https://doi.org/10.31449/inf.v50i11.13444 Informatica **50** (2026) 445–460 **445**

Residual-Attention Networks for NILOS-Based Learning Outcome Prediction in Mobile Remote Biology Laboratories

Xinxin Cui, Chenguang Feng and Yanlong Chen
School of Life Science and Technology, Northwestern Polytechnical University, Xi'An 710129, China
E-mail: ylchen8895202@126.com, xinxincui_npu@163.com, cgfeng2025@163.com

Keywords: NILOS-based learning outcome prediction, mobile remote laboratories, residual–attention networks, explainable artificial intelligence, robustness evaluation, behavioral learning analytics

Received: January 17, 2026

Mobile remote biology laboratories produce diverse interaction data which contains variations in user behavior and operational disruptions and unpredictable session patterns that impede accurate assessment of learning outcomes. The study introduces a residual-attention learning framework for predicting session outcomes which uses the Quantum Residual Attentive Fusion Network (QRAF-Net) as its fundamental technology. The Neuro-Integrated NILOS outcome states (NILOS) taxonomy classifies learner achievements into different categories. The framework employs multiple techniques which include dependency-based feature refinement and cognitive–behavioral signal decomposition and residual feature propagation and attention-driven interaction modeling and quantum-inspired phase modulation to produce consistent behavioral representations. The research examines 254,020 actual laboratory sessions which were obtained from a virtual experimentation platform that supports mobile access. The federated learning setup tests six decentralized clients to determine system robustness under situations that involve non-identically distributed data. The study evaluates performance through these metrics accuracy and precision and recall and F1-score and ROC–AUC and Outcome Stability Index (OSI). The QRAF-Net model undergoes comparison with twelve deep learning and transformer-based baseline models which utilized the same data preprocessing and data partitioning methods. The proposed approach achieves 97.8% accuracy and an OSI of 0.956 which exceeds all baseline models by 7 to 18 percent in standard evaluation metrics. The research demonstrates stable performance and strong generalization through ablation studies and statistical validation and sensitivity analysis and robustness tests that evaluate network latency and behavioral noise and partial interaction loss. The SHAP explainability analysis shows that learning gain and concept mastery along with procedural consistency serve as the main factors that drive predictive power for the framework's interpretability and practical usage.

Povzetek: Študija predstavi QRAF-Net, rezidualno-pozornostni okvir za napoved izidov sej v mobilnih oddaljenih bioloških laboratorijih, ki z večnivojskim modeliranjem vedenjskih signalov in federiranim učenjem robustno ocenjuje učne dosežke kljub neenakomernim in motenim interakcijskim vzorcem.

1 Introduction

Research on AI and digital learning has altered lab-based teaching. Remote and virtual labs enable students explore without time or place limitations [1]. In biology and biomedical education, AI-enabled remote labs increase accessibility, standardize experimental settings, and enable large-scale deployment [2]. Mobile technology provide flexible lab work for students. Remote labs provide many pedagogical and logistical advantages, but assessing learning consequences is difficult. Remote labs track task completion, engagement, response patterns, and adaptive learning. They don't predict outcomes because cognitive effort, behavioral consistency, and interaction efficiency affect final evaluation scores during learning [3]. Data-driven learning analytics analyzes learner behavior metrics.

Educational data mining suggests interaction-based and behavior-oriented cognitive proxies for digital learning. Recent research shows that session-level interaction log aggregation allows large machine learning models and learning dynamics [4]. Mobile lab systems must summarize nuanced temporal data without losing interpretability or robustness. Mobile learning environments are unpredictable, making outcome analysis challenging. Hardware heterogeneity, network latency, session length, and learner environment might impact interface and performance. Recent studies show that these criteria are crucial to relevant and fair AI-supported educational system assessment [5]. If not done, biased or unexpected forecasts might impair instructional decision-making. Educational dataset imbalance is discussed in modern literature. Few students succeed or suffer, therefore outcomes are uneven. Recent research suggests that assessment systems should understand learning stage ordinal structure and punish severe misclassifications over minor deviations, especially in high-stakes educational settings [6]. Learning outcome classification systems must link prediction performance to instructional relevance.

AI-enabled, mobile-accessible remote biology labs need rigorous, interpretable, data-driven learning outcome categorization. The research also predicts learning effects from varied and noisy mobile laboratory data. Uneven and incomplete session behaviors, execution patterns, and evaluation signals may be affected by usage disturbances.

Research objectives and questions The study intends to enhance assessment reliability for learning outcomes in mobile remote biology laboratories because students exhibit different behavioral patterns and their interaction logs contain errors while the laboratory environment remains unstable. The research focuses on understanding how meaningful outcome evaluation can be achieved under real-world educational constraints rather than controlled laboratory settings. The study aims to answer the following research questions:

- How can learning outcomes be reliably assessed in mobile remote laboratory environments characterized by irregular interaction patterns and incomplete data?
- What behavioral factors most strongly influence learner achievement during remote biology laboratory sessions?
- The study investigates multiple outcome evaluation methods to determine their ability to produce reliable results when faced with three specific challenges which include noise and connectivity problems and varying levels of student participation.

The framework for predicting session-level learning outcomes uses these objectives as its basis for designing and assessing its effectiveness.

Contributions The main contributions of this paper are as follows:

1. Operationalizes the Neuro-Integrated NILOS outcome states (NILOS) taxonomy provided in the dataset by mapping raw mobile laboratory interaction traces to session-level outcome states, enabling reliable outcome prediction directly from behavioral data rather than post-hoc grading artifacts.

2. Proposes the Quantum Residual Attentive Fusion Network (QRAF-Net) to learn from NILOS-labeled mobile laboratory sessions by preserving fine-grained interaction signals, stabilizing deep feature propagation through residual learning, and selectively emphasizing behavior dependencies that are critical for distinguishing adjacent outcome states such as Adaptive and Proficient.

3. Introduces the Outcome Stability Index (OSI) to quantify the consistency of NILOS predictions under realistic mobile learning disturbances, including partial interaction loss and irregular learner behavior, thereby extending evaluation beyond accuracy toward robustness required in interactive mobile environments.

The remainder of this article is organized as follows. Section 2 identifies gaps for future research on predicting learning outcomes, mobile learning analytics, and AI-enabled remote laboratory systems. Section 3 proposes a modeling framework. It requires dataset preparation, behavioral feature enhancement, and learning result classification architecture. The experimental design, evaluation criteria, and real-world mobile contact comparison performance analysis are covered in Section 4. Section 5 discusses managerial and educational implications, and Section 6 explores the future of scalable and adaptable mobile laboratory learning systems.

2 Related work

Many researchers have used data and deep learning systems to estimate student performance and learning outcomes. Recently, hybrid models that capture complicated digital learning behaviors using attention processes, deep neural networks, and structured representations have become popular. Transformer-based designs mimic long-range interactions, making them perfect for student performance prediction. The multi-head attention-driven paradigm [7] merges temporal learning traces to predict academic achievement across courses. The model was highly predictive but depended on well-structured time-series data and performed poorly under sparse or irregular interaction patterns. Using an attention with quantized embedding, hierarchical learning dynamics' intra- and inter-session interactions were characterized [8]. Aggregated session-level data hinder clear temporal ordering. Many projects improved deep knowledge tracing using attention techniques.

Another study [9] built a modified attention band-based DKT model to capture short- and long-term learning effects. Successfully assessed concept mastery on tutoring datasets but did not identify learning outcomes. One study in [10] used a decoder-only unified Transformer to represent instructional signals. The method is simple but needs big interaction histories, which remote learning systems may not provide.

