RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2651

Adaptive Komodo Mlipir–Optimized Spatiotemporal Graph Neural Network for Dataset-Specific Physiological-State Classification

R. Kishore Kanna [a,*], Biswajit Brahma [b], N. Aravindha Babu [c], Ramesh Kumar Ayyasamy [d], Bharath Kumar Nagaraj [e], Ayodeji Olalekan Salau [f]

[a] Department of Biomedical Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu 600062, India

[b] McKesson Corporation, 32559 Lake Bridgeport Street, Fremont, California 94555, USA

[c] Department of Oral and Maxillofacial Pathology, Sree Balaji Dental College & Hospital, Bharath Institute of Higher Education and Research, Chennai, India

[d] Department of Information Systems, Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman (UTAR), Malaysia

[e] AI, EXL Business Consulting and Services, New York City, USA

[f] Department of Electrical/Electronics and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria

* Corresponding Author Email: kishorekanna007@gmail.com

DOI: https://doi.org/10.54392/irjmt2651

Received: 25-04-2026; Revised: 22-07-2026; Accepted: 29-07-2026; Published: 10-08-2026

Abstract: Physiological-state classification requires models that can represent dependencies among sensor channels while preserving temporal variation. This study presents a dataset-specific evaluation of an Adaptive Komodo Mlipir Algorithm-optimized spatiotemporal graph neural network (AKMA-ST-GNN) using four public physiological datasets with different modalities and target definitions: STEW for cognitive workload, SEED for emotion, DEAP for affect, and WESAD for stress and amusement. The datasets were analyzed independently. No EEG signal from one dataset was combined with EMG or peripheral signals from another dataset, and workload, emotion, affect, and stress labels were not collapsed into a common target. The proposed workflow included signal filtering, training-set standardization, frequency-domain characterization where physiologically appropriate, within-dataset channel graph construction, spatiotemporal graph learning, and dataset-specific classification. AKMA was used as a hyperparameter-search wrapper rather than as an additional predictive layer. Under the preliminary 80% training and 20% testing protocol preserved in the source manuscript, the reported accuracies were 98.60% for STEW, 98.31% for WESAD, 98.06% for DEAP, and 98.12% for SEED. Precision, F1-score, specificity, and sensitivity were available only for selected datasets. Prediction-level outputs, class supports, confusion matrices, and decision scores were not retained, therefore, macro F1, weighted F1, balanced accuracy, Matthews correlation coefficient, Cohen's kappa, and ROC/AUC could not be reconstructed. Repeated-seed results and fold- or participant-level scores were also unavailable, preventing calculation of standard deviations, confidence intervals, and statistical significance tests. The reported values are consequently interpreted as preliminary evidence of technical feasibility rather than proof of stable or subject-independent generalization.

Keywords: Cognitive Workload, Stress Recognition, Spatiotemporal Graph Neural Network, Komodo Mlipir Optimization, Multimodal Bio Signals.

1. Introduction

Stress, cognitive workload, and affect are psychophysiological phenomena that are related, but not interchangeable. Stress is a response to a perceived demand or threat; cognitive workload is the mental resources needed to complete a task; and affect is the dimensions or categories of valence, arousal, and discrete emotional states. Physiological expressions may overlap, as the autonomic, endocrine, cortical, and muscular systems interact in response to variations in task demands and emotional context [1]. But a model that can distinguish workload levels is not necessarily a stress detector, and an emotion classifier is not necessarily a prospective stress predictor [2]. This distinction is crucial for creating valid models and for clinically responsible interpretation [3].

EEG is a high temporal resolution measure of the electrical activity of the cortex [4]. Electromyography (EMG) measures muscle activation; electrodermal activity, photoplethysmography, respiration, and temperature measure peripheral or autonomic responses; and electrocardiography and inertial signals measure peripheral or autonomic responses. Each modality is physiologically distinct, with distinct noise profiles, sampling rates, and sensitivity to experimental context [5]. Multimodal analysis is thus only useful if signals are synchronized within the participant and across trials. Because EEG and EMG data from different datasets are unrelated and lack a common temporal and participant reference, neural–muscular coupling cannot be determined by combining these data.

High classification performance has been reported for stress and emotion recognition, although the evidence base is challenging to compare. The studies differ substantially in participant numbers, recording hardware, windowing, label construction, preprocessing, class balance, and validation design [6]. Random segment-level splits can be especially confusing, as the same windows from the same participant can appear in both the training and testing sets. The model can then leverage stable participant-specific attributes [7] instead of learning a representation that can be applied to unseen participants [8]. Subject-independent evaluation is therefore crucial when the intended application involves new users or wearable deployment [9].

Graph neural networks provide a principled approach to model multichannel physiological signals. A graph representation of channels is a graph with nodes representing channels and selected relationships as edges, instead of an unordered feature vector. Spatial graph operations can be used to aggregate information across related channels, and temporal convolutions or recurrent operations can model changes across successive windows. This structure is appealing for EEG and multisensor data since the physiological channels are not independent and are not naturally organized as conventional image pixels [10]. For a graph model to be useful, however, its nodes must be well defined, its adjacency rule must be auditable, and graph estimation must be leakage-free and controlled through comparisons with simpler architectures [11].

Many other hyperparameters, such as graph sparsity, hidden dimensions, temporal kernel widths, dropout, learning rate, and batch size, are also involved in deep neural networks. These combinations can be explored using population-based optimization [12], but an optimizer does not itself provide predictive capability; it only searches for suitable model configurations. Comparisons should be based on the same evaluation budget for the search procedures and should choose the candidate configurations based on validation data not contained in the test set.

The present study updates the scope of the AKMA-ST-GNN framework to align its claims with the evidence retained from the original experiments. Four public datasets are considered: STEW, SEED, DEAP, and WESAD, but each is analyzed separately due to differences in their modalities and labels. The cognitive workload classification is done with STEW, the emotion classification with SEED, the affect classification with DEAP, and the stress and amusement classification with WESAD. The study does not combine accuracies across incompatible tasks, does not assume EEG–EMG coupling across datasets, and does not use the term “early stress” as no prospective onset horizon or pre-stress interval was specified.

Updated work is thus deliberately restricted. First, it offers a single spatiotemporal graph-learning workflow for a given dataset. Second, it states that AKMA is a hyperparameter search wrapper [13]. Third, it only fills in the dataset-level metrics recorded in the source manuscript and leaves blank values that were not recorded. Fourth, it distinguishes between technical feasibility in the original 80/20 split and subject-independent evidence. Lastly, it provides a path for reproducibility and validation, detailing experiments that must be conducted before claims of generalization, interpretability, and wearable readiness can be substantiated.

2. Related Work

2.1 EEG-based Recognition of Stress, Workload, and Emotion

EEG-based physiological-state recognition has evolved from handcrafted spectral features and traditional classifiers to convolutional, recurrent, and hybrid deep-learning architectures. Afify *et al.* [14] demonstrated discrimination of stress across different cognitive tasks using a convolutional model, highlighting the potential of spatially organized EEG features. Troyee *et al.* [15] investigated EEG signals elicited by visual and auditory stimuli and achieved high binary classification accuracy using various machine learning techniques. Phutela *et al.* [16] highlighted the importance of temporal sequence modeling for stress and non-stress classification using an LSTM-based model. While these studies demonstrate the general feasibility of EEG-based state classification, the results depend on the acquisition devices, cohorts, label definitions, and validation procedures used.

