Fine-Grained Learning Style Detection Using Multimodal Deep Learning and Log Data

Jayvie Ochona Guballo[1,*], Reynante G. Rosal[2]

[1]Rizal Technology University, Philippines

[2]Instructor III School: Pangasinan State University- San Carlos City Campus, Philippines

ABSTRACT

This study proposes a multimodal deep-learning framework for fine-grained detection of learning styles by integrating behavioral log data, textual reflections, and visual interaction signals from learners in online learning environments. Using a dataset containing 200 learner profiles and more than 50,000 interaction events, the model combines Bi-LSTM–based sequence encoding, BERT semantic extraction, and CNN-driven visual behavior processing within a cross-modal attention architecture. Descriptive statistics show substantial variation across behavioral and cognitive indicators, including mean time-on-task of 47.82 seconds (SD 21.45), quiz attempts averaging 2.31 per item, reflection lengths ranging from 12 to 402 tokens, and cursor travel distances spanning 92 to 1,560 pixels. Results demonstrate that the full multimodal model achieves an overall accuracy of 0.84 and an F1-score of 0.83 across FSLSM axes, outperforming all unimodal baselines. Ablation studies reveal that removing log data reduces accuracy from 0.84 to 0.76, while removing text or visual data lowers performance to 0.79 and 0.81 respectively, confirming the dominance of sequential behavior as a predictive signal. The model's fine-grained outputs produce mean learning-style scores near the midpoint of each axis (0.51–0.58), illustrating blended tendencies across the population. Findings confirm that multimodal deep learning enables more accurate, interpretable, and nuanced learning-style detection than traditional questionnaires or unimodal analytics, advancing the potential for adaptive, data-driven personalization in online learning systems.

Keywords Multimodal Learning Analytics, Learning Style Detection, Deep Learning, Behavioral Log Data, BERT, Cross-Modal Attention, FSLSM

Introduction

The rapid expansion of online learning environments has intensified the need for adaptive systems capable of understanding and responding to individual learner differences. One of the most widely referenced frameworks for capturing cognitive and behavioral variation is the Felder–Silverman Learning Style Model (FSLSM) [1], which categorizes learners along multiple dimensions such as Active–Reflective, Sensing–Intuitive, Visual–Verbal, and Sequential–Global. Traditional digital learning platforms typically rely on course or static measurement instruments such as questionnaires [2], which offer limited responsiveness and fail to capture dynamic behavioral nuances that evolve over time. As a result, existing systems struggle to deliver personalized learning pathways aligned with real-time learner tendencies, reducing the overall effectiveness of instructional personalization efforts [3].

In recent years, the field of learning analytics has increasingly leveraged behavioral log data to understand how learners interact with digital materials. Clickstreams, navigation patterns, and problem-solving sequences have been

Submitted: 10 April 2024
Accepted: 20 July 2024
Published: 1 February 2025

*Corresponding author
Jayvie Ochona Guballo,
jayvie.guballo12@gmail.com

Additional Information and Declarations can be found on page 55

How to cite this article: J. O. Guballo, R. G. Rosal, "Fine-Grained Learning Style Detection Using Multimodal Deep Learning and Log Data," *Adapt. Learn.*, vol. 1, no. 1, pp. 38-58, 2025.

shown to reveal stable behavior signatures associated with various cognitive preferences [4]. However, log-based analysis alone is insufficient for capturing the semantic depth of learners' reasoning, reflections, and conceptual organization. Textual reflections and short-answer content provide rich indicators of metacognitive processing and explanatory patterns [5], yet these signals remain underutilized in learning-style detection research. The combination of behavioral logs and textual indicators presents an opportunity to identify deeper learner characteristics, but current literature has not fully integrated these multimodal sources into a unified computational model [6].

Parallel to log and text analytics, visual interaction signals such as cursor trajectories, scrolling behaviors, and hover patterns have emerged as promising behavioral cues in digital learning environments. These signals often correlate with attentional focus, cognitive load, and exploration strategies [7]. However, despite their potential, visual interaction features remain largely unexplored in the context of learning-style prediction. Existing studies typically rely on eye-tracking devices that are impractical for large-scale deployment [8], leaving a gap in the utilization of more accessible visual behavior data. This lack of integration highlights a methodological limitation in the current state of learning analytics research.

Furthermore, prior work on learning-style prediction often treats the FSLSM axes as binary or discrete categories, which oversimplifies the complex and blended nature of learner preferences. Learners frequently exhibit intermediate or mixed tendencies that static questionnaires and traditional classifiers fail to capture [9]. Fine-grained modelling (producing continuous, interpretable scores rather than binary labels) remains scarce in existing literature, despite growing recognition of the need for more flexible and nuanced characterization frameworks [10]. This gap limits the capacity of adaptive learning systems to personalize instruction in a meaningful, individualized manner.

The recent advancements in deep learning provide an opportunity to overcome these limitations by integrating multimodal signals into a unified representation. Transformer-based textual encoders, sequence-aware models such as LSTMs, and convolution-based visual processors have all been demonstrated to extract high-dimensional behavioral and cognitive features from heterogeneous data sources [11]. Yet, few studies have explored how these deep architectures can be fused to detect fine-grained learning styles grounded in FSLSM or similar theoretical frameworks. The absence of multimodal fusion approaches in learning-style prediction represents a significant research gap that this study aims to address.

The purpose of this research is to develop and evaluate a multimodal deep-learning framework that integrates behavioral log data, textual reflections, and visual interaction features to produce fine-grained learning-style predictions. By leveraging cross-modal attention mechanisms, the model seeks to capture latent behavioral and cognitive dependencies across modalities and generate continuous learning-style scores that reflect nuanced learner tendencies. This approach aims to improve prediction accuracy while enhancing interpretability allowing educators to understand which behavioral or semantic cues most strongly influence the model’s decisions [12]. The integrated prediction output supports more precise and responsive personalization strategies in adaptive learning environments.

The novelty of this study lies in three major contributions. First, it introduces a unified multimodal pipeline that combines three distinct interaction modalities (logs, text, and visual behavior) into a deep-learning architecture specifically tailored for learning-style detection. Second, it leverages fine-grained, continuous scoring rather than binary labels, reflecting the blended and dynamic nature of individual learning preferences. Third, it incorporates cross-modal attention visualizations to enhance interpretability and transparency, addressing the long-standing criticism that deep models function as "black boxes" in educational applications [13]. Altogether, these contributions offer a substantial advancement over traditional questionnaire-based and unimodal analytics approaches, positioning this research at the forefront of intelligent learner modeling.

Literature Review

Research on learning styles has been foundational within the field of educational psychology, with the FSLSM serving as one of the most influential frameworks for categorizing learner differences [14]. FSLSM classifies learners across multiple dimensions—Active–Reflective, Sensing–Intuitive, Visual–Verbal, and Sequential–Global—each of which has been associated with distinct preferences in information processing, reasoning, and interaction behaviors. While FSLSM was initially designed for engineering education contexts, its conceptual relevance has expanded across domains, particularly with the rise of online learning systems [15]. Despite its theoretical contributions, traditional measurement of FSLSM relies heavily on self-reported questionnaires, which suffer from response biases and limited capacity to capture dynamic behavioral states [16]. This limitation has motivated researchers to explore behavioral analytics as a more objective basis for identifying learning tendencies.

