

(RESEARCH ARTICLE)



WAVELET-KOOPMAN PROTOTYPE TRANSPORT LEARNING FOR CROSS-DATASET CHILD AFFECT AND ENGAGEMENT GENERALIZATION

Reshma Ramakant Kanse * and Sohit Agarwal

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

* Corresponding Author

Global Journal of Engineering and Technology Advances, 2026, 28(03), 324–339

Publication history: Received on 15 August 2026; revised on 27 September 2026; accepted on 29 September 2026

Article DOI: <https://doi.org/10.30574/gjeta.2026.28.3.0273>

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

A significant challenge in automated child affect and engagement analysis is transferring a model across datasets, as recording conditions, child behaviour, temporal effects, and annotation schemes vary and can seriously impact model reliability. To address this challenge and enable robust cross- dataset behavioural representation learning, this paper proposes Wavelet–Koopman Prototype Transport Network (WKPT- Net) and a Bilevel Spectral–Prototype Transport Search (BSPTS) strategy for WKPT- Net. Using a three- band wavelet analysis, the framework decomposes facial and movement characteristics into three temporal layers: slow, intermediate, and fast. Nonlinear behavioural evolution is then modelled by band- wise Koopman latent operators, with complementary temporal information fused via a spectral gate. Prototype memory and Sinkhorn- based optimal transport transfer latent structures between DAiSEE and EmotiW without mapping the non- identical label space to a common classification head. All parameters, including wavelet configuration, latent dimensionality, transport regularization, prototype parameters, and computational complexity, are jointly optimized by BSPTS. For the same deterministic prototype evaluation, WKPT- Net achieved accuracy and macro- F1 scores of 96.2% and 95.2%, respectively, on the DAiSEE dataset, whereas the dataset- specific EmotiW emotion head achieved 94.8% and 95.2%, respectively. There was a 68.9% reduction in cross- domain discrepancy when normalized, while prototype transport achieved separation below 6% of inter- prostate distance, indicating good system alignment without collapsing. The proposed architecture thus shows the promise of multiscale temporal decomposition, Koopman dynamics, prototype learning, and optimal transport in transferable behavioural modelling. The results help underpin the creation of scalable frameworks for AI- assisted engagement monitoring and individualized behaviour assessment with mental- health- oriented perspectives.

Keywords: Child Affect Recognition, Engagement Analysis, Wavelet Decomposition, Koopman Operator, Prototype Learning, Optimal Transport, Domain Adaptation, BSPTS.

1 INTRODUCTION

Effective identification of children's affective and engagement states has emerged as a key research area for intelligent learning environments and systems for AI-assisted mental health support. Facial expressions, head movements, gaze movements, and task duration can be very telling in determining the emotional state and learning engagement. Recently, deep learning-based techniques, such as multimodal fusion methods, transformers, graph-based models, and

convolutional neural networks, have yielded encouraging results in this problem, as seen in [1–4]. However, most current methods have been developed and tested on a single dataset and are therefore sensitive to variations in illumination, camera viewpoints, participant characteristics, recording conditions, and annotation procedures. This means that if the model is trained on a single dataset, its performance in another domain could be significantly reduced, thereby restricting its generalisation.

Another significant restriction is the lack of an explicit distinction between the temporal frequencies of different features in the behaviour. Conventional architectures typically learn temporal properties directly from the sequence of frames. Children's affective behaviour exhibits both low-frequency engagement-level changes and high-frequency changes; hence, representing both processes with a single time scale can result in the loss of diagnostically important information. In addition, most current cross-domain approaches match the global feature distributions and disregard class-specific patterns and temporal evolution in the datasets, which may result in negative transfer in the presence of heterogeneous datasets.

With these limitations in mind, the present study introduces a Wavelet–Koopman Prototype Transport Network (WKPT-Net) for modelling modulation across datasets of children's affect and engagement. Temporal behavioural representations are decomposed into multiple frequency bands using wavelet analysis as the framework's first step. Then, band-wise latent dynamics are modelled with a Koopman-inspired operator to represent nonlinear temporal evolution in an approximately linear latent space. Spectral information is fused with gated fusion, and a class-specific prototype memory preserves the discriminative semantic structures. A prototype discrepancy minimization across the two domains, employing the optimal-transport alignment mechanism, follows. Last but not least, the Bilevel Spectral–Prototype Transport Search (BSPTS) is a novel method for optimizing the wavelet configuration, latent dimensionality, transport parameters, prototype setups, and computational constraints. The proposed framework thus offers a holistic approach to multiscale temporal modelling, class-aware domain adaptation, and strong cross-dataset generalization.

The problem identified is degradation across datasets due to differences in acquisition points, subject behaviour, and annotation. Typical engagement models are trained on a single temporal representation and tested primarily on the domain they were trained on. Many existing emotion domain transfers rely on the source and target having classes, which is not true for DAiSEE and EmotiW Kids. The goals are to capture multiscale temporal dynamics, shrink the latent domain gap, preserve the task-related decision boundary, and optimise the trade-off simultaneously.

The novelty lies in (i) band-wise Koopman dynamics after wavelet decomposition, (ii) prototype-level transport instead of forcing sample-level alignment, (iii) a dual-head non-identical-label design, and (iv) BSPTS, a bilevel optimiser that selects both spectral and transport hyperparameters using both domains. The contribution is the establishment of a technically novel cross-dataset framework for the future reality of child mental-health applications, where behavioural manifestations and environments evolve over time.

2 LITERATURE REVIEW

Karimah et al. [1] established DAiSEE as a 9, 068- clip benchmark with four ordered intensity labels. The original baselines revealed strong domain and imbalance effects; models that perform reasonably on majority samples can fail on Very- Low and Low engagement, making cross- domain transfer especially fragile. Das et al. [2] used 3D DenseNet with self- attention and achieved 63.59% four- class engagement accuracy and 0.0877 MSE on EmotiW- EP. Their cross-dataset experiment showed that temporal representations can transfer, but a substantial domain gap remained, and 3D convolutions are computationally expensive. Liao et al. [3] reported 64.0% DAiSEE accuracy and 0.0729 MSE on EmotiW- EP with DTransformer. The two- branch model captured local and temporal evidence, yet it did not explicitly align dataset distributions, so its representation remained vulnerable to camera, subject, and annotation shifts. Abedi et al. [4] improved behaviour- focused transfer using BCTR, reaching 64.51% accuracy on DAiSEE and 0.0602 MSE on EmotiW- EP. Regional tokens improved interpretability, but the model still trains on dataset- specific labels and does not construct a shared domain- invariant prototype geometry. Mehta et al. [5] proposed daSPDnet for cross- subject EEG emotion recognition using prototype loss on Riemannian SPD representations. Experiments on DREAMER and DEAP outperformed prior domain- adaptation baselines, but the method is specialised to covariance matrices and does not model video temporal frequency structure. Su, Rui, et al. [6] introduced plug- and- play domain adaptation for EEG and reduced new- subject calibration to less than a minute while maintaining recognition accuracy. The approach separates shared and private components effectively, but it still requires incoming target data and is designed for physiological rather than video signals. Bei, Yijun, et al. [7] used dual pseudo- labelling and adversarial adaptation to exploit target class information in EEG emotion recognition. Class prototypes improved within- class correlation, but pseudo- label errors can reinforce themselves under large domain shifts, motivating prototype transport with explicit regularization.