Hybrid models that integrate decomposition, convolution, recurrent units, and optimization are more recent. Gupta *et al.* [17] proposed a modified support vector machine approach for detecting stress levels in EEG. Roy *et al.* [18] integrated signal decomposition with CNN, BiLSTM, and GRU components, whereas Wang *et al.* [19] combined differential-entropy feature matrices with a 2D-CNN–LSTM network for EEG-based

emotion recognition. Andhare *et al.* [20] proposed an optimized hybrid deep learning approach for mental stress detection. These techniques show that the spectral structure and temporal dependencies can be informative. They also illustrate why it is not possible to make direct comparisons of accuracy when the studies are based on different data sets or validation methods.

Frontal EEG has also been investigated as a less complex alternative to full-cap EEG. AlShorman *et al.* [21] used machine learning techniques to analyze frontal-lobe EEG for real-time stress detection. Kaminska *et al.* [22] studied stress recognition in a virtual-reality environment, and Kamakshi and Rengaraj [23] investigated augmented EEG for early detection of stress- and anxiety-associated seizures. This type of work can be used in reduced-channel or wearable systems, but channel reduction changes the information available to the model. It should be evaluated directly rather than extrapolated from full-channel results.

One common problem is the interchangeable use of the terms “stress,” “workload,” and “emotion” as objectives. Experimental labels are different questions; underlying EEG features may overlap. Typically, workload labels are derived from task demand, emotion labels from elicitation paradigms, and stress labels from stressor conditions or self-report. A good manuscript, therefore, should remain faithful to the purpose of each set of data and should not attempt to transform one construct into another without an explicit and validated transformation.

2.2 Peripheral, EMG, and Multimodal Physiological Classification

Peripheral physiological signals provide complementary information on autonomic and muscular responses. Wearable biosensors can capture electrodermal activity, heart-rate related signals, respiration, temperature, movement, and muscle activation in daily or experimental activities. These modalities are relatively portable but are motion-sensitive and sensitive to sensor placement, environmental conditions, and individual physiology. For good evaluation, artifact handling, validation, and analysis of missing/corrupted channels on a participant basis are required.

Intramuscular EMG patterns can assist in classifying neuromuscular conditions, and EMG provides discriminative temporal and spectral information, as demonstrated by Jose *et al.* [24]. Li *et al.* [25] employed a graph-based sequential model to predict movements during stroke rehabilitation. That study is a suitable example of neural–muscular modeling as both EEG and EMG were captured in the same experimental environment. It also demonstrates the need for multimodal acquisition in synchrony for interpreting cross-modal relationships.

Multimodal stress-recognition studies are increasingly transforming the heterogeneous signals into representations suitable for deep learning. Modi *et al.* [26] transformed multimodal physiological signals into representations suitable for stress recognition, whereas Xiang *et al.* [27] integrated time- and frequency-domain features within a multimodal framework. But require synchronized, modality-specific pre-processing, and a fusion strategy.

The present study has a more conservative design. DEAP has synchronized EEG and peripheral physiology in the same data set; WESAD has synchronized peripheral and motion channels (including EMG but not EEG).

2.3 Spatiotemporal and Graph-Based Representations

Multichannel recordings are usually converted into a single feature vector for conventional classifiers. Flattening is simple, but it obscures the anatomical, functional, sensor, or statistical dependence of channels. A graph representation preserves channel identities and enables the model to combine information across edges selectively. In the case of EEG, nodes may correspond to electrodes, and edges may represent anatomical proximity or connectivity from the training set. Peripheral multisensor data can be modalities or sensor channels, but the biological meaning of cross-sensor edges must be explicitly defined.

Spatiotemporal graph neural networks combine spatial message passing with temporal modeling. A spatial layer modifies the features of each node based on its neighbours' information, while a temporal layer records short- or long-term changes in the node's features. The combination can be of interest for physiological data, as state-related information can manifest as both a channel pattern and a temporal trajectory. But a graph model is not necessarily interpretable. While a correlation matrix or a global feature ranking captures associations in the inputs, the explanation of an individual prediction requires node and edge attributions, as well as temporal attributions, extracted from the trained model.

The AKMA-ST-GNN workflow contains channel nodes, within-dataset edges, sequential spatiotemporal blocks, pooling, and a dataset-specific classifier. The high-level architecture is feasible for the four tasks, but the contribution of each component cannot be determined without ablation. A graph model should be compared to a temporal model without a graph, a graph model without a temporal component, and a simpler baseline with the same features and split. In the same way, the optimized system should be compared with the same ST-GNN trained with manually selected or conventionally searched hyperparameters.

2.4 Research Gap and Study Position

Three general conclusions can be drawn from literature. First, there is a technical possibility of classifying physiological states using EEG and peripheral modalities. Second, reported accuracy is insufficient to demonstrate generalization due to differences in validation designs, participant leakage, class imbalance, and preprocessing. Thirdly, multimodal claims are meaningful only if the modalities are synchronized, and the fusion is compared against modality-specific baselines.

A remaining methodological gap is the limited evidence for rigorously validated frameworks that can be applied independently across heterogeneous physiological-state classification tasks. Instead, the study investigates whether a common high-level graph-learning workflow can be used independently on heterogeneous public datasets without altering their native tasks. This research also corrects a reporting issue with the previous version, eliminating pooled accuracy, cross-dataset EEG–EMG fusion claims, prospective early-stress language, and unsupported baseline rankings. The other results are reported as preliminary evidence and need to be confirmed by subject-independent analysis.

3. Materials and Methods

3.1 Study Design and Evidentiary Scope

This work is a secondary analysis of four public physiological datasets using a shared high-level modeling framework. Input preparation and evaluation are done separately for each dataset. The unit of analysis, channel set and output classes are still dataset specific. No mixing of recordings from different datasets is done to form a sample and no pooled performance estimate is made. The underlying preprocessing scripts, model code, split manifest, training logs and optimizer histories were not available for reconstruction. Explicitly preserved procedures are described as implemented, and explicitly not preserved procedures are described as limitations of reproducibility and not provided retroactively. The updated article thus introduces the AKMA-ST-GNN as a technically feasible framework and provides performance values.

3.2 Datasets, Modalities, and Target Definitions

The datasets differ in participant composition, sensor modalities, sampling characteristics, elicitation paradigms, and output definitions. Table 1 summarizes only the information preserved in the source manuscript. Fields that were not retained are not reconstructed.

STEW was used solely for cognitive workload classification. Its low- and high-workload labels were retained and were not relabeled as stress. SEED was used for negative, neutral, and positive emotion classification. Emotion labels were not converted into stress labels, and the analysis is not described as an early-stress detection approach. Participant count, session handling, sampling frequency, and final class counts were not recoverable from the source file.

DEAP was used for affect classification. The information for DEAP records 32 participants, 40 audiovisual trials, 32 EEG channels, and synchronized peripheral physiological signals.

The exact retained peripheral channels and the threshold or discretization rule applied to the native affect ratings were not preserved. The results are therefore reported as dataset-specific affect classes without assigning an unverified thresholding procedure.

