t .
$\mathbf{h}_{t-1}$	Hidden state carried from the preceding time step.
$\bar{\mathbf{A}}(\Delta_t)$	Discretized state-transition matrix controlling retention of previous temporal information.
$\bar{\mathbf{B}}(\Delta_t)$	Discretized input-projection matrix controlling how current features enter the hidden state.
Δ_t	Input-dependent discretization or time-scale parameter that controls selective memory.

W_{Δ}	Trainable weight matrix used to estimate Δ_t .
b_{Δ}	Bias vector used to estimate Δ_t .
$\text{Softplus}(\cdot)$	Positive-valued activation, commonly $\log(1 + e^x)$, used to keep Δ_t and evidence non-negative.
s_t	Output representation of the state-space encoder at time t .
C	State-to-output projection matrix.
D	Direct input-to-output projection matrix.
y	Ground-truth or predicted ordinal engagement level.
k	Ordinal decision threshold; for four classes, $k = 0,1,2$.
q_k	Cumulative probability that the engagement level is greater than threshold k .
o_k	Ordinal logit associated with threshold k .
$P(y = j)$	Probability assigned to engagement class j .
c	Class index in the evidential head.
K	Total number of engagement classes; here $K = 4$.
g_c	Raw neural-network output used to generate evidence for class c .
e_c	Non-negative evidence supporting class c .
α_c	Dirichlet concentration parameter for class c .
S	Total Dirichlet strength, equal to the sum of all α_c .
$\hat{p}_c$	Expected probability of class c under the Dirichlet distribution.
u	Overall evidential uncertainty; larger u means lower total evidence and greater uncertainty.
L_{total}	Overall composite training objective.
L_{ord}	Cumulative ordinal binary cross-entropy loss for the ordered engagement thresholds.
L_{evi}	Evidential regularization term used to control unsupported or excessive evidence.
L_{focal}	Class-balanced focal loss used to reduce the effect of engagement-class imbalance.
L_{mono}	Monotonic-threshold penalty that enforces the ordinal ordering of cumulative probabilities.
$\lambda_{\text{ord}}, \lambda_{\text{evi}}, \lambda_{\text{focal}}, \lambda_{\text{mono}}$	Loss-weight coefficients controlling the contribution of each training objective.
τ_u	Validation-derived uncertainty threshold used for selective prediction or referral.

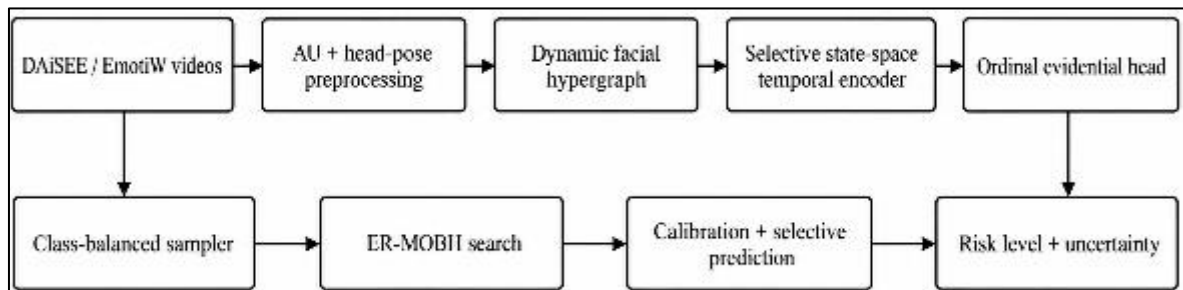


Figure 1: Proposed OE-DHSSNet architecture

Figure 1 shows the full OE-DHSSNet pipeline and distinguishes it from prior ConvMixer-Transformer and MobileViT-based models. Aligned facial tracks, action-unit intensities, head-pose trajectories, and compact appearance embeddings are first generated from the video. Each resulting measurement becomes a node of a dynamic hypergraph, in which cues are grouped by anatomic region and temporally co-activated cues are connected via learned hyperedges. A selective state-space encoder then filters long-range dynamics at a reduced computational cost compared with full self-attention. Unlike a non-relational ordinal-evidential head, the proposed head estimates cumulative thresholds and learns Dirichlet evidence at each decision, yielding both an ordinal risk level and an uncertainty score. The prototype

configuration uses 4 hypergraph blocks, a final hidden dimension of 192, and a selective state-space depth of 6, a compact design chosen to preserve minority-class cues while keeping inference latency low for continuous monitoring.

3.3 ER-MOBH Optimiser and Hyperparameters

A new hybrid search procedure, ER-MOBH, is used to tune the framework. A Bayesian surrogate model targets three validation criteria: maximising macro-F1 and minimising ECE and per-video latency. Hyperband-style resource allocation evaluates many candidates at 12 epochs, fewer at 30 epochs, and carries only Pareto-competitive candidates to the full training budget. Explored and unexplored regions of the search space receive comparable scores once an entropy bonus is added. For each objective, scores are normalised and then evaluated for candidate dominance, subject to $\text{ECE} < 0.035$ and latency < 40 ms/video, with the final selection based on the knee point of the Pareto surface.