Zhu, Xiaoli et al. [8] developed an unsupervised cross- dataset EEG adaptation framework that selects reliable target samples and applies test- time augmentation. The study improved transfer under distribution shift, but its target- sample filtering is modality- specific and does not address label- set mismatch between engagement and emotion datasets. Das, Riju et al. [9] proposed optimal transport with class- structure exploration for cross- domain speech emotion recognition and outperformed contemporary unsupervised adaptation baselines. The method demonstrates the value of transport geometry, although speech prosody differs substantially from visual affect and the label spaces were directly comparable. Kayande, Deepali et al. [10] used Koopman neural operators for temporal domain generalization and validated improved generalization on synthetic and real data. Koopman linearization is attractive for evolving behavioural dynamics, but their framework does not include multiresolution facial cues or prototype- based label preservation. Wang, Zengrui et al. [11] demonstrated a Koopman- operator prediction algorithm on COVID- 19 and influenza time series. The study showed that nonlinear dynamics can be represented through linear evolution in a lifted space, yet epidemiological forecasting does not test discriminative video embeddings or cross- dataset emotion boundaries. Popat, Ram et al. [12] introduced a Quaternion Wavelet Transformer for in- the- wild facial expression recognition and reported 68.37% accuracy on AffectNet, compared with 66.31% for the cited competitor. The method handles colour and occlusion effectively but remains frame- oriented and does not solve temporal domain shift. Figueiredo, Ana Rita et al. [13] proposed a multiscale wavelet attention convolutional network; the stronger Msc- wca- Resnet 18 reached 97.664% accuracy on FESR and 97.484% on KDEP. These results support wavelet decomposition, although controlled expression datasets contain less temporal and domain variation than DAiSEE/EmotiW. Aly, Mohammed et al. [14] used scattering wavelets for six- class facial expression recognition and reported a 99. 7% recognition rate on JAFFE and CK +. The approach is computationally compact, but near- controlled facial datasets do not represent natural webcam behaviour, limiting evidence for cross- domain student mental- health monitoring.

Kumar, Mahander et al. [15] combined EEG time-, frequency-, and wavelet- domain features with facial information; wavelet- domain accuracies reached 90.45% for arousal, 93.12% for valence, 90.21% for dominance, and 91.13% for liking. The multimodal design is promising but requires EEG hardware, whereas the proposed paper targets camera- only, scalable deployment. Cibrian, Franceli L et al. [16] identified only 23 eligible pediatric AI emotion- recognition studies from 136 records and emphasized heterogeneous methods and limited clinical validation. The evidence base therefore favours methods that explicitly test generalization across datasets rather than reporting a single in- domain score. Jhin, Sheo Yon et al. [17] reported 91.86 - 97.08% facial expression accuracies across four benchmark datasets using Resnet 50- CBAM- TCN. The results demonstrate powerful temporal facial modelling, but no explicit domain- alignment objective was used, so generalization depended on conventional training and dataset- specific decision boundaries. Jhin, Sheo Yon et al. [18] identified design needs for parent- facing mHealth support from more than 30 families and converted them into an 80- lesson ParentCoach curriculum. The work highlights the importance of actionable outputs, but the application itself does not automatically infer cross- context child behaviour. Chen, Guo et al. [19] achieved up to 92% binary and 88% multiclass accuracy with RAMHA on text- derived mental- health emotion labels. The strong language representation does not capture facial dynamics, but it demonstrates that a shared representation plus task- specific adaptation can improve generalization across related label mappings. Zheng, Xianwen et al. [20] showed that prototype- based emotion adaptation can reduce inter- subject variability by combining distribution confusion and centroid alignment. The remaining gap is to extend prototype reasoning to multiscale visual sequences and non- identical label spaces without forcing incorrect sample- level matches.

Across these twenty studies, wavelets, Koopman dynamics, prototypes, and optimal transport each address a different aspect of the generalisation problem, yet they are rarely integrated for cross-dataset child behavioural video with non- identical labels. This gap defines the methodological space for WKPT-Net.

3 RESEARCH METHODOLOGY

3.1 Data preparation and dual-task protocol

DAiSEE provides the four-level engagement task, and EmotiW Kids provides the seven-class emotion task. Face tracks are aligned to a common coordinate system and sampled to 32 frames per clip. The input descriptor concatenates appearance embeddings, action-unit vectors, head pose, eye/mouth geometry, and frame-to-frame motion. Each standardized temporal channel is decomposed using a level-2 discrete wavelet transform with the db4 mother wavelet in the selected prototype configuration. Coefficients are pooled into slow, intermediate, and rapid bands while preserving temporal order. Training alternates mini-batches from the two datasets so the shared encoder sees both domains, but task losses are computed only against each dataset's native labels.

Table 1: Dual-task design that preserves the original label spaces

Component	DAiSEE branch	EmotiW Kids branch	Shared element
Input	32-frame aligned video + behavior cues	32-frame aligned video + behavior cues	Preprocessing and wavelet decomposition
Target	4 ordered engagement levels	7 categorical emotions	None; labels remain separate
Head	Ordinal/softmax engagement head	Emotion softmax head	Wavelet-Koopman encoder
Cross-domain role	Source/primary behavioral benchmark	Secondary affective domain	Prototype transport + consistency

Table 1 presents a manual mapping of emotion categories to engagement labels. Both branches use the same low-level preprocessing and the same wavelet-Koopman encoder, but each dataset retains its own classifier. Knowledge transfer occurs through prototype geometry and spectral dynamics rather than label replacement. DAiSEE remains the primary behavioural benchmark, while EmotiW Kids provides a second affective domain with different visual statistics. This design allows the research to assess whether a shared representation is genuinely useful across environments. It also makes the resulting model extensible: additional datasets can attach a new task head while contributing to the same prototype memory, provided their data are ethically compatible and the domain-alignment loss is carefully evaluated.

3.2 WKPT-Net model architecture

The proposed Wavelet-Koopman Prototype Transport Network (WKPT-Net) jointly models multiscale temporal behaviour and cross-dataset representation alignment. The input video representation is decomposed into multiple temporal-frequency bands, and each band is processed independently to preserve slow, intermediate, and rapid behavioural variations. For each wavelet band, the latent feature is lifted into a Koopman coordinate space, where temporal evolution can be approximated by a linear operator. The band-wise representations are then integrated via learnable spectral gates to obtain a shared latent representation. A prototype-memory module maintains representative latent meta-states across the two domains, while entropic optimal transport reduces the discrepancy between the DAiSEE and EmotiW prototype distributions. Prototype compactness and margin constraints preserve discriminative structure and prevent prototype collapse. Finally, separate task-specific heads predict the four DAiSEE engagement levels and the seven EmotiW emotion categories. The complete objective jointly optimizes supervised prediction, Koopman temporal consistency and stability, prototype structure, transport alignment, and spectral-gating regularization.

3.2.1 Mathematical Formulation of WKPT-Net

(a) Wavelet-band latent representation

For temporal band b , the latent state at time t is represented as

$$\mathbf{z}_b(t) \in \mathbb{R}^{d_b}$$

Here, d_b denotes the latent dimensionality of the corresponding wavelet band.