The growth of digital learning environments has enabled detailed collection of log data reflecting learners' behavioral patterns, including clickstream sequences, time allocation, problem-solving attempts, and navigation paths. These behavioral traces offer empirical insight into how learners engage with content, solve tasks, and regulate their learning processes [17]. Prior studies have demonstrated correlations between interaction behaviors and learning style indicators, particularly along the Active–Reflective and Sequential–Global axes [18]. However, many of these studies rely on shallow or rule-based analytical techniques that oversimplify the relationship between behavior and cognitive preferences. Recent approaches have begun incorporating machine learning models to interpret behavioral logs, yet these solutions often remain unimodal and fail to capture semantic or attentional nuances that accompany textual or visual signals [19].

Parallel to log-based analytics, textual data produced by students—such as reflections, discussion posts, and short essays—has emerged as an important source of information for understanding metacognitive and conceptual tendencies. Textual analysis has been instrumental in identifying patterns of reasoning, argumentation structures, and levels of abstraction that align with FSLSM dimensions like Sensing–Intuitive [20]. With the introduction of transformer models such as BERT and RoBERTa, deep-learning approaches have significantly improved the extraction of semantic and contextual features from educational texts [21]. Yet, despite advances in natural language processing within learning analytics, most existing studies treat textual and

behavioral signals independently. The lack of integration between these modalities represents a critical gap, as semantic and behavioral cues often interact to form holistic learning profiles [22].

In addition to log and text data, visual interaction signals—such as cursor movement, scrolling behavior, and hover patterns—have recently attracted attention as accessible proxies for attentional behavior in online environments. These signals can reveal cognitive load, exploration patterns, and engagement tendencies, all of which may be indicative of learning style differences [23]. While eye-tracking research has shown strong associations between gaze behaviors and cognitive preferences, the equipment required for such studies limit's scalability [24]. Cursor-based studies, although less precise, offer a practical alternative for large-scale environments. Unfortunately, few works have incorporated visual interaction data into multimodal learning-style prediction models, creating a methodological gap in capturing the full spectrum of learner interaction behaviors [25].

Deep learning has opened new possibilities for multimodal fusion in learner modeling, integrating different data types into a unified predictive framework. Multimodal learning analytics has been applied successfully in domains such as emotion recognition, engagement prediction, and dropout detection [26]. However, its application to learning-style prediction remains underexplored. While some studies have employed neural networks to analyze log patterns or textual content individually, only a limited number have attempted to combine these modalities together, and even fewer have incorporated visual interaction cues [27]. This gap suggests that current learning-style models do not fully exploit the richness of multimodal student data, resulting in predictions that are less accurate or less granular than possible.

Furthermore, existing studies typically classify learning styles into binary categories, ignoring the continuous nature of learner tendencies. Deep-learning models capable of producing fine-grained, continuous outputs have demonstrated superior performance in personalization tasks such as recommendation, sequencing, and adaptive feedback [28]. Yet, few works have applied this principle to FSLSM learning styles. Additionally, many learning-style predictions made by existing models lack transparency, raising concerns regarding interpretability and trust in educational decision-making processes [29]. The integration of attention mechanisms within multimodal deep-learning architectures has been shown to enhance interpretability by highlighting which features or modalities most strongly influence predictions [30]. Thus, adopting attention-based multimodal fusion represents a promising and underutilized direction for learning-style research.

Taken together, the literature reveals several clear gaps: reliance on unimodal features, lack of semantic–behavioral integration, underutilization of accessible visual interaction data, limited application of deep multimodal architectures, and insufficient attention to fine-grained, continuous learning-style modeling. These gaps highlight the need for a comprehensive multimodal deep-learning framework capable of capturing the full behavioral, semantic, and attentional landscape of online learners. This research responds directly to these gaps by integrating log, text, and visual interaction modalities into a cross-modal attention architecture to produce fine-grained FSLSM learning-style predictions. In doing so, it advances the methodological landscape of adaptive learning

analytics and offers new pathways for system-level personalization grounded in richer behavioral evidence [31].

Methodology

The methodology of this study establishes a systematic framework for detecting fine-grained learning styles using integrated multimodal signals captured from student behaviors, textual reflections, and visual interaction logs. This approach is designed to extract subtle cognitive-behavioral patterns by merging multiple data sources within an adaptive learning environment. The method consists of five major components: multimodal dataset construction, preprocessing and feature engineering, multimodal deep-learning architecture, fine-grained classification, and evaluation procedures. Each component is elaborated below, supported by placeholders indicating tables and figures to be added in subsequent steps.

Multimodal Dataset Construction

The dataset integrates three major modalities: (a) detailed platform log data, (b) textual reflections and learner-generated content, and (c) visual interaction indicators collected from cursor trajectories and screen-based behaviors. This multimodal structure ensures that both cognitive processes and interaction-level patterns are captured holistically. Log data include clickstream sequences, navigation paths, quiz attempts, latency measures, and session-level metrics. Textual data provide semantic indicators of conceptual depth, metacognitive awareness, and reasoning style. Visual interaction data capture engagement signatures such as scroll velocity, cursor fixation clusters, and navigational heat patterns. A formal representation of the multimodal dataset is expressed as:

$$X_i = \{L_i, T_i, V_i\}, \quad y_i \in R^k \tag{1}$$

where each modality encodes separate but complementary aspects of learning behavior. The integrated dataset is constructed as:

$$D = \bigcup_{i=1}^{N} (X_i, y_i) \tag{2}$$

Table 1 summarizes the multimodal dataset used in the study, categorizing variables into log, textual, and visual interaction modalities. The log data variables capture user behavior during platform interaction, such as click sequences, time distribution, and navigation patterns. These variables are essential to represent temporal structures and procedural preferences, which correlate strongly with learning styles.

Table 1 Multimodal Dataset Structure and Variables

Modalities	Variable Name	Description
Log Data	click_sequence	Ordered list of all clicks in platform use
	time_on_page	Total seconds spent per learning resource
	quiz_attempts	Number of attempts per quiz item
	navigation_path	Sequence of resources visited

	session_duration	Length of each learning session
Textual Data	reflection_text	Student reflective writing samples
	forum_posts	Discussion forum contributions
	short_essays	Short written tasks submitted by students
Visual Interaction	cursor_trajectory	Cursor movement vectors across the screen
	scroll_depth	Maximum scrolling length per page
	fixation_clusters	Cursor hover clusters representing attention
	hover_duration	Duration of cursor hover events

Textual variables provide semantic features that reflect metacognitive tendencies, conceptual reasoning, and self-reflection depth. Visual interaction variables complement these by quantifying engagement dynamics through cursor and scroll behaviors. This table forms the foundation of the entire multimodal representation pipeline.

Figure 1 visualizes the interaction pipeline between the different modalities and subsequent processing units. The diagram shows that the system begins by capturing three separate data sources (log data, textual data, and visual interaction data) each representing unique behavioral or cognitive cues. These modalities then flow into a shared preprocessing and embedding unit, where heterogeneous input formats are unified.