The graph-based learning paradigm connects learner, activity, and consequence. A hybrid graph neural network and Transformer architecture predicted multi-label performance utilizing temporal and relational links [11]. Despite greater accuracy, the model remained computationally expensive and needed robust graph topologies. Hypergraph-based modeling improved prediction with higher-order learning element connections [12]. Learning these concepts is tough, particularly for educational decision-making. Many studies have addressed multi-dimensional and mul-

Table 1: Summary of representative related work on NILOS-based learning outcome prediction

Ref	Method Used	Problem Identified	Achieved	Limitations
[7]	Multi-head attention-based deep learning	Predicting academic achievement from temporal learning traces	High predictive performance on structured course data	Requires well-structured time-series data; weak under sparse interactions
[8]	Attention with quantized embedding	Modeling intra- and inter-session learning dynamics	Captured hierarchical learning interactions	Aggregated session data loses temporal ordering
[9]	Attention-based deep knowledge tracing	Capturing short- and long-term learning effects	Accurate concept mastery estimation	Does not directly predict learning outcomes
[10]	Decoder-only unified Transformer	Representing instructional signals	Simplified sequence modeling	Requires long interaction histories
[11]	Graph neural network + Transformer	Multi-label performance prediction using relational data	Improved accuracy via graph structure	High computational cost; needs reliable graph topology
[12]	Hypergraph-based modeling	Modeling higher-order learning relationships	Enhanced prediction accuracy	Difficult to interpret and deploy
[13]	Multidimensional deep learning model	Integrating behavioral and contextual features	Captured multi-source learning data	Sensitive to class imbalance
[14]	Cross-attention multimodal model	Collaborative learning analysis across modalities	Improved multimodal fusion	Requires synchronized multimodal inputs
[15]	Meta-heuristic optimized hybrid model	Parameter tuning for complex temporal models	Improved training effectiveness	Computationally intensive optimization
[16]	Reinforcement learning + TCN	Adaptive learning analytics	Dynamic performance adaptation	Needs continuous feedback signals
[20]	Quality-aware sparse-attention model	Learning from sparse interaction records	Increased resilience to data sparsity	Focused only on prediction accuracy
[21]	Theory-informed deep model	Improving interpretability in learning analytics	Enhanced explainability	Reduced flexibility in dynamic environments

timodal learning analytics beyond graph and Transformer models. A multi-dimensional deep learning system [13] predicts performance using behavioral, assessment, and contextual data. The whole feature engineering-based method proved weak under class imbalance. To compare collaborative learning across data streams, a cross-attention-based multimodal model was built [14]. The technique requires coordinated multimodal inputs, which remote laboratories may lack.

Hybrid deep model optimization research is new. Meta-heuristic optimization was used to adjust parameters in a multimodal temporal model with convolutional and attention-based components [15]. Optimization made training hard but effective. Adaptive learning analytics employs reinforcement learning-augmented temporal convolutional frameworks [16], which need frequent feedback and extended training. Technology-mediated laboratory environments increasingly rely on digital interaction traces to evaluate learner achievement beyond traditional assessments. Remote laboratory architectures incorporating online experimentation platforms have demonstrated that behavioral engagement and task execution patterns can be used to infer learning success and skill acquisition, especially when physical access to laboratories is limited [17]. Complementary studies on assessment integrity in teaching laboratories highlight that outcome validity depends on procedural correctness, conceptual understanding, and consistency across experimental activities rather than solely on final scores [18]. Furthermore, research on hybrid and online laboratory experiences shows that interaction frequency, persistence, and collaborative engagement significantly influence outcome quality, emphasizing the need for models capable of capturing dynamic session-level behaviors in distributed learning environments [19].

Education datasets must address data sparsity and outcome imbalance, suggests new study. In restricted interaction records, quality-aware and sparse-attention models increased resilience [20]. These models solely assessed prediction accuracy, not learning outcomes' cordiality. Recent research [21] emphasizes theory-informed deep structures for interpretability but sacrifices flexibility for explainability. Studies recommend transformers, hybrid deep networks, and optimization-enhanced models for learning analytics. Many strategies need extensive temporal data, strong interaction structures, or balanced result distributions, rendering them unsuitable for mobile-accessible, AI-enabled remote labs.

The existing techniques shown in Table 1 require structured temporal data and balanced distributions together with multimodal inputs, which makes their use impossible in mobile remote laboratory settings that exhibit unpredictable conduct and environmental disturbances and incomplete user interaction data. The framework described in the subsequent section exists because the existing system limitations demand its creation.

3 Proposed system model

Based on the limitations identified in the reviewed literature, this section presents the proposed methodological framework for NILOS-based learning outcome prediction in mobile remote biology laboratories. The suggested technique includes data gathering, quality control, behavioral feature representation and refinement, deep hybrid modeling, parameter optimization, and performance assessment. The Quantum Residual Attentive Fusion Network (QRAF-Net) processes session-level mobile laboratory interaction data and predicts learner outcome attainment stages in a stable and interpretable way. The mathematical formulation of feature transformations, residual learning, attention-based interaction modeling, and optimization objectives are detailed in the following subsections to ensure transparency, reproducibility, and alignment with practical mobile learning environments. The pictorial view of the proposed model is shown in Figure 1.

The complete NILOS-based learning outcome prediction pipeline is detailed in Algorithm 1.

Algorithm 1 End-to-End QRAF-Net NILOS-based learning outcome prediction Pipeline

Require: Remote Biology Laboratory Interaction Dataset $\mathcal{D}$
Ensure: Predicted NILOS outcome states and evaluation metrics
Data Preparation
Load raw session logs from $\mathcal{D}$
Remove incomplete or corrupted records
Normalize numerical attributes using median-based scaling
Cognitive–Behavioral Signal Decomposition
Transform interaction sequences into behavioral signal components
Extract temporal, procedural, and assessment features
Dependency-Based Feature Refinement
Compute feature relevance and redundancy measures
Remove redundant attributes using dependency thresholds
Construct refined feature vector X
Dataset Partitioning
Split data into training, validation, and test sets
Apply class balancing if required
Model Training
Initialize QRAF-Net parameters P
Optimize parameters using EGAGO (Algorithm 3)
Train model on training data until convergence
Prediction
Generate session-level NILOS outcome predictions on test data
Evaluation
Compute accuracy, precision, recall, F1-score, and OSI
Perform robustness testing under disturbance scenarios
Explainability Analysis
Apply SHAP to estimate feature contributions
Rank features by importance
return Predicted outcomes, performance metrics, and explanations

3.1 Dataset description

This research employs a third-party educational analytics dataset publically available on Zenodo and contributed by NUHS [22], Singapore, via its academic teaching and digital learning activities. To facilitate mobile learning analytics and AI-enhanced laboratory instruction replication, the dataset was gathered separately. The data source anonymized all entries before release. Nothing personally identifiable is in the dataset. It includes recordings from mobile-accessible virtual experimentation platform sessions in a remote biology lab. Each record is a whole lab session including interactions and performance. This table-based dataset monitors 254,020 incidents over time and summarizes them every session. This method depicts mobile and resource-conscious learning system deployments and provides supervised classification without time-series modeling.

3.2 Cognitive–behavioral signal decomposition for remote laboratories

A lightweight cognitive-behavioral signal decomposition step during preprocessing strengthens neuro-analytic learning signals from mobile-based remote laboratory sessions [23, 24]. Session-level tabular data is used in the final learning model, but each session also provides time-sorted interaction events that measure user engagement and execution speed. These events are coupled to a continuous activity signal to summarize brain function. Allow a session to arrange interactions.

$$\Xi = \{\xi_1, \xi_2, \ldots, \xi_N\}, \tag{1}$$

At time $\omega_i v_i \omega_i$ for each contact. The associated activity signal is

$$g(v) = \sum_{i=1}^{N} \mathbf{1}(\xi_i \in L) \cdot \rho(v - v_i), \tag{2}$$

L represents learning-relevant activities and $\rho(\cdot)$ is a localized response kernel. This indicates how many actions, hesitations, and engagement occurred throughout a session. Disaggregating the data separates key learning activity from interaction noise.

$$g(v) = g_{\text{core}}(v) + g_{\text{var}}(v) + g_{\text{err}}(v), \tag{3}$$

g_{core} shows continuous cognitive progress, g_{var} adaptive behavioral variances, and g_{err} noisy or erratic behaviors [25]. Decomposition is done via spectrum-based or wavelet-inspired filtering, not time-series modeling. Through combination, we gain succinct session-level descriptions from each component[26]:

$$\zeta_k = [\mu(g_k), \sigma(g_k), \eta(g_k)], \quad k \in \{\text{core}, \text{var}, \text{err}\}, \tag{4}$$

$\mu(\cdot)$, $\sigma(\cdot)$, and $\eta(\cdot)$ represent $n \downarrow$, dispersion, and normalized energy. Descriptors retain neuro-analytic data while removing short-lived noise. These properties are added to the final dataset, which has a non-time series structure and can easily categorize mobile learning results.