Table 1. Dataset-specific participant information, modalities, native targets, and classification scope

Dataset	Verified participant information	Native channels/modalities	Native target	Scope in this study
STEW	48 male participants	14-channel EEG	Low vs. high workload	Binary cognitive-workload classification
SEED	15 participants: 7 male, 8 female	62-channel EEG	Negative, neutral, and positive emotion	Three-class emotion classification
DEAP	32 participants, 40 audiovisual trials	32-channel EEG plus synchronized peripheral physiology	Affect ratings on native scales	Dataset-specific affect classification
WESAD	15 participants	Chest and wrist peripheral/motion signals, including EMG, no EEG	Baseline, stress, and amusement	Three-state stress/affect classification

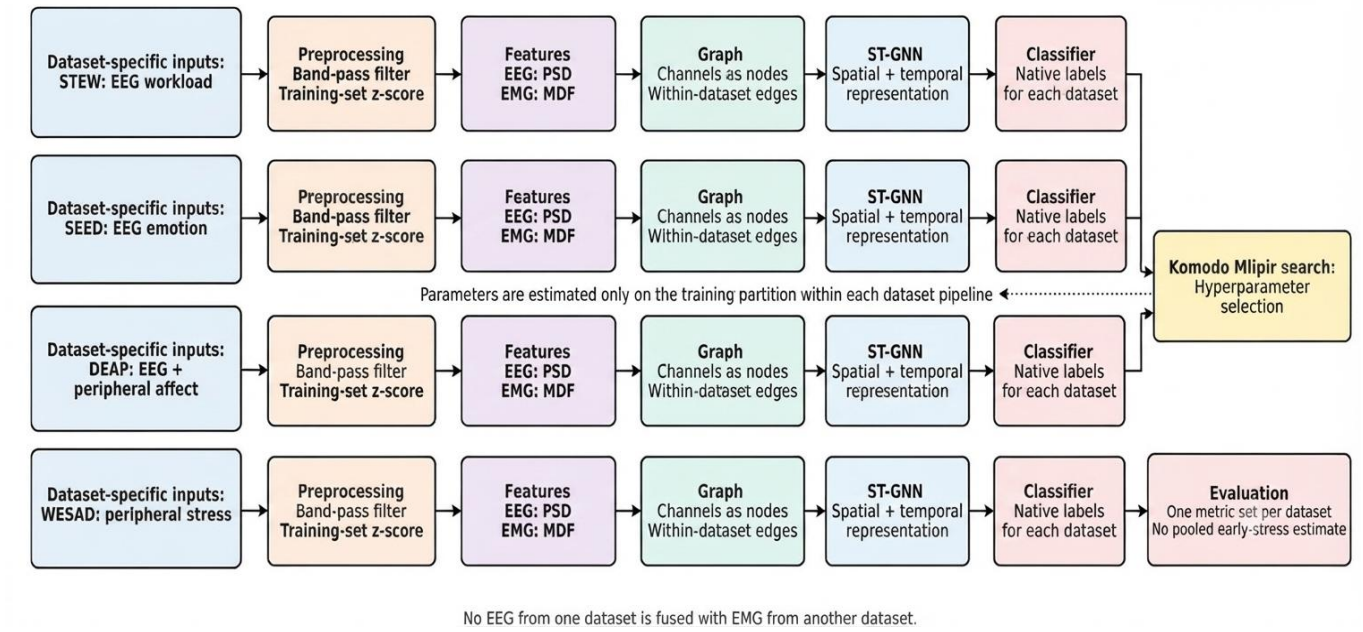

Figure 1. Dataset-specific analysis workflow

WESAD was used for baseline, stress, and amusement classification. It contains synchronized chest and wrist physiological and motion signals, including EMG, but no EEG. Accordingly, WESAD results cannot be interpreted as evidence of EEG–EMG fusion. Modality-specific sampling rates and the exact retained channel set were not preserved.

The public data-access locations recorded in the source were Kaggle mirrors for STEW, SEED, DEAP, and WESAD.

3.3 Dataset-specific preprocessing

The workflow comprises signal filtering, standardization, temporal segmentation, and modality-specific feature extraction. These operations are performed separately for each dataset. Parameters learned from the data, including normalization statistics and any data-driven graph quantities, should be estimated from the training partition and then applied unchanged to validation and test observations. This principle prevents held-out information from influencing preprocessing. The complete dataset-specific analysis workflow, from preprocessing and feature extraction to graph construction, ST-GNN classification, and AKMA-based hyperparameter selection, is illustrated in Figure 1.

Z-score standardization centers each feature and scales it by the standard deviation estimated from the training data. For feature value x, training mean μtrain, and training standard deviation σtrain, the standardized value z is:

$$z = (x - \mu train) / \sigma train \quad (1)$$

The same μtrain and σtrain must be applied to validation and test data. Estimating these values from the complete dataset would introduce leakage, particularly when multiple windows originate from the same participant.

Band-pass filtering is reported in the manuscript, but the filter family, order, cut-off frequencies, and causal or zero-phase implementation were not retained. These settings are physiologically and computationally important because they determine the spectral content available to subsequent feature extraction. A reproducible rerun should report the complete filter specification for each modality and dataset.

For EEG, the source indicates frequency-domain characterization using power spectral density and band-level features. A general relative band-power quantity can be expressed as:

$$RBPb,c = [\Sigma f \in b\ Pc(f)] / [\Sigma f \in F\ Pc(f)] \quad (2)$$

where Pc(f) is the estimated power spectral density of channel c at frequency f, b is the target frequency band, and F is the retained frequency range. The window duration, overlap, taper, frequency resolution, and band boundaries were not preserved and must not be inferred from the figures alone.

Peripheral channels should be processed according to their required attributes. EMG may be represented using amplitude and spectral descriptors, electrodermal activity using phasic characteristics, photoplethysmography using pulse-related measures, and inertial signals using movement descriptors.

3.4 Within-Dataset Graph Representation

For each dataset, retained channels are represented as graph nodes. Edges represent relationships among channels within that dataset. No

cross-dataset nodes or edges are defined. This design preserves the identity of the acquisition system and prevents the model from interpreting unrelated recordings as synchronous observations.

A graph-convolution layer can be written in normalized form as:

$$Hs(l+1) = \varphi(D - 1/2\, A\, D - 1/2\, H(l)\, Ws(l)) \quad (3)$$

where H(l) is the node-feature matrix at layer l, Ã = A + I includes self-connections, D̃ is the degree matrix of Ã, Ws(l) is a trainable weight matrix, and φ is a nonlinear activation. Equation (3) describes the general spatial operation represented in the architecture; it does not reconstruct the unavailable adjacency estimator, sparsification rule, graph-convolution order, or hidden dimensions.

The graph must be estimated without access to held-out participants. If statistical dependence is used to define edges, the dependence matrix and any threshold or top-k selection should be calculated on the training partition within each fold. An adjacency matrix derived from all participants before splitting would leak test information even when the classifier itself is trained only on the training subset.

3.5 Spatiotemporal Graph Neural Network

The model receives a tensor whose dimensions correspond conceptually to time windows, graph nodes, and node features. Figure 2 presents the high-level AKMA-ST-GNN architecture, including three sequential spatiotemporal blocks. Each block combines a spatial graph operation with a gated one-dimensional temporal convolution, followed by normalization, activation, and dropout. The output is pooled and passed to a classifier whose dimensionality matches the native label space of the current dataset.