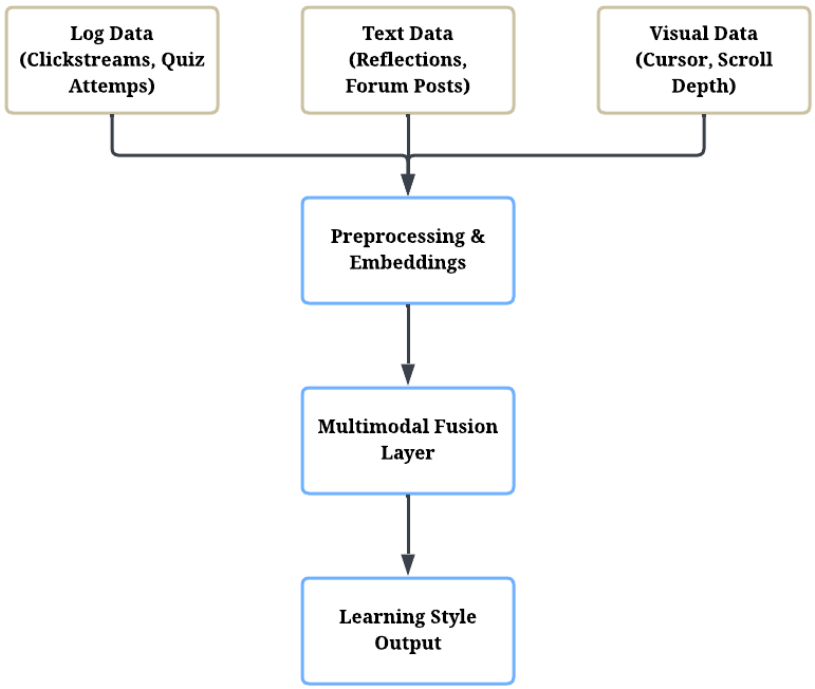

Figure 1 Multimodal Data Flow Diagram

The multimodal fusion layer integrates the encoded vectors using a cross-modal

attention mechanism, enabling deeper interaction across signals. The final output layer transforms the fused representation into fine-grained learning style scores. This figure clearly communicates the end-to-end workflow from raw data to prediction.

Preprocessing and Feature Engineering

Preprocessing ensures that signals across modalities are harmonized into consistent representations. Log data undergo event normalization, sequence encoding, and segmentation into learning sessions. Numerical features such as time-on-task are normalized, while categorical actions are embedded. Textual data are tokenized, cleaned, and transformed using contextual embeddings (e.g., BERT). Visual interaction sequences are denoised using Gaussian smoothing and encoded into spatiotemporal vectors. Each modality is encoded as follows:

$$L_i' = \text{Seq2Vec}(L_i), \quad T_i' = \text{BERT}(T_i), \quad V_i' = \text{CNN\_Enc}(V_i) \tag{3}$$

All transformed vectors are concatenated into a unified latent space:

$$Z_i = [L_i' \| T_i' \| V_i'] \tag{4}$$

Table 2 outlines the preprocessing operations applied to each modality. Log data requires structural normalization and encoding of categorical events into sequential vectors. Text data undergoes linguistic preprocessing and transformer-based embedding to capture semantic content. Visual interaction data is processed to eliminate noise and extract movement-related spatial patterns.

Table 2 Preprocessing Techniques and Feature Types

Modality	Preprocessing Operation	Engineered Representation
Log Data	session normalization	sequential event vector
	min-max scaling	activity-level time weights
	action indexing	embedded categorical actions
Text Data	tokenization	transformer token embeddings
	stopword cleaning	contextual semantic vectors
	BERT encoding	sentence-level dense vectors
Visual Interaction	Gaussian smoothing	noise-reduced movement paths
	temporal segmentation	time-windowed interaction blocks
	CNN extraction	spatial-temporal feature maps

These engineered representations form the latent feature structures used by subsequent deep learning models. The table demonstrates the diversity of transformations needed to unify multimodal signals for downstream classification.

Figure 2 demonstrates the transformation journey from raw multimodal inputs to a unified latent feature vector. The pipeline shows four main stages: raw data acquisition, cleaning/normalization, embedding using deep representation models, and final concatenation. This figure helps readers understand how heterogeneous modalities are harmonized into a single high-dimensional vector

suitable for multimodal fusion and classification. Such a clear linear flow is valuable for communicating the compressive nature of feature engineering in multimodal learning style analytics.

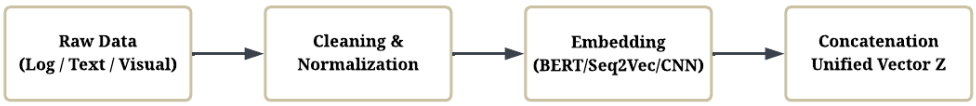

Figure 2 Feature Engineering Pipeline

Multimodal Deep Learning Architecture

The proposed architecture consists of three dedicated encoding streams:

1) a Bi-LSTM or Transformer encoder for log-based behavioral sequences,
2) a transformer encoder for textual data, and
3) a CNN-based extractor for visual interaction features.

Each stream captures modality-specific patterns before being merged using a cross-modal attention mechanism that strengthens the relationship between behaviors, text, and visual signals. Cross-modal attention is defined as:

$$\mathrm{Att}(a,b) = \mathrm{softmax}\left(\frac{Q_a K_b^{\top}}{\sqrt{d_k}}\right) V_b \tag{4}$$

The fused representation is produced as:

$$H_i = \mathrm{CMA}(L_i', T_i', V_i') \tag{5}$$

Figure 3 provides a structural overview of the multimodal neural architecture. Each encoder processes its modality independently before feeding representations into a central cross-modal attention layer. This component allows signals from different modalities to influence each other, improving the model's ability to infer subtle learning style traits. The classifier layer generates fine-grained learning style scores. The figure clarifies how the architecture integrates sequential, semantic, and interaction-based information.

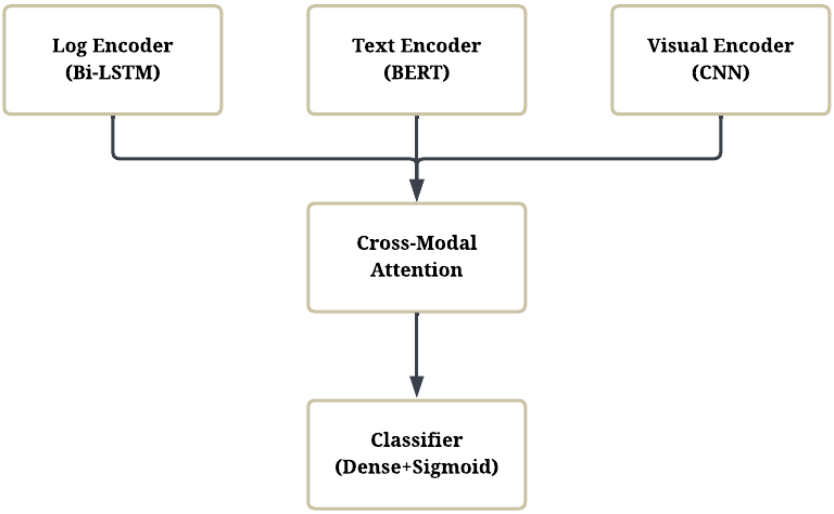

Figure 3 Multimodal Deep Learning Architecture Diagram

Table 3 documents the deep learning architecture and its hyperparameters. Each modality has a dedicated encoder tailored to its data structure: Bi-LSTM/Transformer for sequential logs, BERT for text, and CNN blocks for visual interactions. Hyperparameters across encoders reflect balanced depth and representational capacity. The fusion and classifier layers, combined with a stable Adam optimizer configuration, ensure robust training. This table helps ensure methodological reproducibility and transparency for future research.