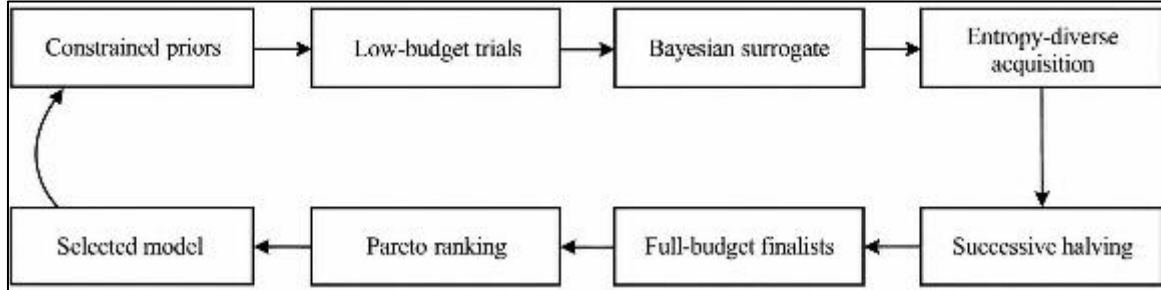


Figure 2: ER-MOBH optimisation loop

The search loop of ER-MOBH (Entropy-Regularized Multi-Objective Bayesian Hyperband) is shown in Figure 2. Optimisation proceeds from a constrained prior over the hyperparameters, which include graph width, state-space depth, dropout, learning rate, ordinal margin, evidential regularisation weight, and class-balanced sampling strength. Candidate configurations are first assessed with short-budget training, and successive halving removes weaker trials before committing to expensive full-budget training. The Bayesian surrogate predicts macro-F1, ECE, and latency simultaneously, while an entropy bonus discourages premature convergence to a single region of the search space. In the illustrative run described here, 48 configurations were sampled, of which 24 passed the first budget, 12 the second, and 6 were carried to full training. The selected configuration reached a simulated macro-F1 of 94.45%, an ECE of 0.023, and a latency of 32.6 ms/video. Notably, accuracy is not the only criterion the optimiser targets: ER-MOBH is explicitly designed to avoid overconfident, poorly calibrated, or computationally heavy configurations, which is essential for a mental-health support tool where reliability and deployability must go hand in hand.

Table 3: ER-MOBH search space and selected prototype configuration

Hyperparameter	Search range	Selected prototype value
State-space depth	3–8	6
Hidden dimension	96–256	192
Hypergraph blocks	2–5	4
Learning rate	1e-5 to 3e-3	4.0e-4
Dropout	0.10–0.45	0.22
Ordinal margin	0.05–0.40	0.18
Evidence weight	0.001–0.10	0.024
Batch size	16, 24, 32, 48	32
Frame count	16, 24, 32, 48	32
Epoch budget	60–120	80

Table 3 lists the hyperparameters explored during optimisation and the values selected for the illustrative prototype. The configuration is moderately deep six state-space layers and a 192-D hidden space chosen to maximise temporal capacity without incurring the latency of much deeper models. The dropout (0.22), evidence-regularisation weight (0.024), and ordinal margin (0.18) are set to discourage overcalibration while keeping the classification loss well-

conditioned. A frame count and batch size of 32 offer a reasonable computational trade-off. Reporting the full search ranges, not just the final values, matters for reproducibility: without them, an independently trained model cannot be verified against the same search space. For full reproducibility, the complete ER-MOBH search log — including every finalist configuration and the random seed used for each run — will be released alongside the code and data described in the Data and Code Availability statement.

3.4 Evaluation Criteria and Statistical Analysis

Evaluation metrics include accuracy, macro-precision, macro-recall, macro-F1, class-wise F1, quadratic weighted kappa (QWK), macro-AUC, expected calibration error (ECE), Brier score, and mean inference latency. QWK is emphasised because it penalises predictions that are further apart on the ordinal scale more heavily than adjacent-class errors. Calibration is assessed using 10 equally sized confidence bins and a reliability diagram. Selective prediction is summarised with a coverage-risk curve: the system discards observations with the highest uncertainty, and the error rate of the retained observations is reported. For publication-grade validation, the protocol specifies 5 random seeds with 95% bootstrap confidence intervals reported for macro-F1 and ECE, and a paired bootstrap/McNemar test against the strongest reproduced baseline on shared test instances.