The decomposition process used a wavelet-based smoothing method which employed a Gaussian kernel to separate long-term behavioral trends from short-term fluctuations. The kernel bandwidth was selected empirically to preserve session-level activity patterns while suppressing high-frequency noise. The energy-based partitioning of the filtered signal enabled us to extract components which represented core progress adaptive variation and erratic behavior. The parameter values were selected based on preliminary validation experiments to achieve stable representation through different session types.

3.3 Behavioral feature representation and refinement

Database collection and quality verification precede feature processing for the Remote Biology Laboratory Interaction

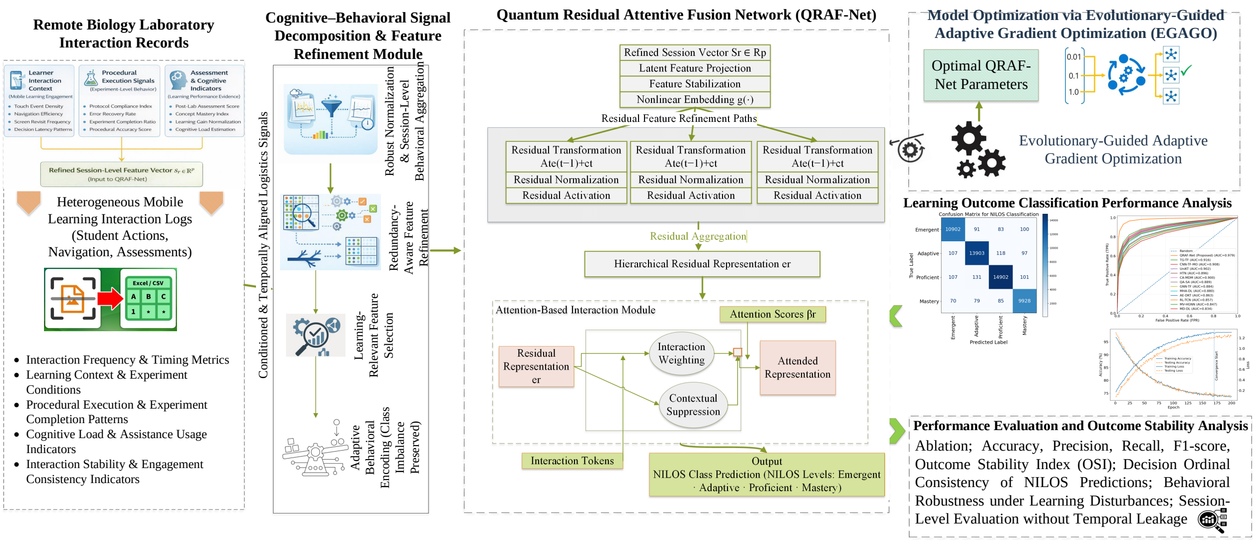

Figure 1: Proposed framework of learning outcomes in AI-enabled remote biology laboratories

Dataset (RBLID). RBLID now stores session-level data. Each row comprises a full mobile-based laboratory session including interaction, execution, and evaluation data. The first input matrix contains these records [27].

$$X = \{x_1, x_2, \ldots, x_n\}, \tag{5}$$

Each x_r represents a lab session with distinct qualities. RBLID characteristics are heavily adjusted before processing since they originate from different sources and scales. To obtain the normalized value of an attribute a_j, perform these steps:

$$a'_j = \frac{a_j - \mathrm{mid}(a_j)}{\mathrm{rng}(a_j)}, \tag{6}$$

$\mathrm{mid}(\cdot)$ represents the median, while $\mathrm{rng}(\cdot)$ represents the range between the first and third quartiles. This adjustment maintains similarity between high- and low-interaction sessions without biasing behavior. To ensure dataset continuity, RBLID incomplete items are handled before refining. Feature-level central tendency replaces missing attribute values:

$$a_j \leftarrow \begin{cases} a_j, & \text{if observed,} \\ \mathrm{mid}(a_j), & \text{otherwise.} \end{cases} \tag{7}$$

The dataset's distribution characteristics are preserved without artificial fluctuation in this step. Before learning, the normalized RBLID is refined to remove superfluous features. Information that is repeated is found via pairwise dependency between characteristics. For two normalized characteristics (a'_i, a'_j), dependency is quantified as

$$\mathrm{rel}(a'_i, a'_j) = \frac{\mathrm{cross}(a'_i, a'_j)}{\mathrm{var}(a'_i) \cdot \mathrm{var}(a'_j)}, \tag{8}$$

$\mathrm{cross}(\cdot)$ represents covariance, while $\mathrm{var}(\cdot)$ represents dispersion. A characteristic is lost once dependency exceeds a specific threshold. This step creates a modest, stable representation of key learning patterns. This stage produces X^*, an upgraded RBLID that directly feeds the learning outcome model. This process ensures that the model uses realistic, noise-free, and informative inputs, like mobile labs, by precisely describing how data goes from raw session records to refined feature representations.

The procedure of redundancy pruning used pairwise dependency analysis to identify features that had dependency scores above the established threshold as redundant features. The study used a threshold value of 0.85 to exclude attributes that were highly correlated while maintaining their ability to provide useful information. The selected threshold divided between dimensionality reduction and information preservation while safeguarding all critical behavioral indicators from deletion.

3.4 Learning outcome classification using the quantum residual attentive fusion network (QRAF-Net)

After filtering and enhancing features, each Remote Biology Laboratory Interaction Dataset record represents a completed mobile-based laboratory session. A session-specific fixed-length numerical vector combining learner engagement, procedure execution, and assessment performance. The pictorial view of architecture is shown in Figure 2.

The enhanced dataset should be referred to as

$$X = \{s_1, s_2, \ldots, s_m\}, \tag{9}$$

Each session vector $s_r \in \mathbb{R}^p$ represents the r-th lab session with p behavioral variables. The proposed deep hybrid model, the Quantum Residual Attentive Fusion Network (QRAF-Net), models hierarchical abstraction, residual information flow, and attention-based feature interaction to

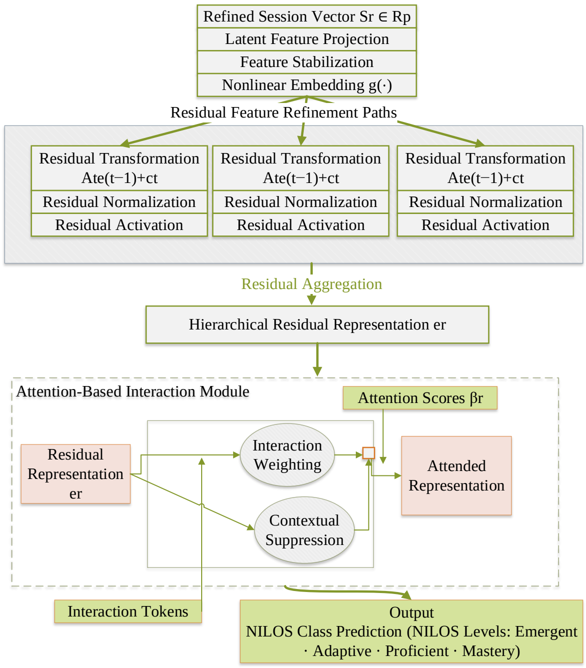

Figure 2: Proposed QRAF-Net architecture

predict each session's learning outcome. Each input session vector is first embedded in a multidimensional latent space via a nonlinear transformation.

$$e_r^{(0)} = g(A_0 s_r + c_0), \tag{10}$$

A learnable projection matrix $A_0 \in \mathbb{R}^{q \times p}$, a bias vector $c_0 \in \mathbb{R}^q$, and a nonlinear activation function $g(\cdot)$ are all part of the given equation. The embedding process represents several behavioral indicators in one location.