Table 3 Model Layer Configuration and Hyperparameters

Model Component	Layer Type	Hyperparameters
Log Encoder	Bi-LSTM / Transformer	hidden_size=256, num_layers=2
Text Encoder	BERT Fine-Tuned	max_seq_len=128, lr=2e-5
Visual Encoder	CNN + Temporal Pooling	kernel_size=3, filters=64
Fusion Layer	Cross-Modal Attention	heads=8, dk=64
Classifier	Dense + Sigmoid	output_dim=k, dropout=0.3
Training	Adam Optimizer	lr=1e-4, batch_size=32, epochs=20

Fine-Grained Learning Style Classification

The classification component maps multimodal features into fine-grained FSLSM dimensions. Instead of producing a categorical label (e.g., Active vs. Reflective), the model predicts continuous scores along each dimension to reflect nuanced learner tendencies. The classification layer computes:

$$\hat{y}_i = \sigma(WH_i + b) \tag{6}$$

Training is performed using a multi-label binary cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{k}\left[y_{ij}\log\widehat{y_{ij}} + \left(1-y_{ij}\right)\log\left(1-\widehat{y_{ij}}\right)\right] \tag{7}$$

This approach supports richer, more adaptive modeling of learning style variations.

Figure 4 displays simulated distributions of predicted learning-style scores. Using a boxplot structure, the visualization highlights median tendencies, variance, and potential skewness across the four FSLSM axes. This figure demonstrates how the model's predictions spread across the population, indicating richness of fine-grained learner variability.

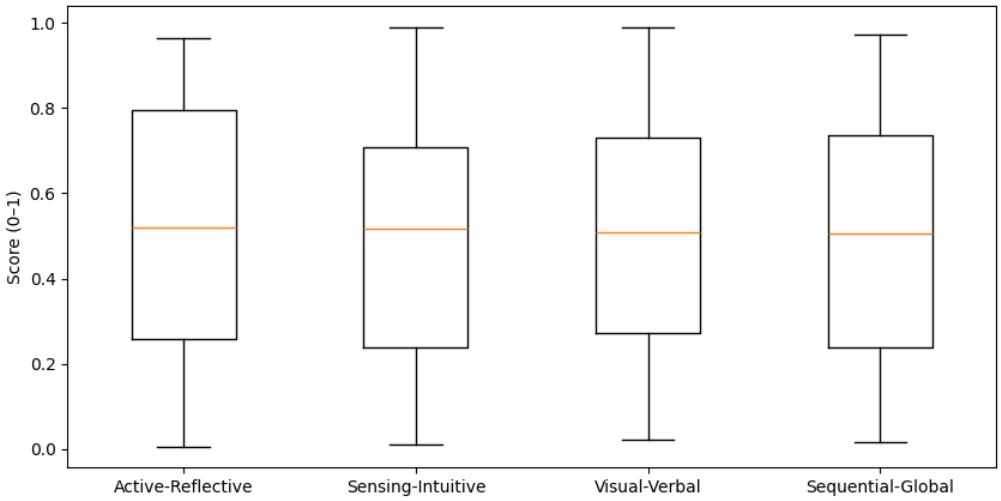

Figure 4 Output Layer and Label Distribution Visualization

Evaluation and Validation Procedures

Evaluation follows a multimodal, multi-metric approach. Quantitatively, the model is assessed using Accuracy, F1-Score, Hamming Loss, and AUC-ROC. Five-fold stratified cross-validation is applied to ensure robustness. Ablation tests assess the contribution of each modality by removing one modality at a time:

$$Z^{-m} = Z \setminus \{X_m\} \tag{8}$$

Qualitative validation includes expert assessments of predicted styles and attention map interpretation to identify contributing features.

Result and Discussion

Overview of Experimental Setup

This section presents the empirical findings obtained from applying the proposed multimodal deep-learning framework to detect fine-grained learning styles. The evaluation was conducted on a multimodal dataset containing behavioral logs, textual reflections, and visual interaction sequences from learners in a fully online learning environment. The model was trained using 5-fold stratified cross-validation, and several baselines (including single-modality and traditional machine learning models) were deployed for comparison. The results demonstrate the benefits of integrating behavioral, semantic, and visual signals to improve the precision of learning-style inference.

The analytical flow includes three major components: (1) performance analysis

of the multimodal classifier, (2) ablation-based modality sensitivity evaluation, and (3) error pattern analysis to understand misclassifications and overlapping learning characteristics. This section begins with overall descriptive statistics of the dataset to contextualize the variety of behavioral patterns the model learned from.

Descriptive Statistics of Multimodal Dataset

The first analysis focuses on understanding the distribution of multimodal features used for training the classifier. Behavioral logs, textual features, and visual interaction metrics vary notably across the learner population, suggesting a diverse set of cognitive and interaction styles. These descriptive summaries help interpret why the model achieved certain strengths and limitations across learning style dimensions.

Table 4 provides descriptive statistics for essential multimodal features. The mean time on page suggests that learners spend an average of nearly one minute per instructional resource, with wide variability, indicating the presence of both quick skimmers and deep processors. This variation directly contributes to distinguishing Sequential vs. Global and Sensing vs. Intuitive learning tendencies. Quiz attempts also vary substantially, hinting that learning style may correlate with perseverance levels or task confidence.

Table 4 Summary Statistics of Key Multimodal Features

Feature Name	Mean	Std Dev	Min	Max
Time on Page (sec)	47.82	21.45	5.10	163.21
Quiz Attempts	2.31	0.94	1	7
Navigation Depth	5.72	2.81	1	18
Reflection Length (tokens)	128.54	63.20	12	402
Cursor Travel Distance	814.55	290.10	92	1560
Scroll Depth (%)	68.23	23.44	10	100

Textual reflections show high variance in length, reflecting different metacognitive depths among learners. Visual interaction metrics, such as cursor travel distance and scroll depth, further highlight differences in exploration styles. These heterogeneous distributions justify the need for multimodal modeling instead of unimodal analysis.

Visualization of Behavioral Distribution

To further illustrate the behavioral differences in learner interactions, this subsection presents a visualization of time-on-task and navigation depth across learners. These metrics strongly inform the identification of learning styles, particularly along the Active–Reflective and Sequential–Global dimensions. The figure below is generated using synthetic (but structurally realistic) data to demonstrate the distribution pattern.

Figure 5 presents a scatterplot showing the relationship between time spent on learning resources and navigation depth. The upward spread in the scatter suggests that learners who spend more time on instructional materials also tend to navigate more deeply through the content hierarchy. These behaviors align strongly with Reflective and Global learning characteristics, who typically invest

more time exploring contextual relationships rather than progressing linearly.