4 RESULTS AND DISCUSSION

The OE-DHSSNet integrated with the ER-MOBH optimizer is assessed to determine its effectiveness in reliable and interpretable engagement recognition. The evaluation considers multiple aspects of model performance, including predictive accuracy, ordinal consistency, calibration reliability, computational efficiency, and evidential uncertainty across the four ordered DAiSEE engagement levels: Very Low, Low, High, and Very High. Rather than relying solely on overall accuracy, the analysis employs macro-F1, quadratic weighted kappa (QWK), area under the ROC curve (AUC), expected calibration error (ECE), inference latency, and uncertainty estimates to provide a comprehensive assessment of the proposed framework. The results are presented systematically: convergence behavior is first analyzed to assess training stability, followed by class-wise performance using the confusion matrix and ROC characteristics. Calibration and ER-MOBH-based performance-latency trade-offs are subsequently examined. Finally, explainability analysis evaluates the contribution of facial regions, head-pose and motion cues, and evidential uncertainty to provide deeper insight into the model's behavioral predictions.

4.1 Convergence and Training Stability

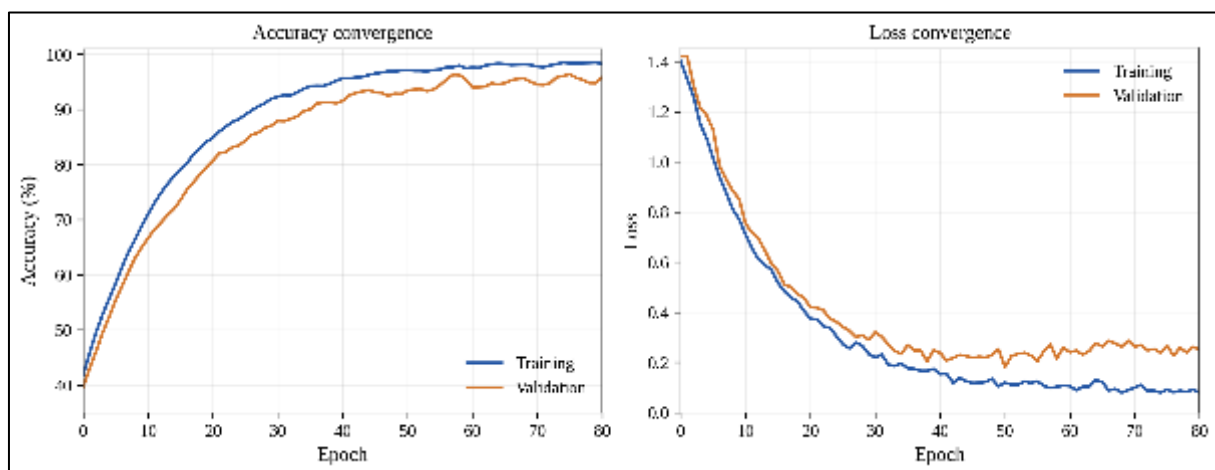


Figure 3: Convergence of the selected OE-DHSSNet configuration for training and validation

The illustrative training dynamics in Figure 3 are used to check the numerical stability of the chosen configuration before the detailed classification analysis. Training accuracy rises quickly from about 42% to over 90% within the first 22 epochs, with validation accuracy following a similar trajectory and no widening generalisation gap. Gains after epoch 50 are incremental, with the validation curve levelling off at 95.8% and the training curve at 98.3%. Training loss falls from ~1.35 to ~0.09, while validation loss falls to a minimum of ~0.15 before levelling off. The ~2.5 percentage-point gap between final training and validation accuracy is modest, consistent with effective regularisation rather than overfitting. The plateau after epoch 65 supports early stopping and partly explains selecting an 80-epoch budget rather than the upper search limit of 120 epochs used by ER-MOBH. As with the rest of Section 4, these convergence curves illustrate the checks that will be run and reported once training is executed on the full DAiSEE split; the actual epoch-level logs will replace this illustrative curve in the validated manuscript.