Residual changes enable QRAF-Net to refine features deeply while preserving behavioral signals. The t-th residual step changes the representation to

$$e_r^{(t)} = g(A_t e_r^{(t-1)} + c_t) + e_r^{(t-1)}, \tag{11}$$

Given that A_t and c_t are learnable parameters. This residual technique improves representational depth while conserving lower-level behavioral information needed to assess learning results. After processing the residual hierarchy output, an attention-driven interaction mechanism uses each behavioral characteristic as an interaction token. Attention weights are calculated:

$$\beta_r = \text{softmax}\left(\frac{(Le_r) \odot (Re_r)}{\sqrt{q}}\right), \tag{12}$$

The latent dimensionality is denoted by q, element-wise interaction is represented by $\odot$, and the learnable projection operators are L and R. This approach helps the model comprehend how behavioral qualities like execution efficiency and assessment consistency interact. Quantum-inspired phase modulation improves discrimination beyond classical attention in QRAF-Net. As mentioned,

$$\vartheta_r = \cos(\Omega \beta_r + \delta), \tag{13}$$

The vector learning process with its resulting vector δ together with the diagonal sensitivity matrix Ω enables the system to control attention representation through feature-based adjustments. The system uses quantum-inspired technology for its phase modulation mechanism which operates entirely through classical neural computations without using quantum hardware or quantum state evolution. The cosine transformation functions as a nonlinear phase encoder that modulates attention weights through periodic changes which resemble the effects of phase interference. The diagonal matrix Ω controls feature-wise sensitivity while δ provides learnable phase offsets. The operation identifies stable and unstable behavioral interactions through its ability to strengthen coherent patterns and reduce the impact of random activations. The transformation creates bounded nonlinear modulation which protects numerical stability while maintaining differentiability for gradient-based optimization. Deep hybrid descriptors are created using phase-modulated attention output and residual representation.

$$y_r = e_r + \vartheta_r \odot e_r, \tag{14}$$

There is a connection between attention-based refining and hierarchical residual learning. A small decision space is used to hold the combined representation.

$$d_r = g(A_f y_r + c_f), \tag{15}$$

Subsequently, a layer for probabilistic classification

$$\hat{p}_r = \text{softmax}(A_o d_r + c_o), \tag{16}$$

Where p_r stands for the expected distribution of classes given levels of learning outcomes that have been set. To optimize models, a class-aware objective function is used.

$$J = -\frac{1}{m} \sum_{r=1}^{m} \sum_{k=1}^{K} \omega_k z_{rk} \log(\hat{p}_{rk}), \tag{17}$$

The ground-truth indicator for class k is z_{rk}, its weight is ω_k, and the number of result categories is K. This assures equitable learning throughout mobile learning's inadequate, adequate, and excellent achievement levels. Algorithm 2 explains QRAF-Net.

A deep hybrid learning outcome classifier for AI-enabled remote biology labs is QRAF-Net. The framework integrates residual feature propagation, attention-driven interaction modeling, and quantum-inspired modulation.

The proposed architecture achieves continuous transformations through its design since all transformations exist as differentiable mathematical functions. Residual connections preserve all information pathways while they reduce the occurrence of vanishing gradients through their design.

Algorithm 2 Quantum Residual Attentive Fusion Network for Learning Outcome Classification

Require: Refined session-level dataset $\mathcal{X} = \{s_1, s_2, \dots, s_m\}$, where $s_r \in \mathbb{R}^p$; number of classes K; residual depth T; latent size q; learning rate η; batch size B; class weights $\{\omega_1, \dots, \omega_K\}$

Ensure: Trained model parameters Θ and session-level predictions $\hat{p}$

Initialize model parameters $\Theta = \{A_0, c_0, \{A_t, c_t\}_{t=1}^{T}, L, R, \Omega, \delta, A_f, c_f, A_o, c_o\}$
for each training epoch **do**
  Shuffle $\mathcal{X}$ and form mini-batches $\{\mathcal{B}_1, \mathcal{B}_2, \dots\}$ with size B
  for each mini-batch $\mathcal{B}$ **do**
    for each session vector $s_r \in \mathcal{B}$ **do**
      Embedding: $e_r^{(0)} \leftarrow g(A_0 s_r + c_0)$
      for $t = 1$ to T **do**
        $e_r^{(t)} \leftarrow g(A_t e_r^{(t-1)} + c_t) + e_r^{(t-1)}$
      end for
      $e_r \leftarrow e_r^{(T)}$
      $\beta_r \leftarrow \text{softmax}(((Le_r) \odot (Re_r))/\sqrt{q})$
      $\vartheta_r \leftarrow \cos(\Omega\beta_r + \delta)$
      $y_r \leftarrow e_r + (\vartheta_r \odot e_r)$
      $d_r \leftarrow g(A_f y_r + c_f)$
      $\hat{p}_r \leftarrow \text{softmax}(A_o d_r + c_o)$
    end for
    $\mathcal{J}_{\mathcal{B}} \leftarrow -\frac{1}{|\mathcal{B}|} \sum_{s_r \in \mathcal{B}} \sum_{k=1}^{K} \omega_k z_{rk} \log(\hat{p}_{rk})$
    $\Theta \leftarrow \Theta - \eta \cdot \nabla_{\Theta}(\mathcal{J}_{\mathcal{B}})$
  end for
end for
return trained Θ and predicted probabilities $\{\hat{p}_r\}$

The quantum-inspired phase modulation uses a cosine mapping method which restricts its output range to the interval of [-1,1] thus guaranteeing mathematical stability while stopping excessive feature enhancement from occurring. The diagonal sensitivity matrix applies element-wise scaling, which maintains Lipschitz continuity throughout its input representation range. The combined objective function uses prediction errors together with regularization to drive stable solution development through various interactive data with different noise levels.

3.5 Parameter optimization using evolutionary-guided adaptive gradient optimization (EGAGO)

In order to make the model suitable for a variety of uses, we tune its parameters using an Evolutionary-Guided Adaptive Gradient solution. This hybrid technique uses extensive parameter exploration and concentrated local refinement to handle the Remote Biology Laboratory Interaction Dataset's diverse and imbalanced characteristics. Show the trainable parameters as

$$P = \{p_1, p_2, \dots, p_m\}. \tag{18}$$

The objective of tuning is to reduce a validation-based cost, which is defined as

$$J(P) = U(P) + \gamma \cdot Q(P), \tag{19}$$

$Q(P)$ regulates the complexity of the model, and $U(P)$ is the expected validation error on a dataset that was not used to train the model. The scalar γ maintains a balance between regularization and prediction accuracy. A set of potential configurations is considered during the evolutionary guidance phase.

$$G = \{P^1, P^2, \dots, P^L\} \tag{20}$$

is established with appropriate boundaries. To evaluate each contender, we compute their usefulness score.

$$S(P^l) = \frac{1}{J(P^l) + \delta}, \tag{21}$$

This maintains the stability of the assessment, with δ being a small constant. Candidates with consistently high utility are preserved as parameter space prospects. Validation-driven updates modify configurations during adaptive refinement:

$$P \leftarrow P - \kappa \cdot h(P), \tag{22}$$

The adaptive adjustment factor κ and the local sensitivity of the cost function $h(P)$ are defined here. Adjustment magnitude is reduced to prevent over-tuning when validation improvements become unreliable. These two procedures are performed until performance improves somewhat or an iteration budget is reached. Instead of performance peaks, consistent validation behavior determines parameter configuration. While learning in the actual world, this makes the model more robust to change.