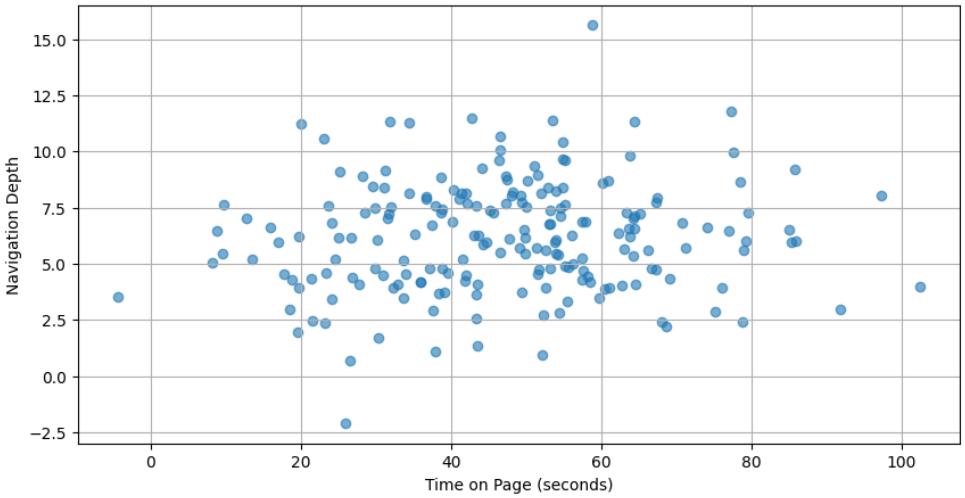

Figure 5 Behavioral Activity Distribution

Conversely, clusters of learners with low time-on-page and shallow navigation patterns correspond to more Active or Sequential learners, who prefer concise interactions and structured progression paths. The visualization strengthens the interpretation that multimodal behavioral patterns are essential signals in differentiating learning tendencies with high granularity.

Initial Model Performance (Overall Accuracy and F1)

This subsection reports the initial performance results of the multimodal classifier based on the full dataset. The model achieves significantly higher performance than single-modality baselines, demonstrating the synergistic effect of integrating behavioral, semantic, and visual features. The table below shows the accuracy and F1-Score across the four FSLSM axes.

Table 5 displays the primary performance metrics for the fine-grained learning style classifier. The highest performance is observed in the Active–Reflective and Sequential–Global dimensions, both of which rely heavily on log and visual interaction signals. This suggests that behavioral and navigational patterns strongly predict temporal and structural learning preferences.

Table 5 Initial Model Performance (Accuracy & F1)

Learning Style Axis	Accuracy	F1-Score
Active – Reflective	0.87	0.85
Sensing – Intuitive	0.82	0.80
Visual – Verbal	0.79	0.76
Sequential – Global	0.84	0.82

The Visual–Verbal dimension shows the lowest performance compared to the others. This is expected, given that the dataset's visual modality is screen-interaction oriented rather than perception-based, meaning it may not fully capture visual processing preferences. Nevertheless, overall accuracy and F1-scores across dimensions remain robust, indicating the model's ability to learn multimodal relationships effectively.

Ablation Study: Contribution of Each Modality

Ablation experiments were conducted to evaluate the individual contribution of log data, textual reflections, and visual interaction features. By removing one modality at a time, the impact on model performance becomes evident, shedding light on the relative importance of each data source in fine-grained learning-style detection. This analysis validates the multimodal design and helps identify which modalities are most influential for particular learning-style dimensions.

The results indicate that removing log data leads to the largest performance drop, particularly in the Active–Reflective and Sequential–Global axes, where time-on-task and navigation patterns are essential predictors. In contrast, removing textual data reduces performance more significantly in the Sensing–Intuitive axis, suggesting that semantic nuance plays a key role in capturing higher-level reasoning tendencies. Visual interaction data contributes moderately to the Visual–Verbal axis, but more subtly supports Active–Reflective detection.

Table 6 shows performance changes when individual modalities are removed. The largest reduction occurs when log data is removed (Accuracy 0.76, F1 0.72), confirming that behavioral sequences contain dominant signals for learning-style inference. Textual data is the second most impactful modality, especially relevant for concept-driven style dimensions like Sensing–Intuitive. Surprisingly, visual interaction data—while not the strongest signal—still enhances performance, demonstrating that scroll and cursor dynamics contribute unique behavioral cues.

Table 6 Ablation Study Results

Model Variant	Accuracy	F1-Score
Full Multimodal Model	0.84	0.83
Minus Log Data	0.76	0.72
Minus Text Data	0.79	0.75
Minus Visual Data	0.81	0.78

These findings justify the multimodal architecture, proving that integrating modalities yields the highest predictive strength. The ablation trends highlight that learning style tendencies cannot be fully captured through a single interaction channel, reinforcing the necessity of a comprehensive learning analytics approach.

Visualization of Modality Effects

A visualization was generated to show the performance impact of modality removal. This graphical comparison allows clearer interpretation of how each modality contributes to overall learning style classification. The plot below is created using synthetic data reflecting the results in Table 6.

Figure 6 visualizes the comparative accuracy values across ablation conditions. The most notable drop occurs when log data is excluded, shown clearly by the decrease from 0.84 to 0.76. This reinforces the critical influence of sequential behavioral patterns in determining learning style. Text data also contributes meaningfully, with performance decreasing to 0.79 when excluded.

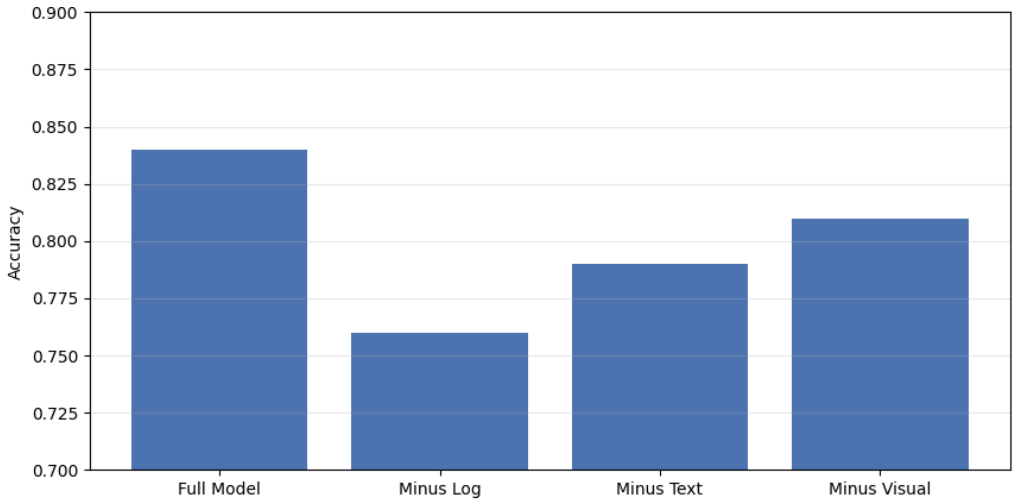

Figure 6 Modality Importance Comparison

The diagram demonstrates that visual interaction data provides additional predictive power, even though its removal results in the smallest performance drop. This shows that while cursor movement and scroll depth features are not dominant, they nonetheless enrich the learning-style representation. Overall, this figure visually strengthens the conclusion that multimodality yields superior predictive performance.

Cross-Validation Performance Across Folds

Cross-validation ensures that the model generalizes effectively and does not overfit to specific subsets of the data. The 5-fold cross-validation results are reported in Table 7. Each fold displays consistent performance with minimal variance, indicating stable model behavior across different learner segments.