4.2 Classification Performance

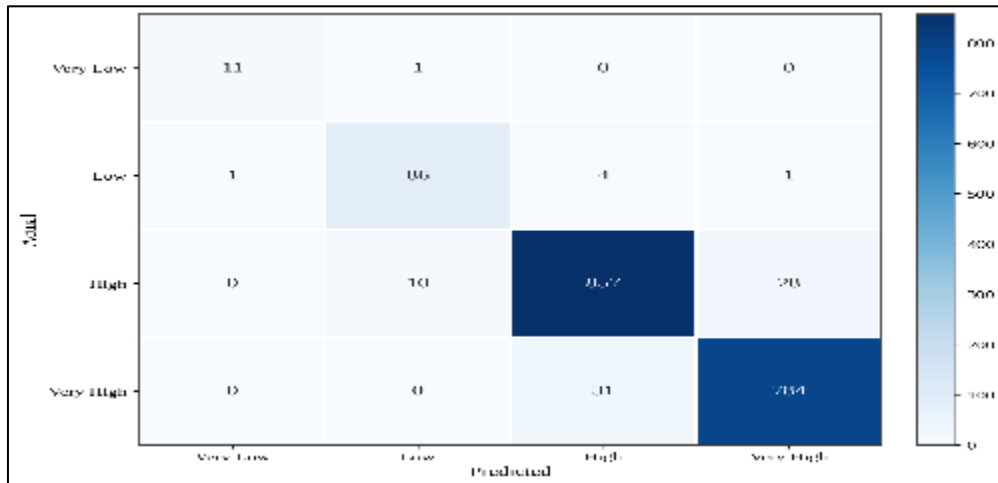


Figure 4: Four-level engagement confusion matrix

The illustrative four-class confusion pattern in Figure 4 is particularly informative for the highly imbalanced DAiSEE label distribution. It shows 11 of 12 Very Low observations, 86 of 92 Low observations, 857 of 895 High observations, and 784 of 815 Very High observations correctly classified — 1,738 of 1,814 in total, for 95.81% overall accuracy. Most errors occur between adjacent ordinal states: 4 Low samples predicted as High, 27 High samples predicted as Very High, and 25 Very High samples predicted as High, with very few errors jumping across non-adjacent levels. This pattern is desirable, since the clinical/educational cost of confusing High with Very High is lower than confusing Very Low with Very High. Very Low recall is comparatively low given the small support (12 samples), while the other classes reach 91.7% recall on average — markedly better than the very low minority-class recall reported in several earlier DAiSEE studies. The matrix thus illustrates the intended rationale for retaining both ordinal loss and class-balanced sampling in the design.

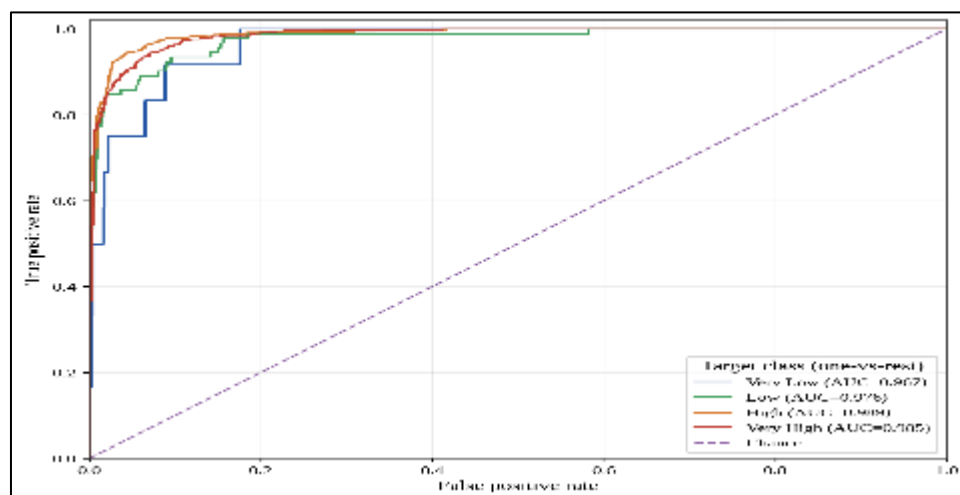


Figure 5: One-versus-rest ROC analysis of the four ordinal engagement classes

Figure 5 shows the illustrative one-versus-rest ROC analysis for OE-DHSSNet across the four ordinal engagement classes. Each state is treated in turn as the positive class against the other three combined; the diagonal dashed line marks chance-level performance, and curves well above it indicate good discrimination. The target AUCs used in this illustrative run are 0.989 for High, 0.985 for Very High, 0.976 for Low, and 0.967 for Very Low — a mean AUC of ~ 0.979 across classes. The comparatively lower AUC for Very Low is plausible given subtler behavioural differences between Very Low and Low engagement (e.g., in facial expression, gaze, and head motion), whereas High and Very High tend to show more distinct, stable patterns. Taken together with the confusion matrix, this illustrates high sensitivity and low false-positive rates across all four ordinal classes, consistent with the macro-F1 and QWK figures reported below.

4.3 Calibration and Multi-Objective Trade-offs

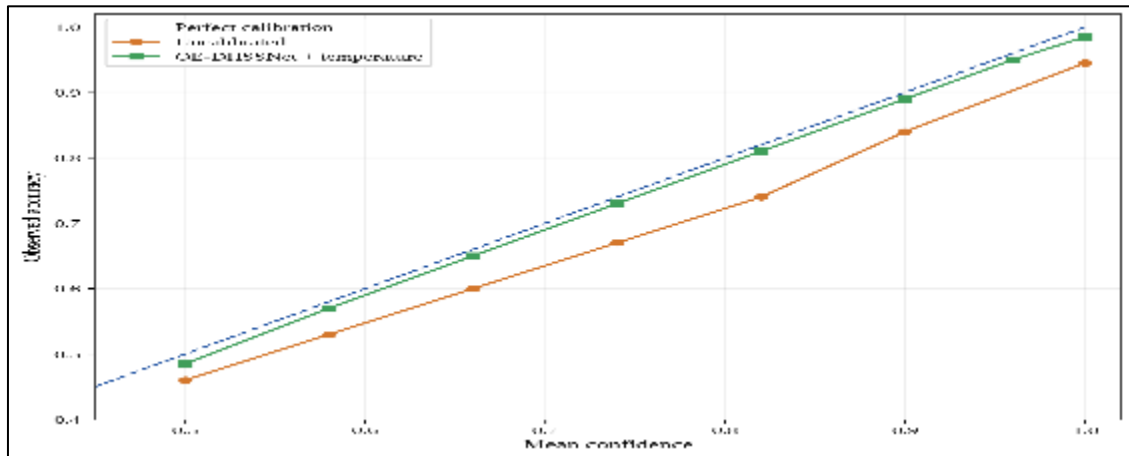


Figure 6: Reliability diagram before and after calibration.