A generic gated temporal operation can be expressed as:

$$Ht = tanh(Convf(Hs)) \odot sigmoid(Convg(Hs)) \quad (4)$$

where Convf and Convg are temporal convolutions and ⊙ denotes element-wise multiplication. The pooled representation is converted to class probabilities using a dataset-specific output layer:

$$\hat{y} = softmax(Wo\, Pool(Ht) + bo) \quad (5)$$

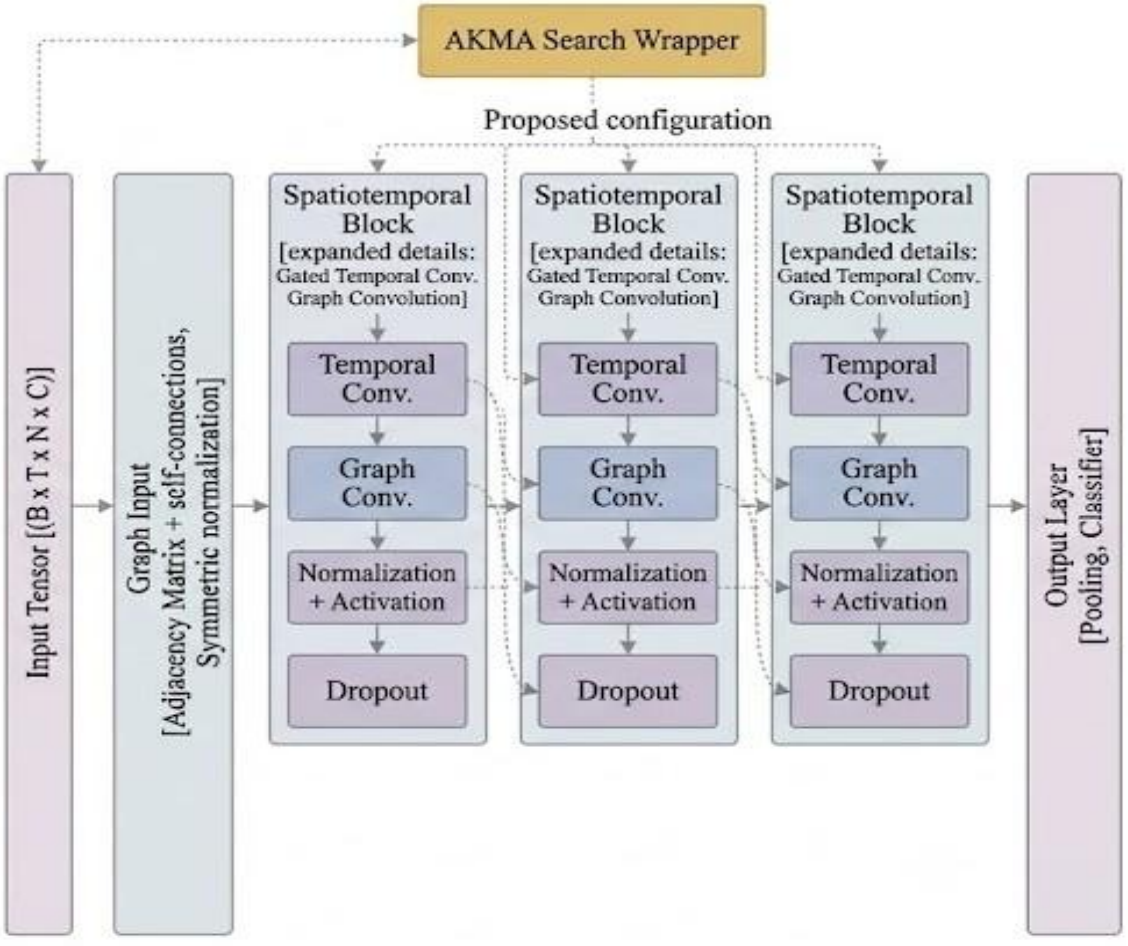

Figure 2. High-level AKMA-ST-GNN architecture.

The exact temporal kernel widths, dilation, padding, hidden dimensions, activation functions, normalization scheme, dropout, pooling operation, and classifier specification were not preserved. Equations (3)–(5) therefore communicate the high-level computational structure and should not be interpreted as a substitute for the original implementation.

3.6 Adaptive Komodo Mlipir Hyperparameter Search

AKMA is used as a population-based wrapper that proposes candidate ST-GNN hyperparameter configurations. Each candidate is trained and evaluated using a training-fold objective, after which the search updates its population and retains the best validation configuration. The selected configuration is then used for final evaluation on the held-out portion of the dataset.

The optimizer is not an independent predictive layer. Its role is limited to selecting hyperparameters such as architecture size or training settings from a predefined search space.

A common analysis for all search methods is required to make a fair assessment of AKMA. Random search, Bayesian optimization, particle swarm optimization, Komodo Mlipir Algorithm without modification and AKMA should be compared for the same number of candidate configurations, same data splittings and same objective. The performance of the classifiers reported without such a comparison cannot be solely credited to the adaptive optimizer.

The workflow can be summarized as follows:

1 Select one dataset and retain its native channels, participants, and labels.

2 Create training, validation, and test partitions without mixing records across datasets.

3 Fit normalization and any data-driven preprocessing on the training partition only.

4 Construct a within-dataset channel graph using training data only.

5 Generate candidate ST-GNN configurations within the predefined AKMA search space.

6 Train each candidate on the training subset and calculate the fitness on validation data.

7 Update the AKMA population until the search budget is exhausted.

8 Select the best validation configuration and evaluate it once on the held-out test subset.

9 Report dataset-native metrics without pooling targets or modalities across datasets.

3.7 Experimental Environment and Validation Protocol

The experiments were conducted in Python 3.10 using a Jupyter Notebook environment on Windows 11 with an Intel i7 processor, 16 GB RAM, and an NVIDIA RTX 3060 GPU. Training histories were reported for 50 epochs. The original study utilized an 80% training and 20% test split, but it does not preserve the split manifest.

Because the split unit cannot be verified, the current results are designated preliminary. They should not be described as subject independent. A confirmatory analysis should use grouped cross-validation or leave-one-subject-out validation. Hyperparameter selection, early stopping, normalization, feature scaling, graph estimation, and any feature selection must occur inside the training process for each fold. The final test participant or fold must remain inaccessible during model selection.

The package does not include repeated-seed outputs, fold-level scores, participant-level predictions, or a split manifest. Accordingly, the reported accuracies are single-point estimates from one 80/20 evaluation. A confirmatory analysis should report the mean, standard deviation, and 95% confidence interval across prespecified repeated seeds and subject-grouped folds. Confidence intervals should be derived from participant- or grouped-fold distributions, rather than from individual windows, because windows from the same participant are not statistically independent.

3.8 Evaluation Metrics

The performance measures include accuracy, precision, sensitivity, specificity, and F1-score. For a binary or one-versus-rest class calculation, the standard definitions are:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (6)$$

$$Precision = TP / (TP + FP) \quad (7)$$

$$Sensitivity = TP / (TP + FN) \quad (8)$$

$$Specificity = TN / (TN + FP) \quad (9)$$

$$F1 = 2 \times (Precision \times Sensitivity) / (Precision + Sensitivity) \quad (10)$$

For multiclass datasets, accuracy should be accompanied by class-wise results and aggregation rules that remain informative under class imbalance. Let C denote the number of classes, nc the number of observations in class c, N the total number of observations, and F1c and Sensitivityc the class-specific F1-score and sensitivity, respectively.

$$Macro\ F1 = (1/C) \times \Sigma c = 1C\ F1c \quad (11)$$

$$Weighted\ F1 = \Sigma c = 1C\ (nc/N) \times F1c \quad (12)$$

$$Balanced\ accuracy = (1/C) \times \Sigma c = 1C\ Sensitivityc \quad (13)$$

The multiclass Matthews correlation coefficient (MCC) can be calculated from the confusion matrix as:

$$MCC = [cs - \Sigma k\, pktk] / \sqrt{[(s^2 - \Sigma k\, pk^2)(s^2 - \Sigma k\, tk^2)]} \quad (14)$$

where c is the sum of the diagonal entries of the confusion matrix, s is the total number of observations, pk is the number of predictions assigned to class k, and tk is the number of true observations in class k.

Cohen's kappa adjusts observed agreement for the agreement expected from the marginal class distributions:

$$\kappa = (po - pe) / (1 - pe) \quad (15)$$

where po is the observed agreement and pe is the expected agreement under the marginal distributions. Confusion matrices should be reported as raw counts and, where useful, normalized by the true class. For multiclass ROC analysis, one-versus-rest class curves and class-specific AUC values should be reported, together with macro- and weighted-average AUC. ROC/AUC requires class labels and continuous decision scores or predicted probabilities; it cannot be recovered from accuracy alone.