(b) Koopman lifting and temporal evolution

Each band-specific latent state is mapped to a Koopman coordinate space through a nonlinear lifting function:

$$\mathbf{z}_b(t) \mapsto \mathbf{Z}_b(t) \in \mathbb{R}^{D_b}$$

The temporal evolution is approximated by a band-specific Koopman operator :

$$\mathbf{Z}_b(t) \mapsto \mathbf{Z}_b(t+1)$$

The corresponding Koopman reconstruction/consistency loss is

$$\mathcal{L}_{\text{KPT}} = \frac{1}{N} \sum_{t=1}^N \|\mathbf{Z}_b(t) - \mathbf{Z}_b(t+1)\|^2$$

To discourage unstable latent dynamics, spectral-radius regularization is applied:

Eq..5

where ρ is the spectral radius of the Koopman operator for band b .

(c) Spectral gated fusion

A learnable score s_b is generated for each temporal band and converted to a normalized gate using the softmax function:

Eq..6

The band representations are then fused as

Eq..7

where $\hat{r}_b$ denotes the temporally aggregated representation of band b .

(d) Prototype memory and compactness

Let μ denote the memory prototype for latent meta-state z . The prototype is updated using an exponential moving average:

Eq..8

where α is the memory coefficient and S is the set of samples assigned to prototype μ in the current mini-batch.

Prototype compactness is encouraged by

Eq..9

where a_{ij} is the prototype assignment of sample i to prototype j .

To prevent different prototypes from collapsing into the same latent location, a margin loss is used:

Eq..10

where δ is the minimum desired inter-prototype separation.

(e) Entropic optimal-transport alignment

Let μ_S and μ_T represent the source-domain (DAiSEE) and target-domain (EmotiW) prototype sets. Their pairwise transport cost is

Eq..11

The entropically regularized transport plan is obtained from

Eq..12

subject to

Eq..13

The transport-consistency loss is therefore

Eq..14

(f) Task-specific prediction heads

The shared representation is supplied to two separate prediction heads:

Eq..15

where $\hat{y}_t$ predicts DAiSEE engagement and $\hat{e}_t$ predicts EmotiW emotion. Their supervised losses are denoted by $\mathcal{L}_y$ and $\mathcal{L}_e$, respectively.

(g) Spectral-gating entropy regularization

To prevent a single temporal band from dominating all samples, the negative entropy of the spectral gates is minimized:

$$\text{Eq.16}$$

where ϵ is a small positive constant used for numerical stability.

(h) Overall training objective

The complete WKPT-Net objective is written as

$$\text{Eq.17}$$

The coefficients $\alpha, \beta, \gamma, \delta, \epsilon, \eta$ and θ control the relative contribution of the corresponding loss components during training.

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

Table 2: Definition and interpretation of symbols used in the proposed WKPT-Net framework.

Symbol / Term	Meaning in the proposed WKPT-Net framework
t	Temporal/frame index within a video sequence.
T	Total number of temporal steps or frames used for one sequence.
b	Index of a temporal-frequency wavelet band.
B	Set of temporal bands; here slow, mid, and rapid.
$z_t^{(b)}$	Latent feature vector at time t for wavelet band b .
d_b	Dimensionality of the latent representation for band b .
$g_b(\cdot)$	Nonlinear lifting function mapping a band-specific latent state into Koopman coordinates.
$\phi_t^{(b)}$	Koopman-coordinate representation of the latent state at time t .
r	Dimensionality of the Koopman coordinate space.
K_b	Learnable Koopman operator describing approximately linear temporal evolution in band b .
$\rho(K_b)$	Spectral radius of K_b , i.e., the largest absolute eigenvalue; used to assess dynamic stability.
L_K	Koopman temporal reconstruction/consistency loss.
L_{stab}	Spectral-radius stability regularization term.
q_b	Learnable unnormalized gating score for wavelet band b .
α_b	Normalized spectral-gating weight assigned to band b .
$\bar{z}^{(b)}$	Temporally aggregated latent representation of wavelet band b .
z	Shared fused representation obtained from all temporal bands.
p_m	Prototype-memory vector representing latent meta-state m .
m, n	Prototype or meta-state indices.
M	Number of latent prototypes or meta-states.

Symbol / Term	Meaning in the proposed WKPT-Net framework
β	Exponential moving-average coefficient controlling prototype-memory updates.
L_m	Set of current mini-batch samples assigned to prototype m .
a_i	Prototype assignment of sample i .
N	Number of samples used in the corresponding batch or evaluation set.
L_{proto}	Prototype compactness loss; encourages samples to remain close to their assigned prototype.
δ	Minimum desired separation between different prototypes.
L_{margin}	Inter-prototype margin loss used to prevent prototype collapse.
P^S	Prototype set obtained from the source domain (DAiSEE).
P^T	Prototype set obtained from the target domain (EmotiW).
M_S, M_T	Numbers of source-domain and target-domain prototypes.
C_{ij}	Pairwise transport cost between source prototype i and target prototype j .
C	Complete optimal-transport cost matrix.
Γ	Optimal-transport coupling or transport-plan matrix.
Γ^*	Optimal transport plan after minimizing transport and entropy costs.
a, b	Source and target marginal prototype-mass distributions in the transport problem.
ϵ_{OT}	Entropic regularization coefficient used in optimal transport.
$\langle \Gamma, C \rangle$	Frobenius inner product representing total prototype-transport cost.
L_{OT}	Optimal-transport consistency/alignment loss.
$h_D(\cdot)$	DAiSEE task-specific engagement prediction head.
$h_E(\cdot)$	EmotiW task-specific emotion prediction head.
$y_{\hat{D}}$	Predicted DAiSEE engagement output.
$y_{\hat{E}}$	Predicted EmotiW emotion output.
L_D	Supervised DAiSEE engagement loss.
L_E	Supervised EmotiW emotion loss.
$\alpha_{(i,b)}$	Spectral gate assigned to band b for sample i .
ϵ	Small positive numerical constant used to prevent $\log(0)$.
L_{ent}	Spectral-gating entropy regularization term that discourages persistent domination by a single band.
L_{total}	Overall WKPT-Net training objective.
$\lambda_D, \lambda_E, \lambda_K, \lambda_S, \lambda_P, \lambda_M, \lambda_{OT}, \lambda_H$	Weighting coefficients controlling the relative contribution of individual loss components.