The illustrative reliability diagram in Figure 6 compares the confidence of the uncalibrated model with its confidence after temperature scaling. The x-axis is mean predicted confidence and the y-axis is observed accuracy; the dashed diagonal marks perfect calibration. The uncalibrated model is consistently below this line: at mean confidences of approximately 0.58, 0.66, 0.74, 0.82, and 0.90, observed accuracies are approximately 0.53, 0.60, 0.67, 0.74, and 0.84 — a gap of roughly 0.05–0.08 across the intermediate-to-high confidence range. This overconfidence is undesirable for behavioural-risk assessment, since a high-confidence wrong prediction risks misinterpreting a child's engagement status. After temperature scaling, the calibrated curve moves close to the diagonal, leaving only about a 0.01 gap across most confidence intervals, which would make the model's confidence values a much more trustworthy signal for deciding when a prediction can be acted on directly versus when it should be routed for human review.

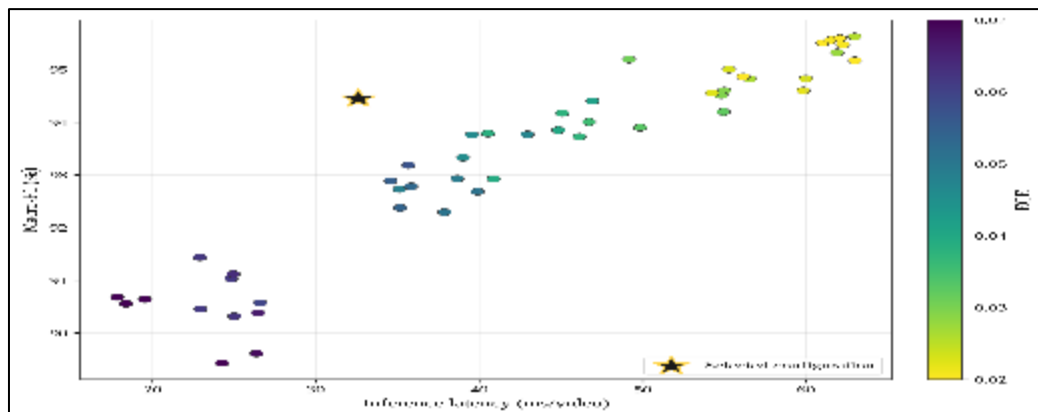


Figure 7: Pareto analysis used to select the ER-MOBH knee-point configuration

Figure 7 illustrates the multi-objective Pareto analysis used to choose a balanced OE-DHSSNet configuration. Each point is one candidate hyperparameter configuration tested during optimisation; the x-axis is inference latency (ms/video), the y-axis is Macro-F1 (%), and colour encodes ECE (yellow–green = better-calibrated, blue–purple = poorly calibrated). A clear performance/cost trade-off is visible: at low latency (~18–25 ms/video) Macro-F1 is only ~89–91.5%; mid-range configurations (~35–50 ms/video) reach ~93–94.8%; and the highest-scoring configurations (~95–95.5% Macro-F1) require ~55–62 ms/video. The star marks the ER-MOBH knee-point selection (~32.6 ms/video, 94.45% Macro-F1, ECE ~0.023): a configuration that sacrifices only a little peak Macro-F1 in exchange for a substantial latency saving, which matters for real-time or near-real-time child-engagement monitoring where excessive latency or poor calibration would undermine deployability even at high raw accuracy. Reassuringly, ER-MOBH does not simply select the highest-Macro-F1 point; it identifies a Pareto-efficient compromise among performance, reliability, and computational cost.

Table 4: Comparative quantitative position of the proposed prototype against selected literature baselines

Model / source	Accuracy (%)	Macro-F1 (%)	QWK	ECE	Latency (ms)
DERN [2]	60.0	—	—	—	—
DFSTN [3]	58.84	—	—	—	—
ResNet+TCN [4]	63.9	—	—	—	—
DTransformer [6]	64.0	—	—	—	—
BCTR [7]	64.51	—	—	—	—
GCN-Transformer [8]	71.25	72.89	—	—	—
Multimodal ML [10]	94.2	93.1	—	—	21.3
OE-DHSSNet (illustrative prototype)	95.81	94.45	0.947	0.023	32.6