3.9 Statistical Uncertainty and Comparative Testing

The experimental package contains one reported performance value per dataset and does not preserve repeated runs, fold-level scores, participant-level predictions, or paired baseline outputs. Standard deviations, 95% confidence intervals, and statistical significance tests therefore cannot be calculated without rerunning the original pipeline.

In a confirmatory analysis, all models should be evaluated on identical subject-grouped folds and repeated under prespecified random seeds. Results should be summarized as mean ± standard deviation with 95% confidence intervals derived from participant- or grouped-fold distributions. Comparisons with baseline or ablation models should use paired tests on the same folds or participants, accompanied by an effect-size estimate and an appropriate correction when multiple models or datasets are tested. Statistical comparisons should be made within a dataset; the four datasets should not be treated as repeated observations of one common classification problem.

4. Results

4.1 Evidence Retained from the Experiments

The source preserves dataset-specific performance values and training histories under the original 80/20 protocol. It does not preserve subject-level fold assignments, prediction files, confusion matrices, decision scores, optimizer trajectories, repeated-seed results, or controlled baseline experiments. The numerical findings are therefore reported exactly as available and are not supplemented with estimated values.

The previous pooled accuracy of 98.51% was removed because averaging results across four tasks with different modalities, labels, and class structures lacks clear inferential meaning. A pooled score implies that all datasets measure the same construct, which they do not.

4.2 Dataset-Specific Classification Performance

Table 2 reports the metrics preserved in the manuscript. “Not reported” indicates that the source did not provide the measure and that it was not reconstructed. For STEW, the retained F1-score did not specify whether a binary, macro, micro, or weighted averaging rule was used; it is therefore reported only as the F1 value.

STEW produced the highest reported accuracy (98.60%), precision (98.37%), F1-score (98.28%), specificity (98.01%), and sensitivity (98.23%). These values indicate internally consistent performance under the split. They do not establish performance on unseen participants because the split unit is unknown.

WESAD reported an accuracy of 98.31%. Other metrics were not preserved. Because WESAD contains baseline, stress, and amusement states, accuracy alone does not show whether the model performs equally across all three classes. Class imbalance or confusion between baseline and amusement could remain hidden without class-wise recall and a confusion matrix.

DEAP reported accuracy of 98.06%, specificity of 98.32%, and sensitivity of 98.12%. Precision and F1-score were not retained. The exact affect-label transformation is also unavailable, so the result is described only as dataset-specific affect classification. It should not be compared directly with studies that use different valence or arousal thresholds.

SEED produced a reported accuracy of 98.12%, specificity of 98.05%, and sensitivity of 98.03% for negative, neutral, and positive emotion classification. Precision and F1-score were not retained. These values describe the emotion task and are not interpreted as evidence of stress detection. The four reported dataset-level accuracies are presented visually in Figure 3.

4.3. Training Histories

Figure 4 presents the training and validation accuracy and loss trajectories for STEW, WESAD, DEAP, and SEED over 50 epochs. Across all four datasets, training and validation accuracy increased, while the corresponding loss values decreased. The curves suggest numerical convergence under the

original data partition. The closeness of the training and validation trajectories is visually reassuring. However, it cannot rule out participant leakage, as correlated windows from the same participant may have appeared in both subsets, which means that correlated windows from the same participant may result in similar behavior in both subsets.

Table 2. Preliminary dataset-specific performance reported for AKMA-ST-GNN under the 80/20 protocol.

Dataset	Precision (%)	F1-score (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Native target
STEW	98.37	98.28	98.01	98.60	98.23	Low/high workload
WESAD	Not reported	Not reported	Not reported	98.31	Not reported	Baseline/stress/amusement
DEAP	Not reported	Not reported	98.32	98.06	98.12	Dataset-specific affect classes
SEED	Not reported	Not reported	98.05	98.12	98.03	Negative/neutral/positive emotion

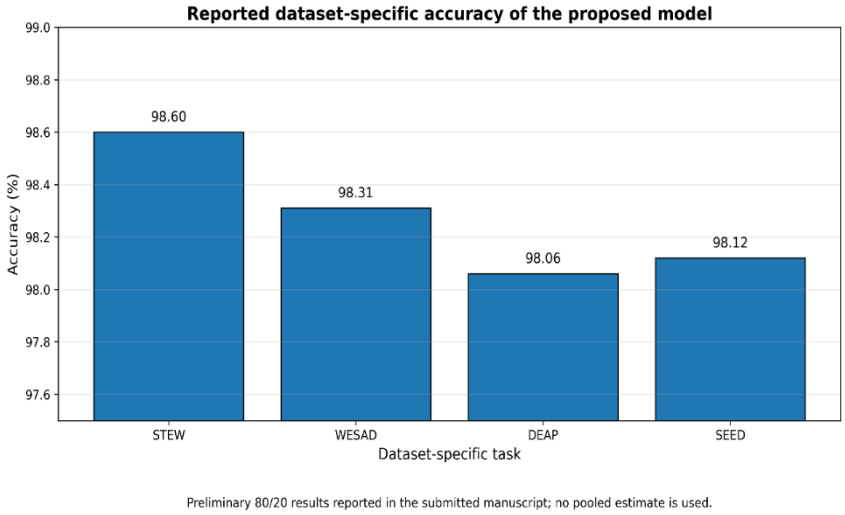

Figure 3. Reported dataset-specific accuracies under the preliminary 80/20 evaluation. The bars are not a pooled estimate and do not constitute a controlled comparison across datasets.

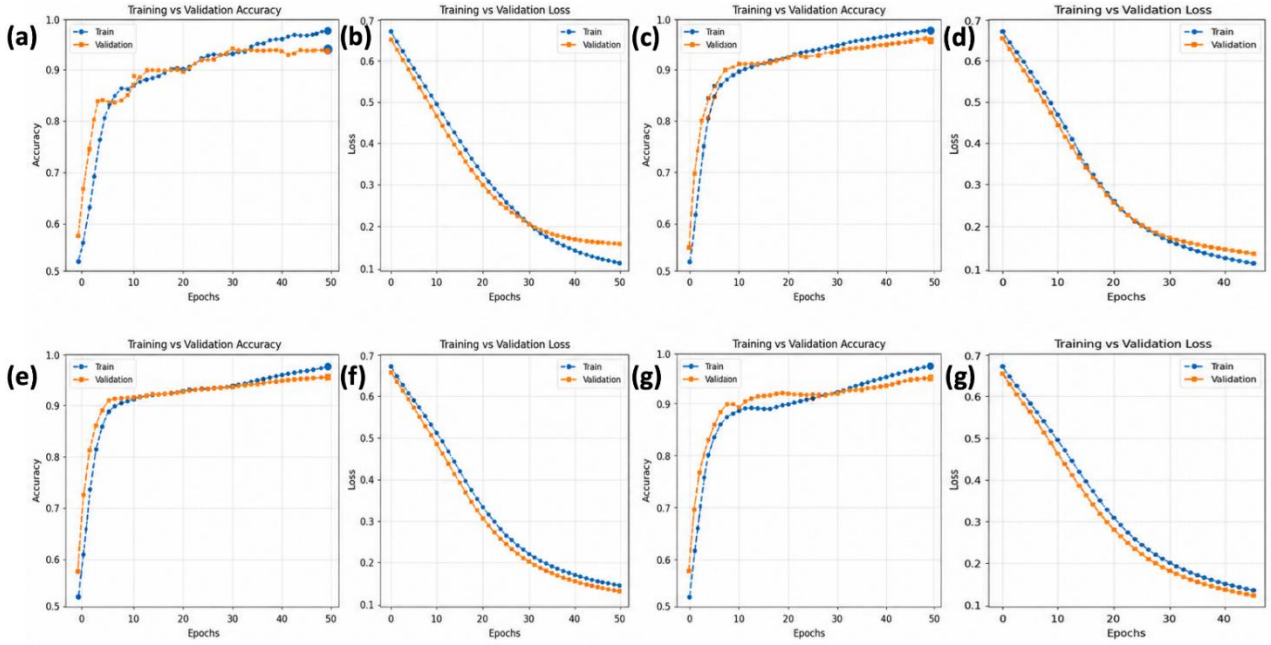

Figure 4. Training and validation accuracy and loss histories for (a-b) STEW, (c-d) WESAD, (e-f) DEAP, and (g-h) SEED over 50 epochs

Table 3. Evaluation outputs in the experimental package.