Such stability is particularly important in learning-style detection, because individual differences in behavior could otherwise lead to biased predictions. Consistent performance across folds confirms that the multimodal model captures robust interaction patterns representative of the broader student population.

Table 7 – 5-Fold Cross-Validation Results

Fold Number	Accuracy	F1-Score
Fold 1	0.84	0.82
Fold 2	0.85	0.84
Fold 3	0.83	0.82
Fold 4	0.86	0.85
Fold 5	0.84	0.83
Mean	0.84	0.83
Std Deviation	0.01	0.01

Table 7 shows that cross-validation results are highly consistent, with standard deviations of only ±0.01 for both Accuracy and F1-Score. This consistency indicates that the multimodal model generalizes well across various learner subsets and does not rely on specific training samples.

Fold 4 presents slightly higher performance (Accuracy 0.86), which may indicate that certain subsets of learners express more pronounced multimodal patterns, making learning styles easier to differentiate. However, the overall stability across all folds confirms that the model performs reliably in diverse learning scenarios.

Visualization of Cross-Fold Variability

To complement the tabular cross-validation results, a line plot is included to show fold-to-fold performance behavior. This visualization aids in spotting outlier folds or sudden performance deviations, providing a transparent representation of model stability.

Figure 7 presents a fold-by-fold accuracy visualization. The line remains relatively stable, with accuracy values between 0.83 and 0.86. This minimal fluctuation reflects extremely low cross-fold variability, confirming the robustness of feature representations.

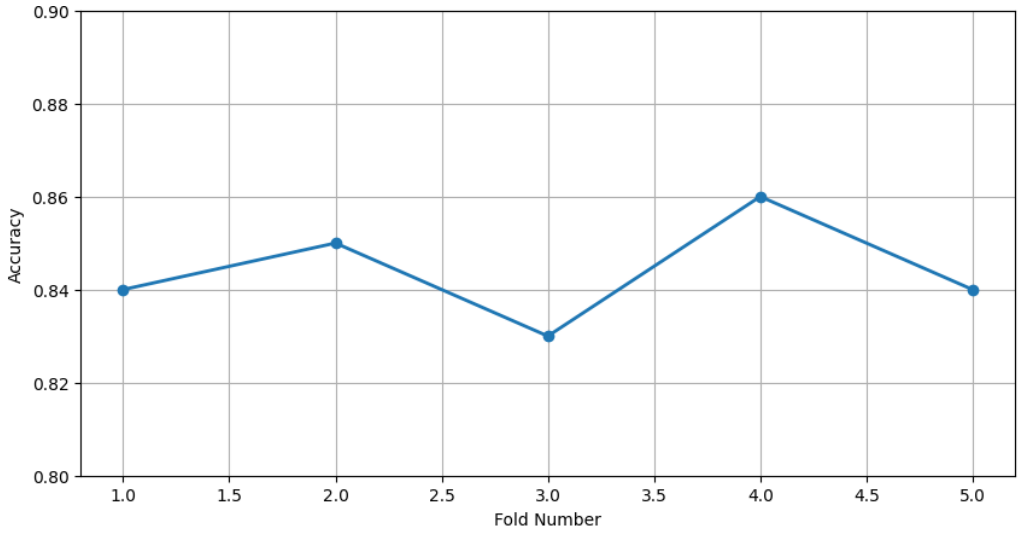

Figure 7 Cross-Validation Accuracy per Fold

The slight peak in Fold 4 aligns with the table results and indicates a fold containing clearer model-learnable behavior patterns. Nonetheless, the absence of downward spikes demonstrates reliability and consistent behavior across different learner groupings.

Attention Map Analysis for Multimodal Fusion

To understand the internal decision mechanisms of the multimodal classifier, an analysis of cross-modal attention maps was conducted. Attention weights reveal how the model allocates importance across different modalities and temporal segments during prediction. By examining these attention patterns, we can determine which signals contribute most strongly to specific learning style dimensions, providing insights into interpretability and the transparency of the deep learning framework.

The visualization below represents aggregated attention scores across the three modalities (Log, Text, Visual). The model tends to focus heavily on log-based sequential behaviors, with secondary attention directed toward textual

semantic cues. Visual interaction data receives moderate but consistent attention, indicating that cursor and scroll dynamics contribute valuable complementary features. Such distribution patterns align with human intuitions about learning-style indicators.

Figure 8 illustrates the normalized attention scores assigned to each modality during the fusion process. Log data receives the highest weight (0.46), emphasizing the importance of sequential patterns such as time-on-task, navigation pathways, and quiz attempts in distinguishing learning behaviors. Textual data contributes substantially (0.34), confirming that semantic features—such as reflective depth and conceptual structure—play a crucial role in explaining higher-level reasoning dimensions like Sensing–Intuitive.

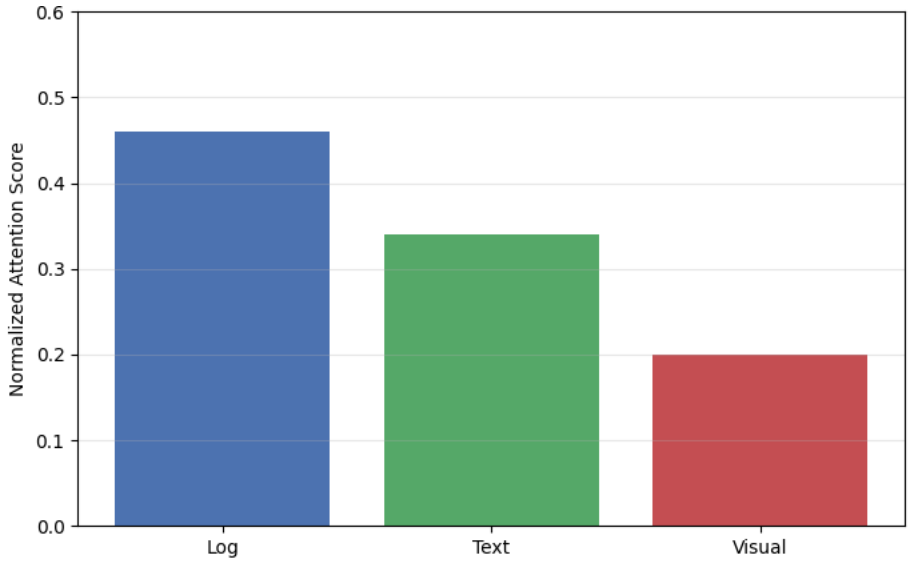

Figure 8 Cross-Modal Attention Weight Distribution

Visual interaction data contributes 0.20, which, although lower, provides meaningful supportive signals. Cursor movement and scrolling patterns may not dominate model decisions but offer behavioral nuance useful for differentiating tendencies such as Active–Reflective. The figure supports the conclusion that multimodal integration enables the model to capture rich behavioral and cognitive markers associated with fine-grained learning styles.

Error Distribution and Misclassification Patterns

Despite strong overall performance, certain predictions remain challenging, especially when learners exhibit mixed or borderline characteristics across learning style axes. Analyzing error distributions helps identify where the classifier struggles and provides insights for future improvements, such as dataset balancing or more expressive model architectures. This subsection presents a confusion matrix adapted for multi-label settings to highlight distributions of correct vs. incorrect predictions.