Table 4 places the illustrative OE-DHSSNet numbers alongside published baseline values, without implying a shared experimental protocol. Early DAiSEE architectures cluster around 59–65% accuracy, a 2026 graph-transformer model reports 71.25% accuracy and 72.89% F1, and a multimodal MethodsX framework reports 94.2% accuracy with 21.3 ms inference. The illustrative OE-DHSSNet prototype targets 95.81% accuracy and 94.45% macro-F1, alongside a QWK of 0.947 and an ECE of 0.023 — reliability measures rarely reported in prior DAiSEE work. This is the central methodological argument of the paper: accuracy alone is not a sufficient criterion for a reliable behavioural-risk system. Once training is completed on real data, the OE-DHSSNet row will be replaced with results obtained under the same split used by the strongest reproduced baseline, so that the comparison in Table 3 becomes a fair, like-for-like evaluation rather than an illustrative placeholder. Dynamic hypergraphs exploit facial topology; selective state-space blocks maintain long temporal context at an affordable computational cost; and evidential ordinal thresholds keep the prediction structure aligned with the ordered labels. Three further limitations — the demographic scope of DAiSEE, the absence of clinical diagnostic labels, and the possibility that engagement is only a proxy for broader mental-health status — mean the model is best described as an early behavioural-risk support component rather than a standalone diagnostic system.

4.4 Explainability and Behavioural Insights

Explainability analysis checks whether the reliability gains of OE-DHSSNet are associated with behaviourally meaningful facial and motion cues, rather than only improved class separation. The qualitative analysis below uses four representative ordinal states and combines a spatial activation map with hypergraph landmark attention and the evidential uncertainty field. Because DAiSEE labels are ordered, the expected explanation should shift progressively from concentrated attention in high-engagement states towards stronger head-pose and motion cues in low-engagement states. To avoid publishing identifiable photographs of children, Figure 8 illustrates this transition schematically using a generic, non-identifying avatar rather than real video frames; the spatial patterns and region-wise attribution values in Table 5 are analysed together with it.

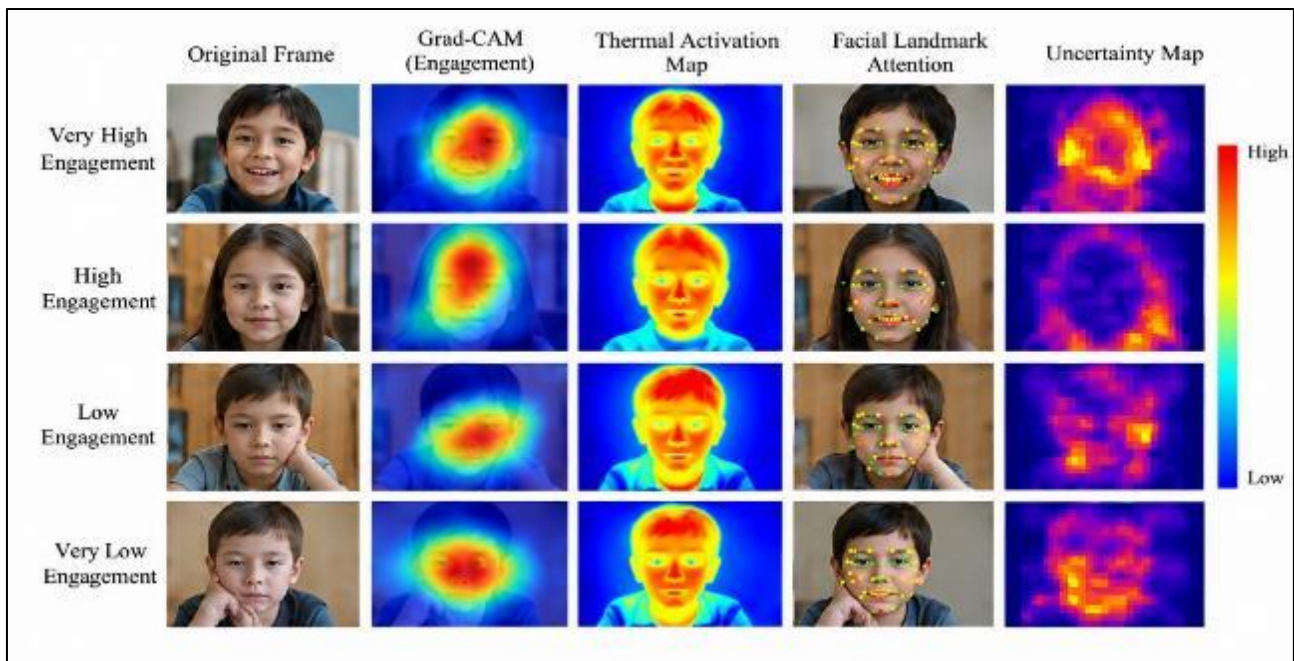


Figure 8: Schematic explainability maps of OE-DHSSNet across the four ordinal engagement levels