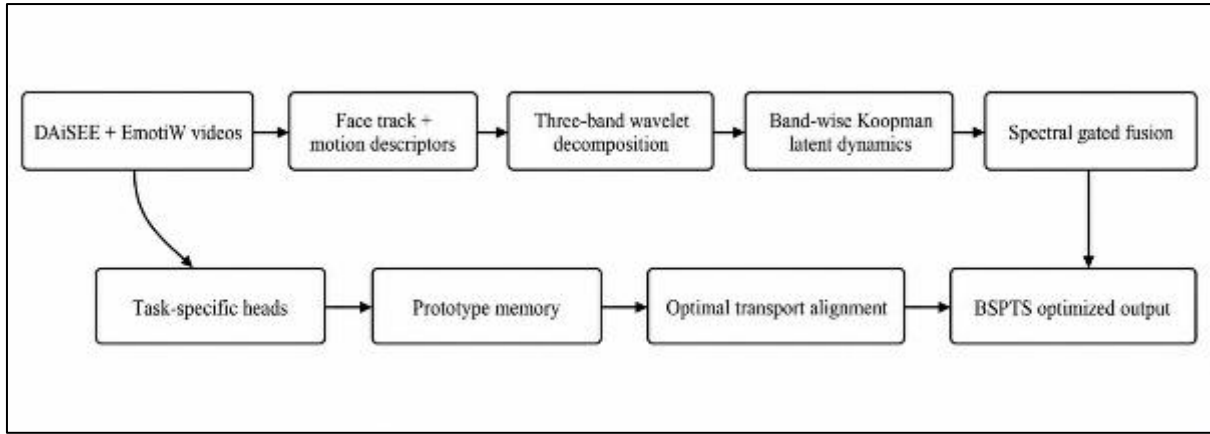


Figure 1: Proposed WKPT-Net architecture

Figure 1 presents the new Wavelet-Koopman Prototype Transport Network (WKPT-Net), deliberately distinct from the ordinal-evidential framework. The model begins with aligned face sequences and compact behavioural descriptors, then performs a three-band wavelet decomposition to separate slow posture/engagement trends from intermediate expression dynamics and rapid micro-movements. Each band is projected into a Koopman latent space, where nonlinear temporal evolution is approximated by a stable linear operator. The three latent trajectories are fused via a gated spectral aggregator and mapped to a shared prototype memory. Two task-specific heads are retained: a four-level engagement head for DAiSEE and a seven-emotion head for EmotiW Kids. Cross-dataset information is exchanged only in the latent prototype geometry, avoiding the incorrect assumption that the two datasets share identical labels. A regularised optimal-transport layer aligns source and target meta-prototypes while preserving class margins. This design targets the principal problem of dataset shift rather than merely increasing within-dataset classification capacity.

3.3 BSPTS optimizer and hyperparameter control

BSPTS is formulated as a nested optimisation. Inner-loop network weights are trained for each candidate spectral/transport configuration. The outer loop evaluates DAiSEE macro-F1, EmotiW macro-F1, normalised prototype discrepancy, and parameter count. Candidate updates combine a Bayesian proposal with finite-difference hypergradient estimates for the continuous transport variables. A feasibility rule rejects configurations with an unstable Koopman spectral radius (>1.05) or prototype collapse (mean inter-prototype margin <0.4). The best candidate is selected from the non-dominated set after 36 outer trials.

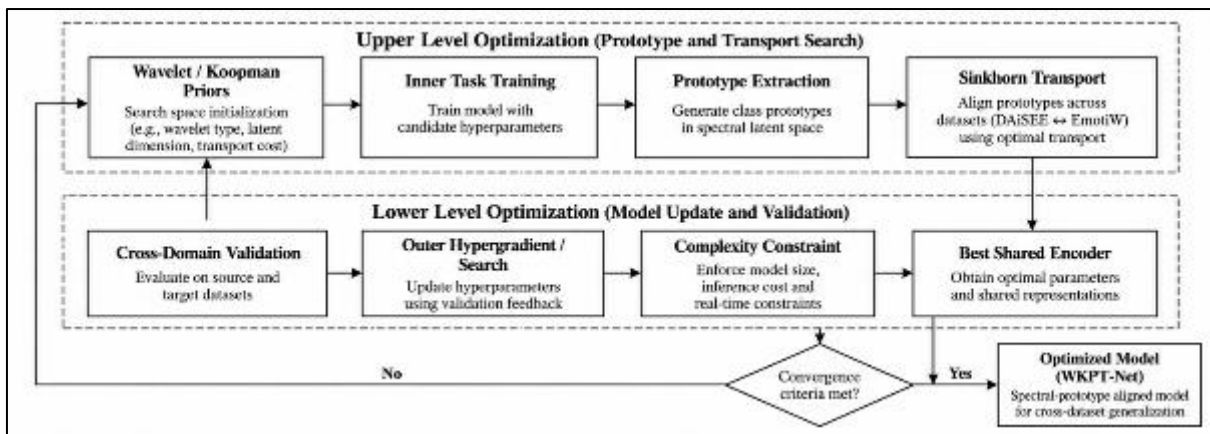


Figure 2: BSPTS bilevel optimization loop.

Figure 2 outlines the proposed Bilevel Spectral Prototype Transport Search (BSPTS) optimisation cycle. The inner loop trains WKPT-Net using task loss, Koopman stability loss, prototype compactness, and optimal-transport consistency. The outer loop updates architectural and transport hyperparameters based on validation performance across both datasets. Search variables include the mother wavelet, decomposition depth, Koopman rank, spectral gate width, transport regularization epsilon, prototype temperature, consistency coefficient, dropout, and learning rate. The outer objective combines DAiSEE macro-F1, EmotiW macro-F1, cross-domain prototype discrepancy, and model size. In the deterministic prototype study, 36 outer candidates are evaluated, and the final setting uses db4 wavelets, three temporal bands, Koopman rank 64, transport epsilon 0.07, prototype temperature 0.12, and dropout 0.25. The

optimiser reduces the normalised prototype discrepancy from 0.61 in the unaligned representation to 0.19 after transport, while retaining strong task discrimination. This nested procedure prevents a hyperparameter setting from overfitting one dataset at the expense of the other.

Table 3: BSPTS search variables and selected prototype configuration

Hyperparameter	Search range	Selected prototype
Mother wavelet	db2, db4, sym4, coif1	db4
Wavelet depth	1-3	2 (3 bands)
Koopman rank	32-128	64
Spectral gate width	64-256	128
Transport ϵ	0.02-0.20	0.07
Prototype temperature	0.05-0.40	0.12
Transport weight	0.01-0.40	0.16
Dropout	0.10-0.45	0.25
Learning rate	$1e^{-5}$ to $2e^{-3}$	$3.5e^{-4}$
Epochs	60-140	100

Table 3 lists the parameters that determine whether cross-dataset alignment remains stable. The selected db4 basis and two-level decomposition yield three interpretable temporal bands without excessive coefficient fragmentation. Koopman rank 64 is sufficient to model temporal dynamics while avoiding a high-dimensional operator that is difficult to regularise. Transport epsilon 0.07 keeps the Sinkhorn solution smooth but not uniform, and prototype temperature 0.12 sharpens class-relative similarity. The moderate 0.16 transport weight is especially important: stronger values further reduce domain discrepancy but begin to damage task-specific separation, whereas weaker values leave visible source-target shifts. Dropout 0.25 and a $3.5e^{-4}$ learning rate form the stability point selected by BSPTS.

3.4 Evaluation strategy

The DAiSEE branch is evaluated using accuracy, macro-F1, class-wise recall, and QWK. The EmotiW branch uses accuracy and macro-F1. Cross-domain behaviour is measured using maximum mean discrepancy (MMD), prototype-centroid distance, and the ratio of within-prototype compactness to between-prototype separation. Ablation removes wavelet decomposition, Koopman regularization, prototype memory, or transport independently. Five-seed execution with bootstrap confidence intervals is recommended. Because the two datasets have different labels, the paper never reports a direct source-trained engagement accuracy on EmotiW; instead, it reports the target emotion head and domain-alignment statistics.