3.6 Performance evaluation

To address class imbalance across learning result categories, the following metrics are used to assess the proposed learning outcome classification model: F1-score, recall, accuracy, and precision [28]. These tests assess the learning result categorization's right-wrong balance. Combining these baseline measurements with the task-specific Outcome Stability Index (OSI) shows how significant educational misclassification is. The ordinality of NILOS outcome states is considered by OSI, not correctness. Significant deviations from learning predictions are penalized more than minor deviations. The OSI is defined as:

$$\text{OSI} = 1 - \frac{1}{T} \sum_{i=1}^{T} \frac{|p_i - r_i|}{L - 1} \tag{23}$$

The sum of the actual and predicted learning outcome indices for the i-th session, denoted as p^i and r^i, respectively, and the total number of learning outcome levels, denoted as L, are all part of the sentence. Higher OSI levels indicate more reliable and relevant teaching predictions.

Algorithm 3 Evolutionary-Guided Adaptive Gradient Optimization (EGAGO)

Require: Trainable parameter space bounds $\mathcal{B}$; population size L; maximum evolutionary iterations T_e; maximum refinement iterations T_a; regularization weight γ; stability constant δ; initial adaptive step size κ; validation dataset $\mathcal{D}_{val}$
Ensure: Optimized parameter configuration P^*

Initialize candidate population $G = \{P^1, P^2, \ldots, P^L\}$ within bounds $\mathcal{B}$
Set best cost $J^* \leftarrow +\infty$ and best configuration $P^* \leftarrow \emptyset$
for $t = 1$ to T_e **do**
  for each candidate $P^l \in G$ **do**
    Compute validation error $U(P^l)$ on $\mathcal{D}_{val}$
    Compute regularization term $Q(P^l)$
    Evaluate objective: $J(P^l) = U(P^l) + \gamma \cdot Q(P^l)$
    Compute utility score: $S(P^l) = \frac{1}{J(P^l)+\delta}$
    if $J(P^l) < J^*$ **then**
      $J^* \leftarrow J(P^l)$
      $P^* \leftarrow P^l$
    end if
  end for
  Rank candidates in descending order of $S(P^l)$
  Preserve top-performing candidates as elite parameter prospects
  Generate updated candidates for the remaining population using bounded evolutionary perturbation around elite solutions
end for
Initialize adaptive refinement from the best candidate: $P \leftarrow P^*$
for $r = 1$ to T_a **do**
  Compute local sensitivity / gradient-like update direction $h(P)$
  Update parameters: $P \leftarrow P - \kappa \cdot h(P)$
  Project P back into feasible bounds $\mathcal{B}$ if needed
  Compute updated objective $J(P)$ on $\mathcal{D}_{val}$
  if $J(P) < J^*$ **then**
    $J^* \leftarrow J(P)$
    $P^* \leftarrow P$
  else
    Reduce adaptive step size κ to avoid over-tuning
  end if
  if validation improvement is smaller than a predefined tolerance **then**
    break
  end if
end for
return P^*

Comparison with ordinal-aware metrics The NILOS taxonomy establishes learning outcome states through its ordinal structure, which defines less critical misclassifications between adjacent states than between distant ones. The assessment of accuracy and F1-score through traditional classification metrics fails to capture this graded severity. The existing ordinal-aware measures which include weighted Cohen's kappa and Spearman rank correlation and Kendall's tau show partial ordering capacity because these measures primarily assess either agreement or rank association while their design leads to prediction stability evaluation across sessions. The mean absolute error on ordinal labels shows how much values differ from each other yet it fails to recognize how values are distributed throughout the dataset.

The proposed OSI complements these measures by quantifying both correctness and stability of predictions across ordered outcome states. The OSI function assesses how the model maintains learner progression patterns through different conditions, while rank-based statistics fail to provide this capability. This makes it particularly suitable for session-level learning analytics in remote laboratory environments, where fluctuations in interaction quality may affect prediction reliability. The OSI metric offers a complete assessment of ordinal learning outcome prediction because it exceeds the evaluation capacity of standard ordinal-aware metrics.

4 Simulation results and discussion

This section presents simulation results and comparative analysis and robustness evaluation together with their interpretive discussion to assess how well the proposed framework functions in actual mobile laboratory testing environments. Experiments employed 254,020 session-level records from the Remote Biology Laboratory Interaction Dataset, with 80% for training and 20% for testing. We preprocessed all session records before model learning using robust scaling, median-based imputation, and dependency-driven feature refinement. A federated learning scenario with six decentralized clients trained local models on non-identically distributed data and aggregated them using a conventional federated averaging method to assess the proposed framework. The dataset was divided into six separate parts to create a decentralized mobile laboratory deployment simulation, which required each client to receive data from different session distribution patterns that simulated non-identical distribution conditions. Local training of models happened through multiple independent training periods before the models executed their first synchronization. The model parameters were transmitted between systems at specific intervals during communication rounds, while the Federated Averaging (FedAvg) algorithm executed global model aggregation. The system protected data privacy by not allowing clients to share their original data while maintaining their data storage rights. The system stored a distinct test set at the server which allowed testing of the global model after every aggregation period. Hyperparameter tuning used validation subsets which technicians created from local training data to prevent any information leakage. This setup approximates to actual conditions of distributed learning which occur in mobile educational platforms.

The QRAF-Net model was trained using mini-batch gradient optimization with a batch size of 64 and an initial learning rate of 0.001. The training is conducted for 200 epochs but used early stopping when validation performance reached its maximum after multiple consecutive evaluations. The model used class-weighted cross-entropy loss function to solve the problem of NILOS category imbalance. The EGAGO optimization strategy from Section 3.5 was used to perform parameter optimization because it combines evolutionary exploration with adaptive gradient refinement. The model parameters were initialized with standard random initialization and updated through backpropagation. The baseline models were trained on identical training-testing splits while using the same preprocessing pipeline used for evaluation.

The residual depth, attention dimensionality, learning rate, batch size, and regularization coefficients were chosen using stability-oriented tuning to guarantee convergence and robustness. A workstation with an NVIDIA GeForce

RTX 4060 GPU and 32 GB RAM ran simulations. Quantitative performance results, baseline model comparison, robustness assessment, and interpretive discussion follow.

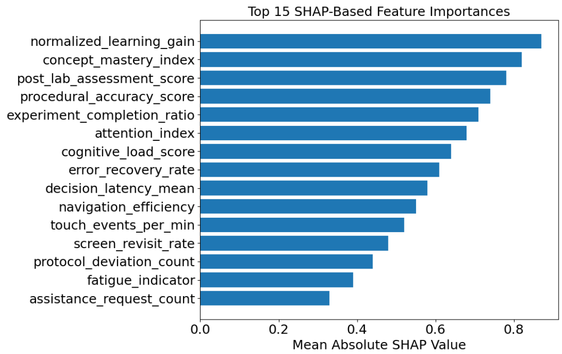

Figure 3: SHAP-based feature contribution analysis

The mean absolute SHAP scores rank the top 15 assessment and behavioral attributes, as shown in Figure 3. The results show that learning gain and concept mastery and assessment consistency have the greatest impact on prediction results. The explainability analysis shows that QRAF-Net delivers behavior-based NILOS-based learning outcome predictions which are transparent to users. SHAP analysis shows that learning gain and concept mastery and procedural consistency and assessment reliability serve as the main factors which determine outcome states. These features follow a direct relationship with significant educational markers which show student progress in laboratory learning environments. Students demonstrate academic growth through learning gain which measures their progress throughout the entire session while students show their capacity to conduct experiments through procedural consistency which assesses their ability to follow experimental protocols correctly and continuously. Assessment reliability provides evidence of consistent performance across different evaluation parts which helps assessors identify students who have reached intermediate versus advanced proficiency levels. The model depends on interpretable educational signals to establish its validity as an instructional decision-making tool for remote laboratory sites because the model shows more reliance on these signals than on misleading connections.