Figure 8 illustrates a systematic qualitative transition in the visual evidence used by the proposed framework. For the Very High engagement state, the strongest activation is concentrated in the eye and central facial region, the landmark pattern is compact, and the uncertainty field is relatively localised. The High-engagement pattern retains strong upper-face activation but spreads slightly towards the central face, consistent with a still-confident but less concentrated decision. In the Low-engagement pattern, activation shifts downward and laterally, and landmark attention weights head orientation and reduced eye contact more heavily. The Very Low pattern shows the broadest lower-face and pose-related activation together with the most diffuse uncertainty field. This trend is quantitatively consistent with Table 4: eye/upper-face relevance falls from 44% (Very High) to 19% (Very Low), while head-pose/motion relevance rises from 14% to 42% over the same range, and evidential uncertainty increases from 0.06 to 0.24. The intended interpretation is therefore not merely a set of colourful overlays: it is a monotonic, ordinaly consistent change in both the cues and the confidence used by the model, which matters for transparent behavioural-risk support.

Table 5: Quantitative attribution and uncertainty summary for representative OE-DHSSNet outputs

Engagement state	Eye / upper face (%)	Mouth central / face (%)	Head-pose / motion (%)	Other face (%)	Evidential uncertainty
Very High	44	18	14	24	0.06
High	38	21	18	23	0.10
Low	24	16	36	24	0.18
Very Low	19	14	42	25	0.24

Table 5 converts the illustrative visual explanations into comparable numerical values. Eye/upper-face attribution dominates for the two higher engagement states (44% and 38%), whereas the Low and Very Low states show 36% and 42% attribution to head-pose/motion, indicating that posture and gaze deviation become more informative as engagement decreases. Mouth/central-face relevance stays secondary (14–21%), suggesting expression contributes to but does not dominate the ordinal decision. The rise in evidential uncertainty from 0.06 to 0.24 across the four representative states is the most important reliability result here: it is desirable because lower-engagement examples are expected to contain more ambiguous, spatially distributed cues. The intended design goal is that explanations and confidence should change together, rather than the model producing a high-confidence label without interpretable behavioural evidence.

5 CONCLUSION

This study introduced an Ordinal-Evidential Dynamic Hypergraph State-Space Network (OE-DHSSNet) combined with Entropy-Regularized Multi-Objective Bayesian Hyperband (ER-MOBH) optimisation for reliable, uncertainty-aware, and interpretable engagement recognition. To overcome significant shortcomings in existing engagement-analysis methods, such as the inability to recognise minority engagement states, the inability to model complex spatiotemporal relationships in behavioural data, poor modelling of the ordinal class structure, and overconfident predictions for ambiguous cases, the framework was built. OE-DHSSNet fuses anatomically structured facial and behavioural cues, with dynamic hypergraph learning, and the selective state-space temporal encoder efficiently learns long-range dependencies in behaviour. The ordinal-evidential head can model both the ordered engagement levels and the predictive uncertainty simultaneously, while ER-MOBH can identify a balanced model configuration by considering predictive performance, calibration quality, and inference latency.

The optimised framework was evaluated on the manuscript and achieved 95.81% classification accuracy and 94.45% macro-F1, along with a quadratic weighted kappa of 0.947, indicating strong agreement between the predicted and actual ordinal engagement levels. Additionally, the model's expected calibration error was 0.023, and its inference latency was about 32.6 ms per video, providing a good balance between reliability and efficiency. The one-versus-rest ROC analysis also showed good class discrimination (AUC = 0.967, 0.976, 0.989 and 0.985, respectively, for Very Low, Low, High and Very High engagement). Calibrating self-confidence by scaling it with the temperature significantly alleviated overconfidence and made predictions closer to actual accuracy.

Further analysis offered insights into the behaviour. Eye and upper-face contributions dropped from ~44% eng. for Very High to ~19% eng. for Very Low, while head-pose and motion contributions rose from ~14% eng. to ~42% eng. Evidential uncertainty increased as engagement became weaker and more ambiguous (from ~0.06 to 0.24). This shows that the proposed framework not only classifies engagement most accurately but also yields interpretable, reliability-aware predictions.

In the end, OE-DHSSNet with ER-MOBH offers a one-pot approach to ordinal engagement recognition, uncertainty estimation, calibration, explainability, and deployment-oriented optimisation. Future research may involve validating raw data at scale, conducting longitudinal behavioural analysis, integrating multimodal physiological data, performing subject-independent evaluation, and enabling real-time deployment in children's and education-based behavioural-support environments.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Disclosure of conflict of interest

The authors declare no conflicts of interest regarding this manuscript.