Requested measure/output	Status in the package	Data required for valid calculation
Macro F1	Not computable	Class-wise F1-scores or true and predicted labels
Weighted F1	Not computable	Class-wise F1-scores and class supports
Balanced accuracy	Not computable	Class-wise sensitivity/recall values
Multiclass MCC	Not computable	Complete confusion matrix or true and predicted labels
Cohen's kappa	Not computable	Observed and expected agreement from the confusion matrix
Confusion matrices	Not available	True and predicted class labels
ROC/AUC	Not computable	True labels and class-wise decision scores or probabilities
Repeated-run mean and SD	Not available	Scores from repeated seeds or grouped folds
95% confidence intervals	Not computable	Participant-, grouped-fold-, or repeated-run score distributions
Statistical significance tests	Not performed	Paired results from the same folds/participants for the proposed and comparison models

The source package does not specify the checkpoint-selection rule, early stopping criterion, best epoch, or if the plotted validation subset was used for final reporting. These curves are therefore considered to be descriptive training histories, rather than unbiased generalization.

4.4. Availability of Multiclass and Uncertainty Measures

Certain extended multiclass and uncertainty measures could not be reconstructed. The calculation requires true labels and model predictions, class supports, continuous decision scores or probabilities, and repeated fold- or seed-level outputs. None of these records was available in the source package. Table 3 documents the status of each requested output so that missing evidence is distinguished from a negative result.

This restriction is directly applicable to the multiclass WESAD, DEAP and SEED tasks and also restricts a full uncertainty analysis for the binary STEW task. These outputs are not present and thus the reported accuracies are not suitable to assess class-balanced performance, calibration, run-to-run stability or statistical superiority.

5. Discussion

5.1 Principal Findings

The new analysis results in a small but important finding: a common high-level spatiotemporal graph-learning workflow can be used for EEG, peripheral, and synchronized multimodal datasets, individually, without changing the native task of each dataset. The 80/20 results were consistently high with a range of 98.06% to 98.60%. The results give a preliminary technical proof-of-concept of combining channel-aware spatial modelling and temporal processing with hyperparameter search on heterogeneous physiological data.

The principal methodological contribution of the revised analysis is the explicit separation of datasets, modalities, and target constructs. The four datasets are no longer offered as a single EEG–EMG corpus. STEW and SEED are EEG tasks, DEAP has synchronized EEG and peripheral physiology, and WESAD has peripheral and motion signals without EEG. This separation ensures that the model is not given credit for neural–muscular fusion when there is no synchronized neural-muscular signal.

The updated interpretation also distinguishes between workload, affect, emotion, and stress. This is important because a high score on one construct does not validate another. STEW performance is related to low and high workload. SEED performance is related to three classes of emotions. DEAP performance is related to the affect-label transformation for the implementation, but this rule also needs to be recovered. WESAD performance is related to baseline, stress, and amusement.

5.2 Rationale for Spatiotemporal Graph Representation

Physiological channels are not independent of each other. EEG electrodes can be coupled due to volume conduction, anatomical proximity, or functional coordination. Autonomic and motor responses co-occur, leading to peripheral signals covarying. A graph model can preserve the channel's identity and enable selective information exchange, whereas a flattened vector treats all features as equally unrelated inputs. One such structural prior is why GNNs are appealing for multichannel biosignals.

Temporal modeling is also crucial. Stress, workload, and affective responses are not a single value, but change over time. Gated Temporal Convolutions can model local changes and regulate information flow. Thus, spatial and temporal operations are better suited

to the general organization of the data than a purely static classifier.

Plausibility is not evidence of benefit, however. The results do not include a temporal-only baseline, a graph-only baseline, or a model using shuffled or identity adjacency. If these controls are not included, the observed performance cannot be attributed to the graph structure. A high-capacity temporal network with the same input features could perform equally well, particularly if there is leakage in the split in terms of participants.

5.3 Dataset-Specific Interpretation of Results

STEW was the most accurate reported and the only dataset to retain all five of the metrics listed above. This apparently good result might be partly because binary classification is less complex than the distinction of three classes or the discretization of continuous affect ratings. This interpretation is still preliminary as class counts, participant allocation, and windowing were not saved.

WESAD also achieved a high reported accuracy, but accuracy alone is insufficient for a three-state task. Class duration and physiological variability may differ among baseline, stress, and amusement, allowing a model to obtain a high overall score while performing poorly on a minority class. A confirmatory evaluation should therefore report class-wise sensitivity, macro F1, weighted F1, balanced accuracy, multiclass MCC, Cohen's kappa, a confusion matrix, and one-versus-rest ROC/AUC where valid decision scores are available.

Label construction is particularly important for the DEAP result. DEAP is continuous, and studies often convert these ratings into classes based on thresholds or exclude trials in the ambiguous mid-range. The sample size and the task difficulty vary under different rules. The threshold cannot be directly compared with published DEAP studies, as it is not available. It is therefore a priority to recover the label-generation code.

SEED performance is related to negative, neutral, and positive emotions. The reported specificity and sensitivity are close to the accuracy, indicating broadly similar aggregate behaviour, but the averaging rule is unknown. Session effects and individual variability also influence emotion recognition at the subject level. A grouped protocol is required to assess the generalization of the representation beyond the participants in the training set.

5.4 Comparison with Previous Physiological-State Classifiers

The accuracies are numerically superior to many of the accuracies summarized in the related literature, such as EEG stress classifiers, multimodal stress models, and graph-based rehabilitation systems. This numerical observation should not be interpreted as superiority. Published studies use different participants, sensors, class definitions, augmentation procedures, and validation units. A model evaluated with random windows can appear substantially stronger than the same model evaluated on unseen participants.

A valid benchmark must rerun competing methods on the same preprocessed data and subject partitions. At minimum, the proposed framework should be compared with a conventional classifier using the same features, a CNN, an LSTM or BiLSTM, a graph convolutional network without temporal blocks, and an ST-GNN without AKMA. The optimizer comparison should use an equal number of candidate evaluations.

5.5 Role and Limits of AKMA

Hyperparameter search can be used to find combinations that might not have been considered when manually tuning the model, which could improve the model. Population-based procedures can be helpful if the search space contains both discrete and continuous variables, and if the interactions between the parameters are strong. The AKMA is a reasonable search strategy for the ST-GNN.

The performance of the final model is not a measure of the quality of the optimizer, since the same architecture might have been able to find an equivalent solution by random search or by conventional tuning.

In the future, search performance and predictive performance should be distinguished. Search performance includes the best validation score as a function of evaluation budget, variance across optimizer runs, convergence speed, and computational cost. Predictive performance includes the final held-out metrics of the selected model. Reporting both would allow readers to determine whether AKMA contributes a practical advantage.