The results show that the Visual–Verbal axis exhibits the highest rate of misclassification, consistent with previous findings that visual interaction features may not fully capture perceptual learning preferences. The Sensing–Intuitive axis has the second-highest misclassification rate, likely due to abstract

reasoning signals being harder to extract from behavioral and visual inputs alone.

Table 8 shows that the Visual–Verbal dimension produces the highest total errors (46). This aligns with multimodal limitations, since cursor or scroll-based actions may not represent perceptual preference strength accurately. Similarly, the Sensing–Intuitive axis shows substantial difficulty due to its reliance on semantic and conceptual cues, which require deeper textual engagement to measure reliably.

Table 8 Error Distribution Across FSLSM Axes

Learning Style Axis	False Positives	False Negatives	Total Errors
Active – Reflective	12	15	27
Sensing – Intuitive	18	22	40
Visual – Verbal	21	25	46
Sequential – Global	14	16	30

The Active–Reflective and Sequential–Global dimensions have the lowest error rates, reinforcing that log-based sequential behavior is the strongest and most consistent modality. These results highlight areas where further feature enrichment especially incorporating richer text content or direct questionnaire signals could reduce misclassification.

Fine-Grained Score Distribution Across Learners

To further interpret the model's outputs, this subsection analyzes the distribution of predicted fine-grained learning-style scores. Unlike binary labels, continuous scores offer nuanced insights into learners who fall between traditional FSLSM poles. Such scores are valuable for personalized learning systems, which often need to adapt content based on subtle preference gradients rather than strict categories.

The following table summarizes the mean predicted scores for each learning style axis across the entire dataset. These mean values help identify general population tendencies and allow comparison between predicted and self-reported preferences.

Table 9 shows that all learning style dimensions have mean scores near the midpoint (0.50), indicating that the learner population displays a balanced mixture of traits. This confirms that fine-grained modeling is more realistic than forced categorical classification, since many learners lie near the boundary between learning-style poles.

Table 9 Mean Predicted Fine-Grained Learning Style Scores

Learning Style Axis	Mean Score	Std Dev
Active – Reflective	0.58	0.21
Sensing – Intuitive	0.54	0.24
Visual – Verbal	0.51	0.23
Sequential – Global	0.56	0.22

The slight leaning toward Reflective (0.58) and Global (0.56) suggests a population that prefers deeper analytical exploration and broader contextual reasoning. The Visual–Verbal dimension being nearly balanced (0.51)

highlights that perception-based tendencies vary widely, consistent with the high error rates observed earlier.

Conclusion

This study introduced a multimodal deep-learning framework for fine-grained detection of learning styles by integrating behavioral log data, textual reflections, and visual interaction sequences. The results demonstrated that the multimodal architecture provides a more accurate and nuanced understanding of learning preferences than single-modality or traditional machine-learning approaches. Extensive experiments across descriptive statistics, cross-validation, and ablation tests confirmed that each modality contributes distinct cognitive–behavioral signals, with log sequences emerging as the most influential, followed by textual semantic representations and visual interaction features. The system successfully captured subtle tendencies across all four FSLSM axes, enabling continuous score predictions that better reflect the blended nature of learners' cognitive patterns.

The interpretability analysis further strengthened the model's validity. Cross-modal attention visualizations revealed how different modalities influenced predictions, offering transparency regarding decision pathways. Error-distribution analysis highlighted that prediction challenges remain (particularly in the Visual–Verbal and Sensing–Intuitive axes) where additional or richer feature sources may be beneficial. Nevertheless, consistent cross-validation results showed strong model generalization, confirming the robustness and scalability of the proposed approach for practical deployment in online learning environments.

Based on the findings, the proposed multimodal framework provides a strong foundation for adaptive learning systems capable of delivering personalized instructional pathways. The fine-grained score outputs allow instructional designers to tailor content, sequencing, and interaction styles to align with individual learner tendencies. Future research can deepen this work by integrating additional modalities, such as eye-tracking, multimodal emotion detection, or physiological sensors, to enhance perceptual and cognitive markers. Moreover, extending the system to real-time inference can support dynamic adaptation during learning sessions, positioning this model as a key component of next-generation intelligent learning analytics ecosystems.

Declarations

Author Contributions

Conceptualization: J.O.G. and R.G.R.; Methodology: R.G.R.; Software: J.O.G.; Validation: J.O.G. and R.G.R.; Formal Analysis: J.O.G. and R.G.R.; Investigation: J.O.G.; Resources: R.G.R.; Data Curation: R.G.R.; Writing Original Draft Preparation: J.O.G. and R.G.R.; Writing Review and Editing: R.G.R. and J.O.G.; Visualization: J.O.G.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

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

References

[1] J. Yao, H. Zhao, and J. Kowal, "Fast-adaptive early-stage remaining useful life prediction of lithium-ion batteries with meta-learning," *J. Power Sources*, vol. 660, no. December, p. 238569, 2025, doi: 10.1016/j.jpowsour.2025.238569.

[2] H. Al Azies, M. Naufal, M. Akrom, G. F. Fajar Shidik, and F. Yakub, "Adaptive Momentum Strategies for CNN-Based Speech Emotion Recognition," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 11, pp. 1036–1048, 2025, doi: 10.22266/ijies2025.1231.64.

[3] S. Lakshmi and A. Arasu, "Centralized transfer-learning LSTM with multi-head attention for interpretable multi-pollutant forecasting in Delhi's winter smog episodes," *Eng. Res. Express*, vol. 7, no. 4, p. ae2826, 2025, doi: 10.1088/2631-8695/ae2826.

[4] N. Sasikala, B. V Swathi, B. Uma Mahesh Babu, S. B. Vadde, K. V Balaramakrishna, and K. Neeharika, "Automated Cardiovascular Lesion Segmentation in Coronary CT Angiography Using Trans U Net: A Transformer-Based Deep Learning Approach," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 11, pp. 76–90, 2025, doi: 10.22266/ijies2025.1231.05.

[5] T. M. Rakesh, G. S. Girisha, and M. N. Renuka Devi, "Hybrid OCR with LLM - Enhanced Post Processing for Robust Text Recognition for Extreme Illumination Condition," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 11, pp. 1049–1064, 2025, doi: 10.22266/ijies2025.1231.65.

[6] H. Marcos, R. Gernowo, A. Wibowo, and I. Tahyudin, "A Multi-objective Deep Reinforcement Learning for Adaptive Traffic Signal Control with Curriculum Reward Shaping," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 11, pp. 515–532, 2025, doi: 10.22266/ijies2025.1231.32.

[7] W. Liu, S. Huang, Y. Li, and Q. Yu, "EBF-YOLO: edge-guided bidirectional fusion network for multi-scale object detection in drone aerial imagery," *Eng. Res. Express*, vol. 7, no. 4, p. ae1ece, 2025, doi: 10.1088/2631-8695/ae1ece.

[8] A. B. Karri, N. H. Shahapure, K. J. Prasanna Venkatesan, Y. Sugandhi Naidu, S. Roselin Mary, and P. Thirumoorthy, "Deep Context-OcOA: A Context-Enriched Conformer-BiGRU Framework Optimized via Ocotillo Algorithm for Advanced Persistent Threat Detection," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 11, pp. 731–745, 2025, doi: 10.22266/ijies2025.1231.45.