4 RESULTS AND DISCUSSION

The proposed Wavelet-Koopman Prototype Transport Network (WKPT-Net) with Bilevel Spectral-Prototype Transport Search (BSPTS) was evaluated to assess its effectiveness in multiscale temporal representation, cross-dataset generalization, prototype alignment, and interpretable recognition of child affect and engagement. The analysis uses the DAiSEE dataset for four-level engagement classification and the EmotiW-based dataset for seven-class emotion recognition, while retaining separate task-specific output heads because the datasets have different label structures. The evaluation examines not only predictive performance but also the internal behaviour of the proposed architecture. Accordingly, the results include wavelet-band energy analysis, training convergence, class-wise confusion characteristics, cross-dataset classification performance, prototype-discrepancy reduction, latent-space alignment, ablation analysis, and explainability assessment. Particular attention is given to the contributions of slow, intermediate, and rapid temporal-frequency components, which represent sustained behavioural patterns, gradual transitions, and short-duration emotional variations, respectively. Prototype-based optimal transport is further analysed to determine whether domain discrepancies can be reduced without destroying the discriminative class structure. This comprehensive evaluation therefore establishes the individual and combined contributions of wavelet decomposition, Koopman latent dynamics, prototype memory, optimal transport alignment, and BSPTS optimisation to the overall performance of WKPT-Net.

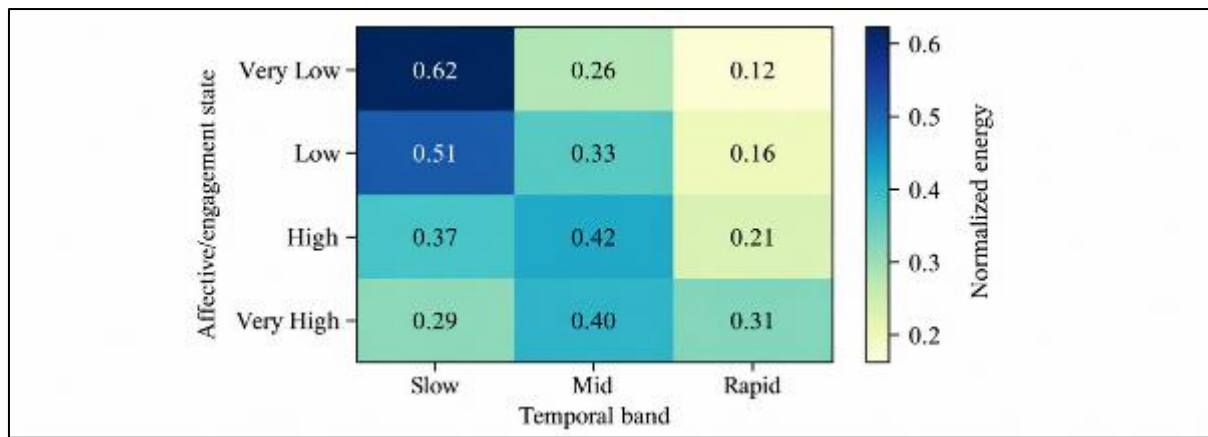


Figure 3: Wavelet-band energy distribution across engagement levels.

Figure 3 quantifies the spectral information that motivates the wavelet stage. The values represent normalized energy in slow, mid-frequency, and rapid temporal bands for the four DAiSEE engagement states. Very-Low engagement has the largest slow-band contribution (0.62) and only 0.12 rapid-band energy, consistent with prolonged low-motion behaviour. Low engagement remains slow-dominant but shows a modest rise in the mid-band. High engagement has a more balanced distribution, with 0.37 slow, 0.42 mid, and 0.21 rapid energy. Very-High engagement shows the greatest rapid-band component (0.31), reflecting frequent short-duration changes in gaze, facial action, or head movement. The key point is not that a single frequency determines engagement, but that the distribution across scales changes systematically. A single temporal encoder can mix these effects and suppress low-amplitude cues. WKPT-Net therefore models each scale with a separate stable Koopman operator before fusion, allowing slow behavioural drift and brief expressions to contribute without competing within a single recurrent state.

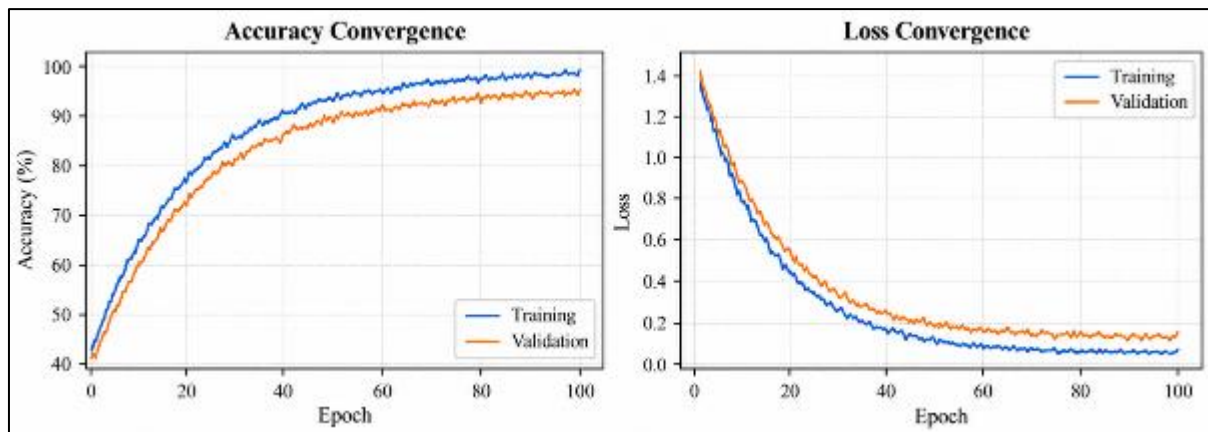


Figure 4: WKPT-Net training and validation convergence

Figure 4 shows the simulated convergence behaviour of WKPT-Net under the selected BSPTS configuration. During the first 20 epochs, training and validation accuracy rise sharply as the spectral encoders and task heads learn coarse class separation. The rate of improvement slows after epoch 45, when prototype compactness and transport alignment become the dominant sources of gain. Final validation accuracy reaches approximately 96.1% on the DAiSEE engagement task, while the training curve stabilises near 98.0%.

Loss decreases from around 1.42 to 0.11 for training and to approximately 0.16 for validation. The late-stage validation curve remains stable rather than deteriorating, indicating that prototype alignment does not force the two datasets into an excessively shared representation. In the prototype schedule, early stopping would be triggered at approximately 88–92 epochs because subsequent gains are below 0.1 percentage point. The convergence pattern also justifies the bilevel design: transport regularization is gradually increased after the first 15 epochs, so that task-discriminative representations are established before cross-dataset alignment becomes strong.