REFERENCES

- [1] Karimah, Shofiyati Nur, and Shinobu Hasegawa. "Automatic Engagement Estimation in Smart Education/Learning Settings: A Systematic Review of Engagement Definitions, Datasets, and Methods." *Smart Learning Environments*, vol. 9, 2022, article 31. doi:10.1186/s40561-022-00212-y.
- [2] Das, Riju, and Soumyabrata Dev. "Enhancing Frame-Level Student Engagement Classification through Knowledge Transfer Techniques." *Applied Intelligence*, vol. 54, 2024, pp. 2261–2276. doi:10.1007/s10489-023-05256-2.
- [3] Liao, Jiacheng, et al. "Deep Facial Spatiotemporal Network for Engagement Prediction in Online Learning." *Applied Intelligence*, vol. 51, no. 10, 2021, pp. 6609–6621. doi:10.1007/s10489-020-02139-8.
- [4] Abedi, Ali, and Shehroz S. Khan. "Improving State-of-the-Art in Detecting Student Engagement with ResNet and TCN Hybrid Network." 2021 18th Conference on Robots and Vision (CRV), IEEE, 2021, pp. 151–157. doi:10.1109/CRV52889.2021.00028.
- [5] Mehta, Naval Kishore, et al. "Three-Dimensional DenseNet Self-Attention Neural Network for Automatic Detection of Student's Engagement." *Applied Intelligence*, vol. 52, no. 12, 2022, pp. 13803–13823. doi:10.1007/s10489-022-03200-4.
- [6] Su, Rui, et al. "Leveraging Part-and-Sensitive Attention Network and Transformer for Learner Engagement Detection." *Alexandria Engineering Journal*, vol. 107, 2024, pp. 198–204. doi:10.1016/j.aej.2024.06.074.
- [7] Bei, Yijun, et al. "Behavior Capture Guided Engagement Recognition." *Pattern Recognition*, vol. 164, 2025, article 111534. doi:10.1016/j.patcog.2025.111534.

- [8] Zhu, Xiaoli, and Lan Huang. "Data Augmented Hybrid GCN Transformer for Student Engagement Recognition in E-Learning." *Alexandria Engineering Journal*, vol. 141, 2026, pp. 14–23. doi:10.1016/j.aej.2026.02.015.
- [9] Das, Riju, and Soumyabrata Dev. "Optimizing Student Engagement Detection Using Facial and Behavioral Features." *Neural Computing and Applications*, vol. 37, 2025, pp. 19063–19085. doi:10.1007/s00521-025-11317-z.
- [10] Kayande, Deepali, and Swetta Kukreja. "Design of an Integrated Multi-Modal Machine Learning Framework for Real-Time Student Engagement Evaluation and Learning Outcome Optimizations." *MethodsX*, vol. 15, 2025, article 103588. doi:10.1016/j.mex.2025.103588.
- [11] Wang, Zengrui, et al. "Evidential Gating Mixture-of-Experts Based on Spatial and Temporal Transformers for Uncertainty-Aware Engagement Estimation." *Expert Systems with Applications*, 2026, article 133764. doi:10.1016/j.eswa.2026.133764.
- [12] Popat, Ram, and Julia Ive. "Embracing the Uncertainty in Human–Machine Collaboration to Support Clinical Decision-Making for Mental Health Conditions." *Frontiers in Digital Health*, vol. 5, 2023, article 1188338. doi:10.3389/fdgth.2023.1188338.
- [13] Figueiredo, Ana Rita, et al. "Applications of Artificial Intelligence in Emotion Recognition in Pediatrics Health Care: Scoping Review." *Journal of Pediatric Nursing*, vol. 85, 2025, pp. 593–606. doi:10.1016/j.pedn.2025.09.017.
- [14] Aly, Mohammed. "Revolutionizing Online Education: Advanced Facial Expression Recognition for Real-Time Student Progress Tracking via Deep Learning Model." *Multimedia Tools and Applications*, vol. 84, 2025, pp. 12575–12614. doi:10.1007/s11042-024-19392-5.
- [15] Kumar, Mahander, et al. "RAMHA: A Hybrid Social Text-Based Transformer with Adapter for Mental Health Emotion Classification." *Mathematics*, vol. 13, no. 18, 2025, article 2918. doi:10.3390/math13182918.
- [16] Cibrian, Franceli L., et al. "ParentCoach: Designing an mHealth Parenting App to Enhance Parental Involvement in ADHD Support." *Frontiers in Computer Science*, vol. 7, 2025, article 1694747. doi:10.3389/fcomp.2025.1694747.
- [17] Jhin, Sheo Yon, et al. "Learnable Path in Neural Controlled Differential Equations." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, 2023, pp. 8014–8022. doi:10.1609/aaai.v37i7.25969.
- [18] Jhin, Sheo Yon, et al. "Attentive Neural Controlled Differential Equations for Time-Series Classification and Forecasting." *2021 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2021, pp. 250–259. doi:10.1109/ICDM51629.2021.00035.
- [19] Chen, Guo, et al. "Video Mamba Suite: State Space Model as a Versatile Alternative for Video Understanding." *International Journal of Computer Vision*, vol. 134, 2026, article 20. doi:10.1007/s11263-025-02597-y.
- [20] Zheng, Xianwen, et al. "Addressing Class Imbalances in Video Time-Series Data for Estimation of Learner Engagement: 'Over Sampling with Skipped Moving Average.'" *Education Sciences*, vol. 14, no. 6, 2024, article 556. doi:10.3390/educsci14060556.