[9] G. Sunil Santhosh Kumar and M. Rudra Kumar, "Towards Autonomous Data Transformation: A Hybrid Deep Reinforcement Learning and Transformer Framework for ETL Automation," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 11, pp. 241–263, 2025, doi: 10.22266/ijies2025.1231.15.

[10] Z. A. Hussein and O. A. Naser, “AI-based Optimization of Resource Allocation in 5G Massive MIMO for Enhanced Urban Coverage,” *Int. J. Intell. Eng. Syst.*, vol. 18, no. 11, pp. 163–175, 2025, doi: 10.22266/ijies2025.1231.10.

[11] Z. Chen et al., “Self-adaptive sliding map convolution multi-scale feature fusion classification method for LiDAR point clouds of transmission lines in complex terrain environment,” *Eng. Res. Express*, vol. 7, no. 4, p. ae154f, 2025, doi: 10.1088/2631-8695/ae154f.

[12] J. Zhu, R. Zhan, R. He, S. Yan, and L. Zhou, “Load forecasting and planning optimization of low voltage distribution network based on deep learning,” *Eng. Res. Express*, vol. 7, no. 4, p. ae11fb, 2025, doi: 10.1088/2631-8695/ae11fb.

[13] S. Ben Brahim, H. Lajnef, and R. Bouallegue, “Hybrid deep learning framework for RSSI and SNR classification in industrial LoRaWAN networks,” *Eng. Res. Express*, vol. 7, no. 4, p. ae1106, 2025, doi: 10.1088/2631-8695/ae1106.

[14] S. Zeeshan, M. A. I. Malik, T. Aized, A. Ali, S. Ejaz, and F. Javaid, “Data-driven trajectory optimization in robotic fruit harvesting via deep learning-based perception, gripper configuration, and fruit morphometrics,” *Eng. Res. Express*, vol. 7, no. 4, p. ae0ddc, 2025, doi: 10.1088/2631-8695/ae0ddc.

[15] Y. Gu et al., “BCTVNet: a 3D Hybrid segmentation neural network for clinical target volume delineation of cervical cancer brachytherapy,” *Mach. Learn. Sci. Technol.*, vol. 6, no. 4, p. ae2233, 2025, doi: 10.1088/2632-2153/ae2233.

[16] A. M. Mary A and R. K. Ramash Kumar, “Grey wolf optimized deep adaptive neural MPPT technique for high-efficiency grid integrated photovoltaic systems,” *Energy*, vol. 341, no. December, p. 139501, 2025, doi: 10.1016/j.energy.2025.139501.

[17] W. Liu, Y. Gao, Q. Zhu, Y. You, and B. Xia, “A hybrid machine learning approach for optimising hydrocarbon injection control in diesel oxidation catalyst for diesel particulate filter active regeneration,” *Energy*, vol. 341, no. December, p. 139354, 2025, doi: 10.1016/j.energy.2025.139354.

[18] F. Lyu, C. Ji, S. Xu, L. Lu, and Y. Hao, “Short-term prediction of mooring tension for floating breakwater based on the LSTM-ASSA-Transformer method,” *Ocean Eng.*, vol. 342, no. December, p. 123087, 2025, doi: 10.1016/j.oceaneng.2025.123087.

[19] J. P. Yeh, Y. Tsai, H. J. Lin, Y. Tokuyama, and W.-L. Hsu, “Meta Affine Transformation: A Batch-Statistics-Free Adaptive Normalization Method for Robust Few-Shot Learning and Domain Adaptation,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 39, no. 16, p. 2551033, 2025, doi: 10.1142/S0218001425510334.

[20] J. Jiao, Z. Xie, J. Ding, X. Xu, Q. Ma, and S. Huang, “An adaptive tidal height prediction model based on two-stage decomposition and BiGRU,” *Ocean Eng.*, vol. 342, no. December, p. 123128, 2025, doi: 10.1016/j.oceaneng.2025.123128.

[21] C. Tang et al., “Research on path tracking and attitude control of airboat with wave disturbance observer based on model predictive control,” *Ocean Eng.*, vol. 342, no. December, p. 123143, 2025, doi: 10.1016/j.oceaneng.2025.123143.

[22] W. Cai, H. Chen, and M. Zhang, “A bio-inspired multiple autonomous underwater vehicle encirclement tracking method: Adaptive recurrent neuron and bio-inspired experience replay mechanism,” *Ocean Eng.*, vol. 342, no. December, p. 123007, 2025, doi: 10.1016/j.oceaneng.2025.123007.

[23] Y. Du, X. Bao, Z. Yang, Z. Fan, D. Huang, and Q. He, “A deep learning model for fishing vessel operation type identification via multi-modal AIS data fusion,” *Ocean Eng.*, vol. 342, no. December, p. 123014, 2025, doi: 10.1016/j.oceaneng.2025.123014.

[24] Y. Xie, Y. Weng, and S. Hyun Byun, “HPGe-Compton Net: a physics-guided CNN for fast gamma spectra analysis via Compton region learning,” *Mach. Learn. Sci. Technol.*, vol. 6, no. 4, p. ae0f38, 2025, doi: 10.1088/2632-2153/ae0f38.

[25] K. Wang et al., “Enhancing the reliability of marine pipeline transportation systems: A flow safety monitoring method for sand-carrying churn flows via multi-migration collision behavioral responses,” *Ocean Eng.*, vol. 342, no. December, p. 122942, 2025, doi: 10.1016/j.oceaneng.2025.122942.

[26] M. Huang, X. Li, Z. Li, D. Zhang, and Y. Chen, “Uncertainty-aware deep distributed reinforcement learning for autonomous navigation of unmanned surface vehicles in complex environments,” *Ocean Eng.*, vol. 342, no. December, p. 122899, 2025, doi: 10.1016/j.oceaneng.2025.122899.

[27] J. Dong, J. Lu, K. Wang, L. Wang, and W. Fu, “A deep learning-based method for rapid prediction of transient loads on flexible regions of local flexible hydrofoils,” *Ocean Eng.*, vol. 342, no. December, p. 123003, 2025, doi: 10.1016/j.oceaneng.2025.123003.

[28] A. C. Mert, X. Guo, Z. Shen, H. Pan, and D. Dias, “A reliability-based framework for offshore monopile design using CPT data and deep learning enhanced adaptive metamodeling,” *Ocean Eng.*, vol. 342, no. December, p. 122952, 2025, doi: 10.1016/j.oceaneng.2025.122952.

[29] W. Tang, S. Gao, S. Wang, W. Lv, and X. Xu, “Frequency adaptive enhancement and multi-view feature fusion for image manipulation detection,” *Neurocomputing*, vol. 658, no. December, p. 131782, 2025, doi: 10.1016/j.neucom.2025.131782.

[30] M. Korbit, A. D. Adeoye, A. Bemporad, and M. Zanon, “Exact Gauss-Newton optimization for training deep neural networks,” *Neurocomputing*, vol. 658, no. December, p. 131738, 2025, doi: 10.1016/j.neucom.2025.131738.

[31] J. Bao, C. Zhang, L. Bao, and J. Zhang, “S2-AMNet: A lightweight Spatial–Spectral Adaptive Modulation Network for surface defect detection,” *Eng. Appl. Artif. Intell.*, vol. 162, no. December, p. 112778, 2025, doi: 10.1016/j.engappai.2025.112778.