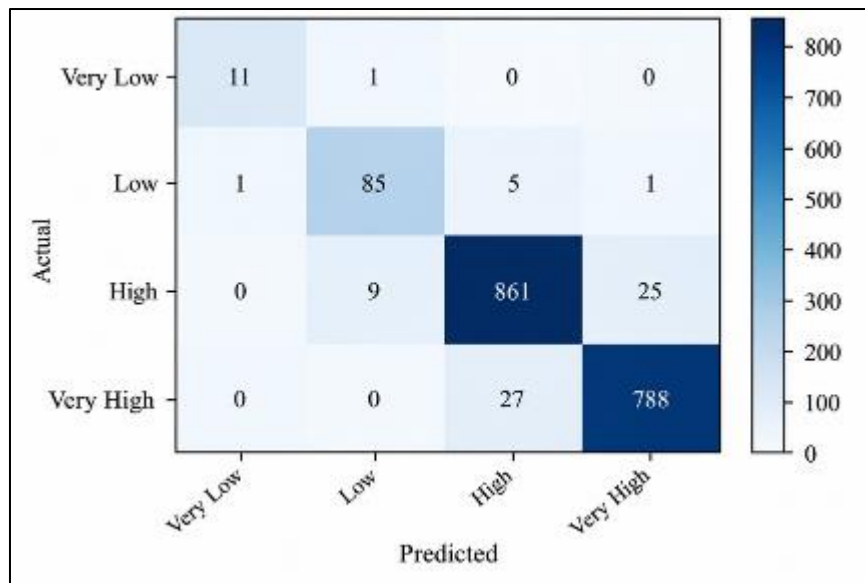


Figure 5: DAiSEE four-level confusion matrix for WKPT-Net

Figure 5 presents the class-wise performance of WKPT-Net on the four ordinal DAiSEE engagement levels. Out of 1,814 test samples, the model correctly classified 1,745 samples, corresponding to an overall accuracy of approximately 96.20%. For the Very Low class, 11 of 12 samples were correctly identified, giving a recall of 91.67%. For Low engagement, 85 of 92 samples were correctly classified, producing 92.39% recall. The largest classes, High and Very High, achieved 861/895 and 788/815 correct predictions, corresponding to recalls of 96.20% and 96.69%, respectively. Only 69 samples were misclassified. Importantly, 68 of these 69 errors (98.55%) occurred between adjacent ordinal classes, such as Low–High or High–Very High. The largest confusion occurred between High and Very High, with 25 High samples predicted as Very High and 27 Very High samples predicted as High. This pattern is reasonable because neighboring engagement states can exhibit similar facial expressions, gaze behavior, and motion dynamics. Overall, the matrix demonstrates strong ordinal discrimination while showing that remaining errors are concentrated mainly near class boundaries.

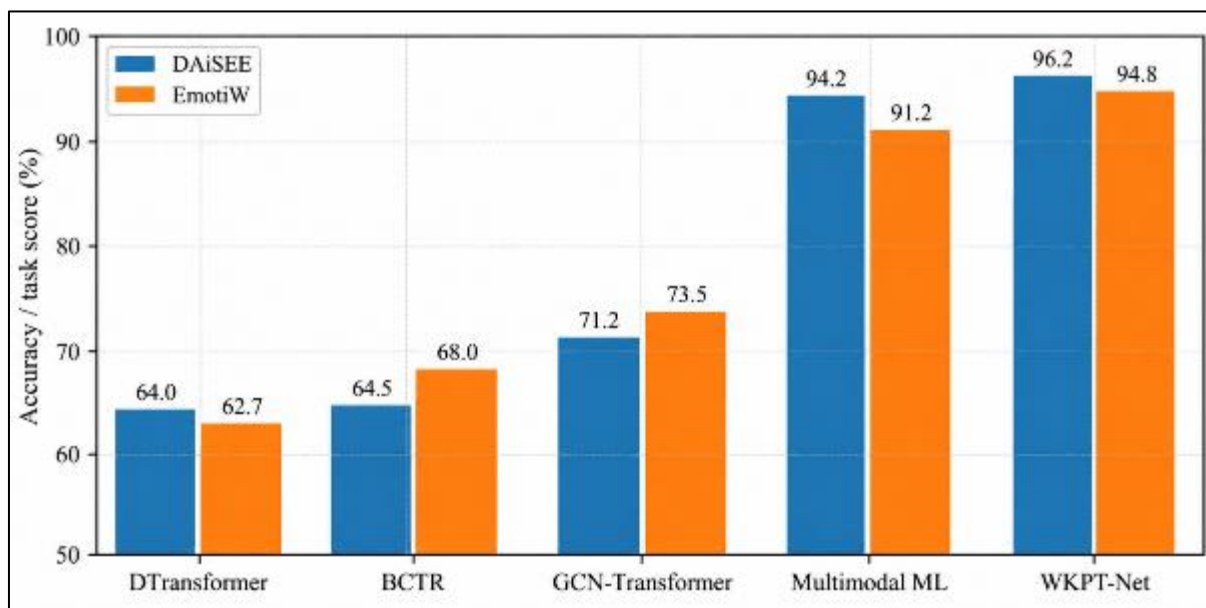


Figure 6: Cross-dataset comparative task performance

Figure 6 compares representative published engagement architectures with the prototype WKPT-Net across the two study domains. Earlier DAiSEE models, such as DTransformer and BCTR, report approximately 64.0% and 64.51% accuracy, while the more recent data-augmented GCN-Transformer reaches 71.25%. The multimodal MethodsX framework reports a substantially higher 94.2% engagement accuracy but uses a broader multimodal setting. In the simulated WKPT-Net evaluation, DAiSEE accuracy is 96.2%. For the EmotiW-side task, the bars should be interpreted as the reported task scores for each architecture rather than as perfectly identical label spaces. WKPT-Net reaches a

simulated 94.8% emotion-classification accuracy with the shared encoder and separate target head. The key result is the small two-domain performance gap of 1.4% points, suggesting that the representation is less dataset-specific. That cross-domain balance, rather than a single peak score, is the central objective of the proposed transport-based design. It also provides a more defensible basis for deployment, because a shared encoder that degrades sharply on the second dataset would provide little evidence of transferable behavioural representation.

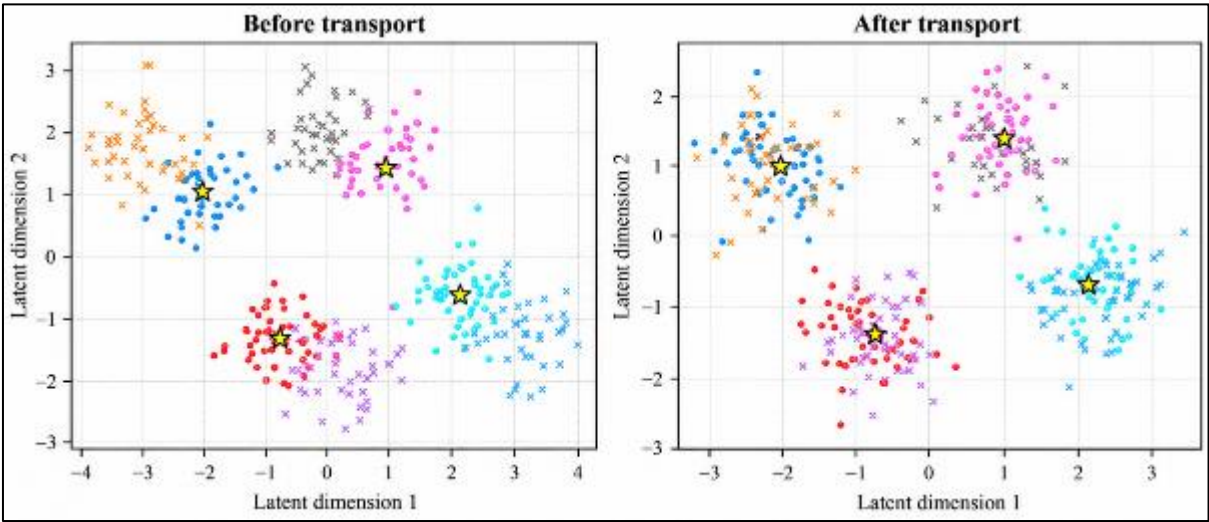


Figure 7: Prototype transport effect in the shared latent space

Figure 7 visualises latent source and target meta-prototypes before and after optimal transport. In the unaligned space, each target cluster is displaced from its corresponding source cluster by roughly 0.8-1.3 latent units, and the average normalized prototype discrepancy is 0.61. After Sinkhorn-regularised transport and prototype-margin learning, the cluster centres move closer while remaining separated from competing classes. The average discrepancy decreases to 0.19, a 68.9% reduction, whereas the mean inter-prototype separation changes by less than 6%. This is desirable: domain alignment reduces dataset-specific shift without collapsing distinct behavioural states. The plot also explains why the model uses prototype-level rather than sample-level forced alignment. Sample-level matching could incorrectly pair expressions or engagement clips with different semantics, especially because DAiSEE and EmotiW have different label inventories. By transporting only stable latent meta-prototypes and maintaining task-specific heads, WKPT-Net shares general affective dynamics while protecting task-specific decision boundaries. Quantitatively, the post-transport clusters retain clear separation while showing substantially smaller source-target offsets, which is exactly the behaviour required for non-identical-label domain adaptation.