5.6 Interpretability and Physiological Validity

Physiological interpretation is not inherent in graph architecture. Input correlations show relationships between features, and a global feature ranking shows which features were deemed important by a separate importance procedure. Neither explains how the ST-GNN arrived at a particular decision. For each prediction, influential nodes, edges, frequencies, and time intervals should be identified through model-specific interpretation.

Gradient-based node and temporal attribution, integrated gradients, edge masking, attention analysis (if attention is part of the model), and perturbation tests (remove channels or time segments) are useful. The explanations should be evaluated for consistency between folds and participants.

There is a need for extra caution when interpreting across modalities. In DEAP, synchronized EEG and peripheral channels could be analyzed together if the channels are retained and the fusion mechanism is described. In WESAD, the interaction between peripheral and motion signals can be investigated. EEG–EMG coupling is not possible with the current datasets, as they do not have the same participants or trials.

5.7 Generalization, Leakage and Uncertainty

The biggest validity problem is the lack of a subject-level split manifest. Physiological windows of the same person are highly correlated due to stable anatomy, sensor contact, baseline amplitude, and recording conditions. Those windows are not independent as required for deployment to a new user if they are divided randomly. The model can identify participants' signatures rather than the intended state.

Grouped cross-validation is thus not an optional improvement, but rather a change of the scientific question. A random-window split is a question that asks whether the model can classify other windows from participants it has already seen. A subject-independent split is a question that determines whether a new participant can be classified. The latter is required for claims of generalization.

A single accuracy value provides no information about run-to-run or participant-to-participant stability, and the four dataset values cannot serve as replicates because they correspond to different tasks. The absence of repeated seeds, grouped-fold scores, and participant-level predictions prevents calculation of standard deviations, 95% confidence intervals, and paired significance tests. These analyses must be generated from a leakage-controlled rerun. Calibration curves and expected calibration error should also be reported when predicted probabilities are retained.

5.8 Failure Conditions

Motion artifacts, electrode movement, sweat, varying contact impedance, sensor dropout, hardware variations, environmental temperature, and participant behaviour may affect real-world performance. Classifiers trained in a controlled laboratory environment may not perform well when the state distribution or acquisition context changes. Domain shift is especially important if public datasets are gathered using different devices and protocols.

The number of parameters, the size of the model when serialized, the inference latency, the peak memory usage, the energy consumption, and the throughput should be measured on the target hardware. Masking/corruption of channels should be performed to mimic real sensor failure in robustness experiments.

False-alarm rates and class-transition behaviour should also be assessed. Even if the window level accuracy is high, a physiological monitor with rapid state transitions may be unusable. For the intended use, the temporal smoothing, decision latency and clinical/operational cost of false positives and false negatives should be defined.

5.9 Reproducibility Roadmap

Before making any larger claims, a confirmatory study should follow this sequence:

- Normalize, extract features, estimate graphs, and select hyperparameters within each training fold.
- Report class counts, window counts, class-wise metrics, confusion matrices, repeated-seed variability, and confidence intervals.
- Compare the proposed model against conventional, temporal, graph, and spatiotemporal baselines using the same inputs and folds.
- Conduct ablations for graph edges, temporal blocks, frequency features, modality groups, and AKMA optimization.
- Compare the performance of AKMA and random search, Bayesian optimization, particle swarm optimization, and unmodified KMA under the same budget.
- Make code, versions of dependencies, random seeds, trained configurations, and scripts available that reproduce all tables and figures.
- Assess model attribution, reduced channel performance, missing-sensor robustness, and embedded-device resource requirements.

5.10 Strengths and limitations

A strength of the proposed research is its explicit separation of datasets and constructs. It prevents scientifically misleading pooling and focuses the interpretation on what the evidence can support.

The restrictions are significant. The source package does not contain the original code, split manifest, full model configuration, exact preprocessing parameters, DEAP label-generation rule, class counts, prediction-level outputs, continuous class scores, repeated-seed results, or controlled baselines. Consequently, macro and weighted F1, balanced accuracy, MCC, Cohen's kappa, confusion matrices, ROC/AUC, standard deviations, confidence intervals, and statistical significance tests cannot be reconstructed. The 80/20 split also cannot be verified as participant independent. The figures do not replace these missing records.

6. Conclusion

In this study, a dataset-specific AKMA-ST-GNN framework is proposed for the classification of cognitive workload, emotion, affect, and stress using STEW, SEED, DEAP, and WESAD. The new scope keeps the original participants, modalities, acquisition conditions, and labels for each dataset. It does not combine EEG from one source with EMG or peripheral signals from another, merge incompatible targets, or claim prospective early-stress prediction. The main dataset characteristics, such as the number of participants, sensor configuration, sampling rates, and recording protocols, were obtained from public documentation.

The manuscript reported accuracies of 98.60% for STEW, 98.31% for WESAD, 98.06% for DEAP, and 98.12% for SEED under an 80%/20% split. These are single estimates from a single assessment. Prediction-level outputs, class supports, decision scores, repeated seeds, fold-level results, controlled baseline runs were not saved and macro or weighted F1, balanced accuracy, MCC, Cohen's kappa, confusion matrices, ROC/AUC, standard deviations, 95% confidence intervals and statistical significance tests cannot be reported. The values are thus only indicative of the technical feasibility and do not indicate class-balanced performance, stability, superiority or subject-independent generalization.

A confirmatory study should be performed with the same partitions for all comparison models, repeated seeds, training-fold-only preprocessing and graph estimation, and the whole pipeline repeated with subject-grouped cross-validation. It should contain class-wise metrics, macro and weighted F1, balanced accuracy, MCC, Cohen's kappa, confusion matrices, and ROC/AUC (where applicable), mean ± standard deviation, 95% confidence intervals, effect sizes, and paired statistical comparisons. The reported accuracies should be used with caution until these analyses are completed.

References

[1] M. Bonanno, D. Papa, A. Cerasa, M.G. Maggio, R.S. Calabrò, Psycho-neuroendocrinology in the rehabilitation field: Focus on the complex interplay between stress and pain. Medicina, 60(2), (2024) 285. https://doi.org/10.3390/medicina60020285

[2] Y. Liu, M.I. Palacio, T. Bikki, C. Toledo, Y. Ouyang, Z. Li, Z. Wang, F. Toledo, H. Zeng, M.T. Herrero Machine learning, physiological signals, and emotional stress/anxiety: Pitfalls and challenges. Applied Sciences, 15(21), (2025) 11777. https://doi.org/10.3390/app152111777

[3] C. Su, M. Zangeneh Soroush, N. Torkamanrahmani, A. Ruiz-Segura, L. Yang, X. Li, Y. Zeng Measurement and quantification of stress in the decision process: A model-based systematic review. Intelligent Computing, 3, (2024) 0090. https://doi.org/10.34133/icomputing.0090

[4] R. Pillalamarri, U. Shanmugam, A review on EEG-based multimodal learning for emotion recognition. Artificial Intelligence Review, 58(5), (2025) 131. https://doi.org/10.1007/s10462-025-11126-9

[5] C. Chatzaki, M. Tsiknakis, An overview of stress analysis based on physiological signals: Systematic review of open datasets and current trends. Sensors, 25(23), (2025) 7108. https://doi.org/10.3390/s25237108

[6] K. Jin, A. Rubio-Solis, R. Naik, D. Leff, J. Kinross, G. Mylonas, Human-centric cognitive state recognition using physiological signals: A systematic review of machine learning strategies across application domains. Sensors, 25(13), (2025) 4207. https://doi.org/10.3390/s25134207

[7] J.Y. Wu, C.T.S. Ching, H.M.D. Wang, L.D. Liao, Emerging wearable biosensor technologies for stress monitoring and their real-world applications. Biosensors, 12(12), (2022). 1097. https://doi.org/10.3390/bios12121097