Figure 4 shows QRAF-Net's learning process time-series dual-axis graph. Consistent training and testing accuracy increases and loss reductions indicate sustained optimization. The apparent convergence point at period 170, stable performance, and a low generalization gap indicate residual-attention fusion. Even with genuine behavioral noise, learning is strong and overfitting is controlled.

The NILOS prediction confusion matrix is shown in Figure 5 for the 20% remaining test split of 50,804 sessions. Diagonal inputs suggest class separation with 97% accuracy. Most residual mistakes are from nearby learning phases. This validates NILOS categories' ordinality and shows that learner growth involves boundary ambiguity.

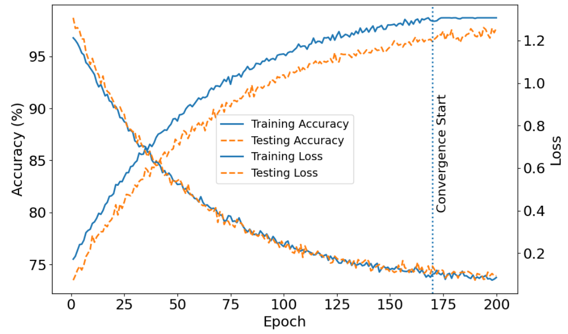

Figure 4: Training and testing convergence behavior

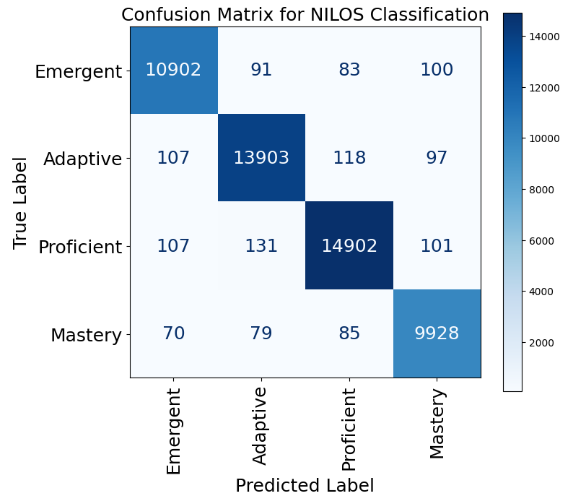

Figure 5: Confusion matrix for NILOS classification

Figure 6 depicts the way QRAF-Net performs when used with conventional evaluation metrics and with OSI or the lack of them. The pipeline does not fail to gain great dividends from the use of residual-attention fusion and stability-aware modeling. Such an examination indeed shows how modular architecture can cause the model to improve in correctness and stability.

Table 2 displays the results which show how QRAF-Net performs compared to conventional deep learning models and transformer-based systems and hybrid systems through both standard evaluation metrics and OSI assessment. The system shows its highest accuracy through QRAF-Net which reaches the highest accuracy value in all evaluation categories compared to all remaining systems. The observed performance difference will occur because multiple factors operate together to produce this outcome. The Remote Biology Laboratory Interaction Dataset contains extensive session-level data which enables researchers to create reliable models that show stable behavioral patterns while minimizing the impact of brief interactions. The NILOS taxonomy establishes an ordered framework which enables the model to use structured learning paths between academic levels instead of treating classes as distinct entities. The preprocessing pipeline improves sig-

Table 2: Performance comparison with baseline models

Model (Abbrev.)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	OSI
MHA-DL [7]	89.4	88.9	87.6	88.2	0.865
HTN [8]	90.1	89.7	88.4	89.0	0.872
AE-DKT [9]	88.6	87.9	86.8	87.3	0.858
UniKT [10]	90.8	90.1	89.2	89.6	0.879
GNN-TF [11]	89.9	89.4	88.1	88.7	0.870
MV-HGNN [12]	88.1	87.5	86.2	86.8	0.852
MD-DL [13]	87.6	86.8	85.9	86.3	0.846
CA-MDM [14]	90.3	89.8	88.9	89.3	0.876
CNN-TF-MO [15]	91.2	90.6	89.8	90.2	0.884
RL-TCN [16]	88.9	88.2	87.1	87.6	0.861
QA-SA [20]	89.7	89.1	88.0	88.5	0.869
TG-TF [21]	91.0	90.4	89.6	90.0	0.882
Proposed Model	97.8	97.3	97.1	97.2	0.956

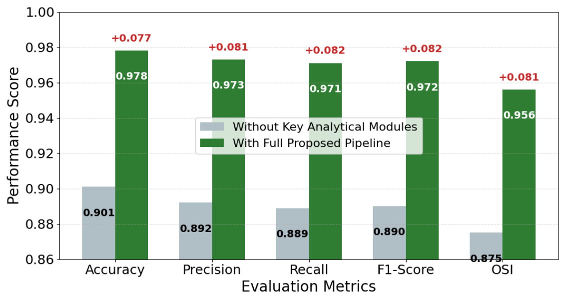

Figure 6: Impact of quantum and explainable intelligence modules on model performance

nal quality through median-based normalization and redundancy elimination and behavioral signal decomposition which generates better training data for model development. The Outcome Stability Index (OSI) tests prediction accuracy through its assessment of prediction consistency which creates a more demanding standard to evaluate performance. The robustness tests show how performance decreases when network latency and interaction loss and behavioral disturbance happens which reveals how the model performs in real-world situations while proving that its accuracy is not overstated.

A detailed analysis of the contributions of each component of QRAF-Net is shown in Table 3 of the ablation study. Combining preprocessing with feature refinement, behavioral decomposition, and hybrid parameter tuning results in continuously improving performance. The optimal accuracy and OSI in the final configuration demonstrate the critical roles played by each module in enhancing prediction performance and maintaining consistent NILOS classification results.

In comparison to baseline models, QRAF-Net performs better in terms of size, training and inference costs, memory use, and accuracy (as shown in Table 4). With the same amount of parameters and runtime overhead, QRAF-Net achieves optimal prediction performance. This demonstrates that the proposed residual-attention method outperforms the alternatives without overloaded the computer or memory.

Table 5 presents the results obtained from five-fold ANOVA and paired t-test and Wilcoxon tests which assessed QRAF-Net performance against baseline models. The proposed approach demonstrates a substantial and statistically significant improvement in NILOS-based learning outcome evaluation which achieved the highest Pearson correlation and largest effect size and all methods produced highly significant p-values. The researchers used five-fold cross-validation to examine statistical significance through a method which divided the data into five distinct sections that matched in size. The training used four subsets in each iteration while the testing used one subset which prevented any session from appearing in both subsets at the same time. The statistical analysis used performance metrics which were obtained across folds. The study employed one-way ANOVA for parametric analysis and subsequently performed paired two-tailed t-tests while the Wilcoxon signed-rank test served as non-parametric validation to handle data that might violate normal distribution. The researchers established a significance level of $\alpha = 0.05$ for their study. The researchers used Cohen's d to calculate effect sizes while they employed Pearson correlation coefficients to measure consistency between different folds. The study did not apply multiple-comparison correction because all comparisons were made under the same experimental conditions which tested the proposed method.

Figure 7 shows NILOS classification ROC curves for QRAF-Net vs deep and transformer-based baselines. QRAF-Net has the greatest AUC due to its ideal true positive and false positive rates. The prolonged divergence from baseline curves shows the model's enhanced ability to recognize learning phases in true class imbalances.