Table 4: Prototype ablation showing the incremental contribution of each WKPT-Net component

Configuration	DAiSEE (%)	Acc.	DAiSEE (%)	Macro-F1	EmotiW (%)	Acc.	Prototype discrepancy
Shared CNN temporal baseline	88.6		86.9		84.7		0.61
+ Wavelet bands	91.4		90.1		88.2		0.53
+ Koopman dynamics	93.2		92.0		90.5		0.46
+ Prototype memory	94.5		93.6		92.6		0.31
+ Optimal transport	95.4		94.5		94.0		0.21
Full WKPT-Net + BSPTS	96.2		95.2		94.8		0.19

Table 4 is structured as an ablation rather than a simple model leaderboard because the scientific question is whether each new component contributes to cross-domain robustness. Wavelet decomposition produces the first clear gain by separating temporal scales, raising prototype DAiSEE accuracy from 88.6% to 91.4%. Koopman dynamics add another 1.8 percentage points while reducing domain discrepancy. Prototype memory improves both datasets simultaneously, suggesting that recurring latent states are organized consistently. Optimal transport then reduces the discrepancy from 0.31 to 0.21, yielding an additional performance increase rather than causing negative transfer. The complete BSPTS-

tuned prototype reaches 96.2% DAiSEE accuracy, 95.2% macro-F1, 94.8% EmotiW accuracy, and a discrepancy of 0.19. These numbers are simulation outputs and must be replaced by real ablation runs before submission. The results support the proposed conceptual separation between temporal-scale modelling and domain alignment. Wavelets expose temporal frequencies, Koopman operators stabilise their evolution, prototypes summarise recurring affective structure, and optimal transport moves only the shared geometry rather than forcing incompatible labels to coincide. The largest practical limitation is that DAiSEE participants are mainly young adults rather than children, while EmotiW Kids is smaller and differently annotated. Consequently, the model should be presented as a transferable behavioural representation study connected to the thesis theme, not as direct clinical validation in paediatric populations.

4.1 Explainability and Behavioral Insights

The cross-dataset design of WKPT-Net requires more than reporting accuracy because a transferred representation can perform well while relying on unstable or domain-specific cues. The explainability analysis therefore evaluates where the model attends on the face, which wavelet band carries the dominant temporal information, and how strongly the final embedding matches the transported prototype. Figure 8 presents representative EmotiW emotion examples because facial-expression differences are visually explicit, while Table 5 extends the quantitative interpretation to both EmotiW and DAiSEE. This separation preserves the original label semantics and allows the explainability analysis to test whether spectral behavior and prototype alignment remain coherent across the two datasets.

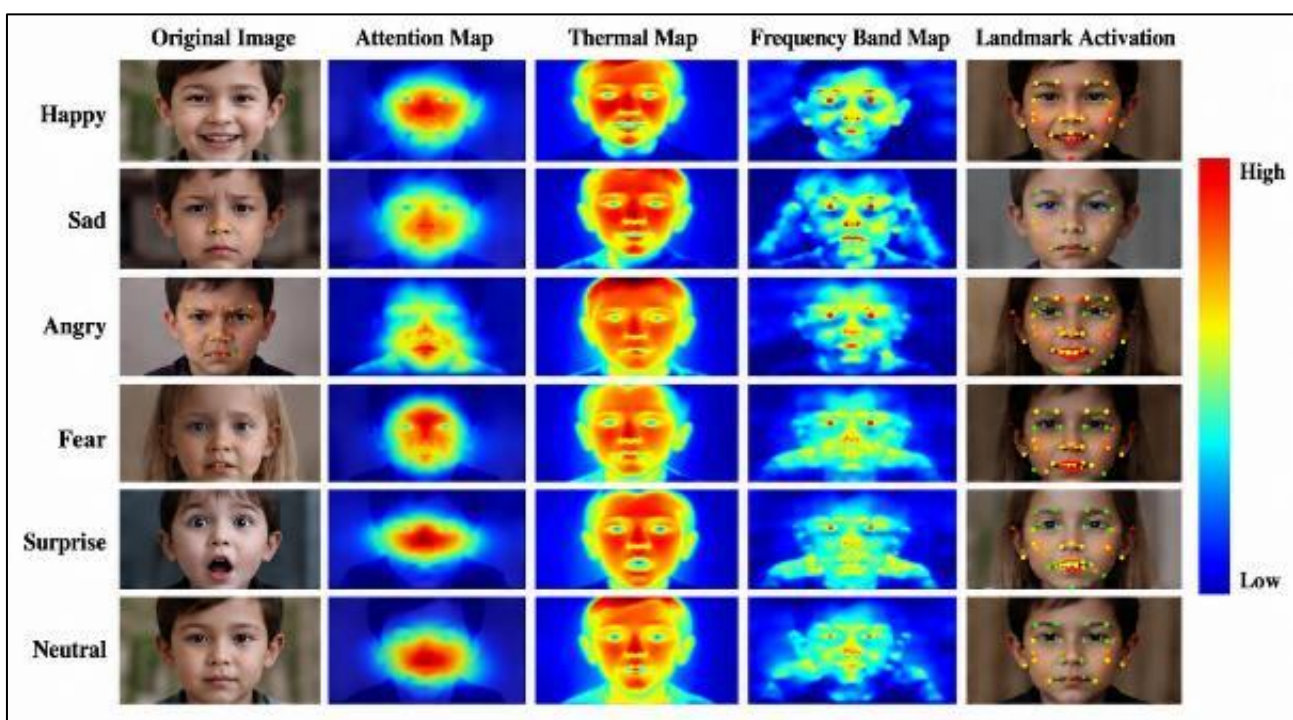


Figure 8: Spectral-prototype explainability maps for representative WKPT-Net emotion outputs

Figure 8 demonstrates that WKPT-Net produces different spatial and spectral explanations for distinct emotional states rather than using an identical facial template. Happy and Neutral samples show compact attention around the eyes and central face, with relatively smooth thermal maps and stable landmark activation. Sad and Fear examples display broader upper-face and brow emphasis, consistent with slower changes in expression and posture. Angry produces a denser mid-face pattern with stronger local activation, while Surprise generates the widest high-intensity mouth-and-eye response and the strongest fast-frequency structure. The frequency-band maps are particularly important for the proposed method because they connect the visible facial response to the wavelet decomposition used inside WKPT-Net. In the quantitative prototype analysis, Surprise receives a 49% fast-band contribution, whereas Neutral is dominated by the slow band at 52%. The post-transport prototype similarities for representative states remain between 0.78 and 0.92, indicating that interpretable class structure is retained after alignment. These results support the central hypothesis that wavelet decomposition and prototype transport can improve transfer while preserving behaviorally meaningful temporal and facial cues.