[8] J. Su, J. Zhu, T. Song, H. Chang, Subject-independent EEG emotion recognition based on genetically optimized projection dictionary pair learning. Brain Sciences, 13(7), (2023) 977. https://doi.org/10.3390/brainsci13070977

[9] G. Vos, K. Trinh, Z. Sarnyai, M. Rahimi Azghadi, Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review. International Journal of Medical Informatics, 173, (2023) 105026. https://doi.org/10.1016/j.ijmedinf.2023.105026

[10] D. Klepl, M. Wu, F. He, Graph neural network-based EEG classification: A survey. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 32, (2024) 493–503. https://doi.org/10.1109/TNSRE.2024.3355750

[11] J. Zhu, X. Jin, Y. Ming, W. Kong, EEG-DGRN: Dynamic graph representation network for subject-independent ERP detection. Brain-Apparatus Communication: A Journal of Bacomics, 4(1), (2025) 2447576. https://doi.org/10.1080/27706710.2024.2447576

[12] A. Morales-Hernández, I. Van Nieuwenhuyse, S. Rojas Gonzalez (2023). A survey on multi-objective hyperparameter optimization algorithms for machine learning. Artificial Intelligence Review, 56(8), 8043–8093. https://doi.org/10.1007/s10462-022-10359-2

[13] S. Suyanto, A.A. Ariyanto, A.F. Ariyanto Komodo Mlipir Algorithm. Applied Soft Computing, 114, (2022) 108043. https://doi.org/10.1016/j.asoc.2021.108043

[14] H.M. Afify, K.K. Mohammed, A.E. Hassanien,

Stress detection based EEG under varying cognitive tasks using convolution neural network. Neural Computing and Applications, 37(7), (2025) 5381–5395. https://doi.org/10.1007/s00521-024-10737-7

[15] T.G. Troyee, M.H. Chowdhury, M.F.K. Khondakar, M. Hasan, M.A. Hossain, Q.D. Hossain, M.A.A. Dewan, Stress detection and audio-visual stimuli classification from electroencephalogram. IEEE Access, 12, (2024) 145417–145427. https://doi.org/10.1109/ACCESS.2024.3471590

[16] N. Phutela, D. Relan, G. Gabrani, P. Kumaraguru, M. Samuel, Stress classification using brain signals based on LSTM network. Computational Intelligence and Neuroscience, 2022, (2022) 7607592. https://doi.org/10.1155/2022/7607592

[17] R. Gupta, M.A. Alam, P. Agarwal, Modified support vector machine for detecting stress level using EEG signals. Computational Intelligence and Neuroscience, 2020, (2020) 8860841. https://doi.org/10.1155/2020/8860841

[18] B. Roy, L. Malviya, R. Kumar, S. Mal, A. Kumar, T. Bhowmik, J.W. Hu, Hybrid deep learning approach for stress detection using decomposed EEG signals. Diagnostics, 13(11), (2023) 1936. https://doi.org/10.3390/diagnostics13111936

[19] T. Wang, X. Huang, Z. Xiao, W. Cai, Y. Tai, EEG emotion recognition based on differential entropy feature matrix through 2D-CNN-LSTM network. EURASIP Journal on Advances in Signal Processing, 2024(1), (2024) 49. https://doi.org/10.1186/s13634-024-01146-y

[20] M.S. Andhare, T. Vijayan, B. Karthik, S. Urooj, A novel optimized hybrid deep learning framework for mental stress detection using electroencephalography. Brain Sciences, 15(8), (2025) 835. https://doi.org/10.3390/brainsci15080835

[21] O. AlShorman, M. Masadeh, M.B.B. Heyat, F. Akhtar, H. Almahasneh, G.M. Ashraf, A. Alexiou, Frontal lobe real-time EEG analysis using machine learning techniques for mental stress detection. Journal of Integrative Neuroscience, 21(1), (2022) 20. https://doi.org/10.31083/j.jin2101020

[22] D. Kamińska, K. Smółka, G. Zwoliński, Detection of mental stress through EEG signal in virtual reality environment. Electronics, 10(22), (2021) 2840. https://doi.org/10.3390/electronics10222840

[23] K. Kamakshi, A. Rengaraj, Early detection of stress and anxiety based seizures in position data augmented EEG signal using hybrid deep learning algorithms. IEEE Access, 12, (2024) 35351–35365. https://doi.org/10.1109/ACCESS.2024.3365192

[24] S. Jose, S.T. George, M.S.P. Subathra, V.S. Handiru, P.K. Jeevanandam, U. Amato, E.S. Suviseshamuthu Robust classification of intramuscular EMG signals to aid the diagnosis of neuromuscular disorders. IEEE Open Journal of Engineering in Medicine and Biology, 1, (2020) 235–242. https://doi.org/10.1109/OJEMB.2020.3017130

[25] H. Li, H. Ji, J. Yu, J. Li, L. Jin, L. Liu, Z. Bai, C. Ye, A sequential learning model with GNN for EEG–EMG-based stroke rehabilitation BCI. Frontiers in Neuroscience, 17, (2023) 1125230. https://doi.org/10.3389/fnins.2023.1125230

[26] N. Modi, Y. Kumar, K. Mehta, N. Chaplot, Physiological signal-based mental stress detection using hybrid deep learning models. Discover Artificial Intelligence, 5(1), (2025) 166. https://doi.org/10.1007/s44163-025-00412-8

[27] J.Z. Xiang, Q.Y. Wang, Z.B. Fang, J.A. Esquivel, Z.X. Su, A multi-modal deep learning approach for stress detection using physiological signals: Integrating time- and frequency-domain features. Frontiers in Physiology, 16, (2025) 1584299. https://doi.org/10.3389/fphys.2025.1584299

Acknowledgments

The authors acknowledge the providers of the public datasets used in the study and the institutional support described in the funding statement.

Has this article screened for similarity?

Yes

Ethics statement

This study reports secondary analysis of publicly available physiological datasets and did not recruit new participants or collect new biological samples. Ethical approval and informed-consent procedures are governed by the original dataset providers.

Authors Contribution Statement

R. Kishore Kanna: Conceptualization, Methodology, Investigation, Formal analysis, Writing - Original Draft, Visualization, Project administration. Biswajit Brahma: Writing - Review & Editing. N. Aravindha Babu: Supervision. Ramesh Kumar Ayyasamy: Methodology. Bharath Kumar Nagaraj: Formal analysis. Ayodeji Olalekan Salau: Writing - Review & Editing. All authors reviewed and approved the final version of the manuscript.

Code availability

The code used for the reported experiments was not available in the source package reviewed for this revision. No replacement code or reconstructed experiment was used to generate the reported numerical results.

Funding

This work was supported by Vel Tech Research Park, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, under the Research Development Fund Policy, Ref. No. VTU/RDF/FY 2026-27/009.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The datasets were accessed from public mirrors identified in the manuscript: STEW (https://www.kaggle.com/datasets/mitulahirwal/mental-cognitive-workload-eeg-data-stew-dataset), SEED (https://www.kaggle.com/datasets/daviderusso7/seed-dataset), DEAP (https://www.kaggle.com/datasets/harshilgupta28/deap-dataset), and WESAD (https://www.kaggle.com/datasets/orvile/wesad-wearable-stress-affect-detection-dataset).

Declaration of generative AI use

OpenAI ChatGPT (GPT-5.6 Sol) was used for language refinement, grammatical correction, and improvement of manuscript organization. The tool was not used to generate or modify raw physiological data, participant records, class labels, experimental outcomes, or missing performance metrics. All scientific statements, numerical values, figures, citations, and references were independently reviewed and verified by the authors, who take full responsibility for the final content.