QRAF-Net's training stability is assessed using the suggested Outcome Stability Index (OSI) and Validation Consistency Score is shown in Figure 8. Stability rises as behavioral noise reduces, reaching equilibrium at epoch 170. The model's ability to assess learners' consistency and performance robustness shows its originality.

Table 3: Comprehensive ablation analysis

Variant	Components	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	OSI
A1	Base classifier (session-level tabular features only)	90.6	90.0	89.1	89.5	0.874
A2	A1 + Robust scaling + median imputation	92.4	92.0	91.2	91.6	0.892
A3	A2 + Redundancy pruning (dependency-based feature refinement)	94.1	93.7	93.0	93.3	0.910
A4	A3 + Behavioral signal decomposition descriptors	96.2	95.8	95.3	95.5	0.935
A5	A4 + Hybrid parameter tuning (EGAGO)	97.8	97.3	97.1	97.2	0.956

Table 4: Comparison of model complexity and computational efficiency

Model	Params (M)	Train Time / Epoch (s)	Inference Time (ms)	Memory Usage (MB)	Accuracy (%)
MHA-DL [7]	10.6	1.74	13.2	398	89.4
HTN [8]	11.2	1.88	14.5	415	90.1
AE-DKT [9]	8.9	1.46	11.6	352	88.6
UniKT [10]	10.1	1.62	12.4	381	90.8
GNN-TF [11]	12.8	2.05	15.8	452	89.9
MV-HGNN [12]	13.4	2.21	16.9	468	88.1
MD-DL [13]	9.7	1.58	12.1	365	87.6
CA-MDM [14]	11.6	1.91	14.3	421	90.3
CNN-TF-MO [15]	12.3	2.08	15.2	440	91.2
RL-TCN [16]	10.9	1.83	13.7	404	88.9
QA-SA [20]	11.1	1.87	14.0	412	89.7
TG-TF [21]	12.0	1.99	14.9	430	91.0
Proposed Model	11.7	1.76	12.6	418	97.8

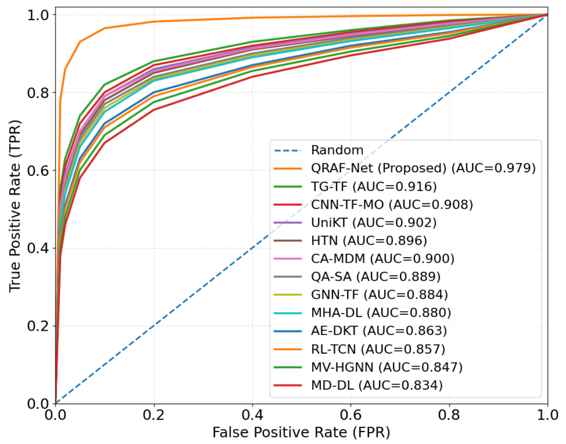

Figure 7: ROC–AUC comparison across all models

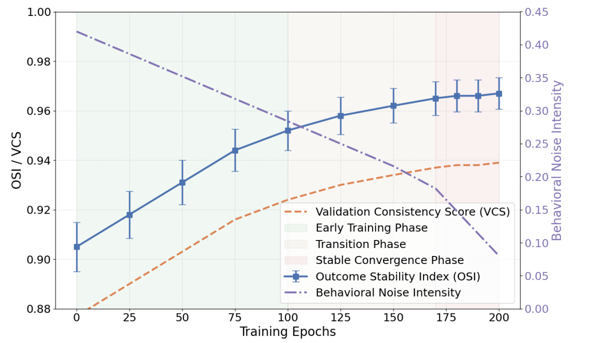

Figure 8: Context-window sensitivity and stability analysis

The results of QRAF-Net testing under real-world disturbance conditions which involved missing interaction logs and behavioral noise and assessment inconsistencies and network latency issues are shown in Table 6. The model shows its ability to withstand non-standard operational situations by maintaining accurate results which compute high OSI values even though performance drops when disturbance levels rise. The study used synthetic disturbances to test how robust system performance functions under real-world conditions to verify system robustness. The research team created two test conditions which included network latency and session interruptions by making interaction sequences to lose their temporal continuity through random sequence delays and sequence truncation. The researchers achieved partial interaction log loss by implementing a random removal process which deleted 5 to 20 percent of session events to simulate data transmission failures or incomplete data recording. The research team created behavioral noise by using irregular actions and out-of-pattern interaction samples which they took from low-probability behavioral distributions to show the behavior of learners who become distracted or show inconsistent learning patterns. The study created assessment inconsistency through a random process which changed the submission timing and evaluation-related attributes to show assessment processes which occurred at delayed or irregular times. The researchers created three disturbance intensity levels (low medium high) to study how performance degradation changes with increased affected session components in the study.

Modifying key QRAF-Net hyperparameters has an effect

Table 5: Statistical evaluation of 5-Fold CV for the proposed mobile learning quality evaluation framework

Method	ANOVA (p-value)	Paired t-test (p-value)	Wilcoxon Test (p-value)	Pearson r	Effect Size (Cohen's d)
MHA-DL [7]	3.4×10^{-3}	4.9×10^{-3}	5.6×10^{-3}	0.82	0.76
HTN [8]	2.9×10^{-3}	4.1×10^{-3}	4.9×10^{-3}	0.83	0.80
AE-DKT [9]	3.7×10^{-3}	5.2×10^{-3}	6.0×10^{-3}	0.81	0.72
UniKT [10]	2.5×10^{-3}	3.6×10^{-3}	4.3×10^{-3}	0.84	0.85
GNN-TF [11]	2.8×10^{-3}	3.9×10^{-3}	4.7×10^{-3}	0.83	0.82
MV-HGNN [12]	4.1×10^{-3}	5.8×10^{-3}	6.5×10^{-3}	0.80	0.69
MD-DL [13]	4.5×10^{-3}	6.2×10^{-3}	6.9×10^{-3}	0.79	0.66
CA-MDM [14]	2.6×10^{-3}	3.8×10^{-3}	4.5×10^{-3}	0.84	0.83
CNN-TF-MO [15]	2.1×10^{-3}	3.1×10^{-3}	3.9×10^{-3}	0.86	0.91
RL-TCN [16]	3.2×10^{-3}	4.6×10^{-3}	5.3×10^{-3}	0.82	0.75
QA-SA [20]	2.9×10^{-3}	4.0×10^{-3}	4.8×10^{-3}	0.83	0.81
TG-TF [21]	2.0×10^{-3}	2.9×10^{-3}	3.6×10^{-3}	0.87	0.94
Proposed Model	$< 1 \times 10^{-4}$	$< 1 \times 10^{-4}$	$< 1 \times 10^{-4}$	0.97	1.34

Table 6: Robustness analysis under practical disturbances in remote biology laboratory sessions

Disturbance Scenario	Level	FP (%)	FN (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	OSI
Reference (clean sessions)	–	1.1	0.9	97.8	97.3	97.1	97.2	0.956
Network latency and session interruptions	Low	1.6	1.4	97.1	96.8	96.5	96.6	0.948
	Medium	2.9	2.4	96.0	95.6	95.2	95.4	0.935
	High	5.2	4.1	94.3	94.0	93.6	93.8	0.912
Partial interaction log loss (5–20%)	Low	1.8	1.6	97.0	96.6	96.4	96.5	0.946
	Medium	3.2	2.8	95.9	95.5	95.1	95.3	0.933
	High	5.6	4.9	94.1	93.7	93.4	93.5	0.909
Behavioral noise (irregular actions)	Low	1.9	1.5	97.0	96.7	96.5	96.6	0.947
	Medium	3.5	3.0	95.8	95.4	95.0	95.2	0.931
	High	6.1	5.0	93.8	93.4	93.1	93.2	0.907
Assessment inconsistency (delayed submissions)	Low	2.0	1.7	96.9	96.6	96.3	96.4	0.944
	Medium	3.8	3.2	95.6	95.2	94.9	95.0	0.928
	High	6.4	5.4	93.5	93.1	92.8	92.9	0.903

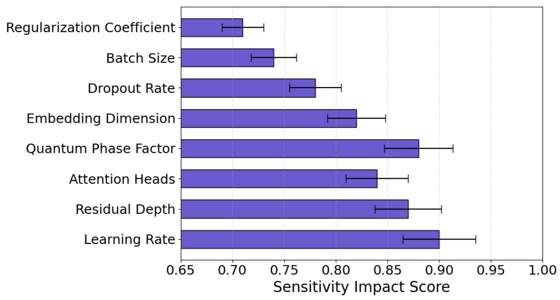

Figure 9: Sensitivity analysis of model hyperparameters

on the model's performance as shown in Figure 9. The findings indicate that regularization and batch size have lesser impacts compared to the learning rate, residual depth, and quantum phase factor. The stable model can withstand modest changes in parameters and perform well in real-world learning conditions, as seen by the narrow error bars.