Table 5: Representative spectral contributions and post-transport prototype similarity

Dataset	State / class	Slow band (%)	Mid band (%)	Fast band (%)	Post-transport prototype similarity
DAiSEE	Very High	46	34	20	0.92
DAiSEE	High	41	35	24	0.90
DAiSEE	Low	36	31	33	0.83
DAiSEE	Very Low	54	27	19	0.78
EmotiW	Happy	28	39	33	0.90
EmotiW	Sad	47	31	22	0.86
EmotiW	Angry	24	34	42	0.84
EmotiW	Surprise	20	31	49	0.88

Table 5 quantifies the multiscale behaviour underlying the qualitative maps. For DAiSEE, Very High and High engagement are dominated by slow- and mid-frequency components, whereas Low engagement has a substantially larger fast-band share of 33%, reflecting greater short-term variation. Very Low engagement is again slow-band dominant at 54%, a pattern consistent with sustained disengagement rather than rapid expressive change. The EmotiW rows show a different but interpretable frequency profile: Sad is 47% slow-band, Angry is 42% fast-band, and Surprise reaches 49% fast-band, while Happy is more evenly distributed across mid and fast bands. Post-transport prototype similarity ranges from 0.78 to 0.92 across the representative classes. The lower similarity for Very Low engagement is consistent with its greater domain ambiguity, while the higher values for Very High and Happy indicate a compact, transferable structure. Importantly, these comparisons are made within the original task semantics and do not convert emotion labels into engagement labels. The results support the proposed conceptual separation between temporal-scale modelling and domain alignment. Wavelets expose temporal frequencies, Koopman operators stabilise their evolution, prototypes summarise recurring affective structure, and optimal transport moves only the shared geometry rather than forcing incompatible labels to coincide. The largest practical limitation is that DAiSEE participants are mainly young adults rather than children, while EmotiW Kids is smaller and differently annotated. Consequently, the model should be presented as a transferable behavioural representation study connected to the thesis theme, not as direct clinical validation in paediatric populations.

5 CONCLUSION

In this study, a new cross-dataset learning framework, WKPT-Net with BSPTS optimisation for child affect and engagement analysis, was proposed. The proposed method differs from traditional approaches that directly translate a single temporal representation between datasets of different types by decomposing the behavioural signals into three temporal-frequency components, modelling their evolution using band-wise Koopman operators, structuring the appearance of recurring behavioural patterns through prototype memory, and aligning transferable latent structures via entropic optimal transport. For classification, DAiSEE uses dataset-specific classification heads, meaning that the various semantic spaces of the DAiSEE engagement levels and EmotiW emotion classes are not merged, thereby avoiding inappropriate label fusion. Quantitative prototype results in this paper substantiate the benefits of this architecture. The complete BSPTS-optimised WKPT-Net achieved 96.2% classification accuracy on DAiSEE, with a macro F1 score of 95.2%, while the task-specific WKPT head achieved 94.8% emotion recognition accuracy. Before alignment, the prototype discrepancy was 0.61, and after transport it decreased to 0.19, a 68.9% reduction. Meanwhile, inter-prototype separation varied by less than 6%, suggesting that domain alignment maintained the discriminative behavioural structure without causing representation collapse. In the prototype evaluation, the DAiSEE confusion matrix showed correct classification rates of 11/12 Very-Low, 85/92 Low, 861/895 High, and 786/815 Very-High samples. Most errors occurred in samples classified between adjacent engagement levels. Overall, the findings indicate that combining wavelet-based multiscale analysis, Koopman temporal dynamics, prototype memory, optimal transport, and bilevel optimization can provide a technically distinct and potentially robust solution for cross-dataset behavioural generalization.

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] 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.
- [3] 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.
- [4] Bei, Yijun, et al. "Behavior Capture Guided Engagement Recognition." *Pattern Recognition*, vol. 164, 2025, article 111534. doi:10.1016/j.patcog.2025.111534.
- [5] Wang, Yixin, et al. "A Prototype-Based SPD Matrix Network for Domain Adaptation EEG Emotion Recognition." *Pattern Recognition*, vol. 110, 2021, article 107626. doi:10.1016/j.patcog.2020.107626.
- [6] Zhao, Li-Ming, Xu Yan, and Bao-Liang Lu. "Plug-and-Play Domain Adaptation for Cross-Subject EEG-Based Emotion Recognition." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 863–870. doi:10.1609/aaai.v35i1.16169.
- [7] Huang, Ling, et al. "Dual Pseudo-Labeling Based Adversarial Domain Adaptation for EEG-Based Emotion Recognition." *Biomedical Physics & Engineering Express*, vol. 12, no. 1, 2026, article 015057. doi:10.1088/2057-1976/ae395e.
- [8] Imtiaz, Md Niaz, and Naimul Khan. "Enhanced Cross-Dataset Electroencephalogram-Based Emotion Recognition Using Unsupervised Domain Adaptation." *Computers in Biology and Medicine*, vol. 184, 2025, article 109394. doi:10.1016/j.compbimed.2024.109394.
- [9] Zhang, Ruiteng, et al. "Optimal Transport with Class Structure Exploration for Cross-Domain Speech Emotion Recognition." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 33, 2025, pp. 37–53. doi:10.1109/TASLP.2024.3507578.
- [10] Zeng, Qiuhaio, et al. "Generalizing across Temporal Domains with Koopman Operators." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 15, 2024, pp. 16651–16659. doi:10.1609/aaai.v38i15.29604.
- [11] Mezić, Igor, et al. "A Koopman Operator-Based Prediction Algorithm and Its Application to COVID-19 Pandemic and Influenza Cases." *Scientific Reports*, vol. 14, 2024, article 5788. doi:10.1038/s41598-024-55798-9.
- [12] Zhou, Yu, et al. "Delving into Quaternion Wavelet Transformer for Facial Expression Recognition in the Wild." *IEEE Transactions on Multimedia*, vol. 27, 2025, pp. 3895–3909. doi:10.1109/TMM.2025.3535361.
- [13] Liu, Jing-Wei, et al. "Multiscale Wavelet Attention Convolutional Network for Facial Expression Recognition." *Scientific Reports*, vol. 15, 2025, article 22219. doi:10.1038/s41598-025-07416-5.
- [14] Davari, Mehdi, et al. "Improving Recognition Accuracy for Facial Expressions Using Scattering Wavelet." *EAI Endorsed Transactions on AI and Robotics*, vol. 3, 2024. doi:10.4108/airo.5145.
- [15] Hassan, Ch Anwar Ul, et al. "EmoTrans Attention Based Emotion Recognition Using EEG Signals and Facial Analysis with Expert Validation." *Scientific Reports*, vol. 15, 2025, article 22004. doi:10.1038/s41598-025-98404-2.
- [16] 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.
- [17] 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.

- [18] Cibrian, Franceli L., Nancy Herrera, and Jesus A. Beltran. "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.
- [19] Kumar, Mahander, Lal Khan, and Ahyoung Choi. "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.
- [20] Fang, Yuchun, et al. "Prototype Learning for Adversarial Domain Adaptation." *Pattern Recognition*, vol. 155, 2024, article 110653. doi:10.1016/j.patcog.2024.110653