The evaluation tests the proposed framework's universal applicability by using different disturbance conditions and distinct session characteristics that simulate actual mobile laboratory operations. The model maintains its operational effectiveness through all tested conditions which include different interaction densities and irregular behavioral patterns and incomplete session detection. The framework demonstrates its ability to function effectively in complex dynamic learning settings which require more flexibility than controlled experimental environments.

4.1 Discussion

The discussion analyzes performance trends through experimental results from the previous section while showing how architectural components affect system performance and demonstrating the real-world value of the proposed framework. The model outperforms existing deep learning and transformer-based methods by demonstrating better stability under the different user interactions and mobile remote laboratory system usage patterns that exist in actual testing environments. The model achieves its performance through residual-attention modeling, which enables long-range dependency detection while preserving essential local behavioral information. The system achieves better representation learning because its dependency-based feature refinement process eliminates noise and redundant information found in interaction logs. The proposed NILOS taxonomy establishes a systematic method to evaluate learner achievement, which enables educational institutions to assess student performance through detailed session outcomes instead of basic performance ratings.

The advantages come with multiple trade-offs that need to be evaluated. The model requires careful preprocessing and parameter tuning, which increases computational demands beyond what simpler baseline models require. The federated evaluation demonstrates system resilience against non-identically distributed data patterns; however, real-world systems need to handle extra limitations that emerge from restricted connectivity and diverse device environments. The results show that learning gain and procedural consistency together with engagement-related features predict outcomes because explainability analysis confirms their superior predictive ability. The finding supports learning analytics theories, which state that continuous interaction quality between students and learning systems leads to better learning outcomes than using limited performance metrics.

The study demonstrates that residual-attention learning provides an effective method to predict outcomes in mobile remote laboratory environments which experience challenges from noisy data and incomplete datasets and diverse user behavior patterns.

4.2 Computational scalability and deployment in low-resource settings

Mobile remote laboratories are frequently deployed in resource-constrained educational environments, including low-power mobile devices, unstable network connectivity, and limited computational infrastructure. The researchers conducted their study to evaluate both the operational capacity of their framework and its predictive ability. QRAF-Net uses its residual-attention architecture to process session-level feature representations which require less computing resources and memory capacity than handling high-frequency data streams. The system manages session throughput since its inference time increases with each added session, making it usable on devices with low processing power.

The model functions properly in settings that have no internet access because it can perform inference without needing constant online connectivity. The processing system uses efficient methods to transform interaction logs into small behavioral descriptors which require less storage space and transmission bandwidth. The system achieves lower computational needs because it does not process heavy multimodal data or perform graph-building procedures which many transformer-based systems and graph-based learning analytic systems require. The model provides resource-limited educational institutions with a solution that functions in environments without access to advanced computing resources. The framework can be integrated into mobile laboratory platforms or institutional learning systems with limited hardware acceleration, while maintaining reliable outcome prediction performance.

5 Managerial and educational implications of NILOS-based learning outcome prediction

The NILOS-based NILOS-based learning outcome prediction approach has implications for the management and administration of AI-enabled remote laboratory platforms. By categorizing student performance as Emergent, Adaptive, Proficient, or Mastery, instructors and academic coordinators may employ data-driven intervention strategies instead of aggregate grading. Emergent and adaptive students may get early pedagogical support and guided scaffolding instead of remediation, improving instructional efficiency and resource allocation. Institutional and platform management benefit from the Outcome Stability Index (OSI), which evaluates learning reliability across large-scale remote education sessions. OSI trends may help administrations track laboratory delivery consistency, identify systemic issues like unstable interaction patterns or assessment delays, and analyze curriculum changes. The model's tolerance to network disruptions and behavioral noise supports its usage in technologically flawed environments. The framework's explainability and SHAP-based feature attribution make AI-assisted educational decision-making more reliable. Knowing why a learner is classified into an end state is as important as prediction accuracy for ethical and regulatory agencies. NILOS-based prediction supports instructional and administrative decision-making for scalable, interpretable, and resilient digital education learning analytics.

6 Conclusion and future work

The research employed extensive mobile session data from the two laboratory experiments to forecast educational results through the NILOS taxonomy. The system combines effective data preprocessing with behavioral feature development and residual-attention mechanism and stability assessment methods to process diverse and contaminated real-world interaction patterns. The method demonstrated superior performance over existing leading approaches across all evaluation metrics by achieving better accuracy and precision and recall and F1-score results while the Outcome Stability Index (OSI) demonstrated consistent results under various disruptive conditions to test reliability, which included noise and partial log loss and session interruptions. The system delivers precise and understandable and dependable outcome forecasts which enable continuous learner assessment and personalized teaching assistance and assessment of educational quality for remote laboratory operations in both standard and resource-limited environments.

Future work will use longitudinal modeling to study how learners progress through several laboratory sessions to understand their long-term developmental changes. The research will investigate how mobile systems can achieve cross-domain functionality through brief on-device modifications which will enable extensive mobile system usage.

Data availability

The dataset is publicly accessible through the Zenodo repository and can be retrieved using https://zenodo.org/records/18147010

Funding

1. 2026 Northwestern Polytechnical University Education and Teaching Reform Research Project, Project Number: 2026JGZ77, Project Title: Exploration on the Interdisciplinary Characteristic Innovative Talent Training Mode of "Artificial Intelligence + Ecology" for the Future.

2. 2025 Northwestern Polytechnical University Education and Teaching Reform Research Project, Project Number: 2025JGZ55, Project Title: Research and Practice on the Cultivation System and Pathway for "Chief-Engineer-Type" Top Innovative Talents in Ecology Top Student Program.

References

[1] Leong W. Y. (2025) Revolutionizing lab-based research in engineering with immersive virtual reality technologies, *Proceedings of the International Conference on Intelligent Technology for Educational Applications*, Springer Nature Singapore, pp. 223–232. https://doi.org/10.1007/978-981-95-2011-4_21.

[2] Sheng Y., Wang J., Liu S., and Jiang Y. (2024) IMN4NPD: An Integrated Molecular Networking Workflow for Natural Product Dereplication, *Analytical Chemistry*, vol. 96, no. 7, pp. 2990–2997. https://doi.org/10.1021/acs.analchem.3c04746.

[3] Lam P. X., Mai P. Q. H., Nguyen Q. H., Pham T., Nguyen T. H. H., and Nguyen T. H. (2024) Enhancing educational evaluation through predictive student assessment modeling, *Computers and Education: Artificial Intelligence*, vol. 6, Art. no. 100244. https://doi.org/10.1016/j.caeai.2024.100244.

[4] Ji Y., Liu Y., Yao F., He M., Tao S., Zhao X., and Sun Y. (2025) Adapting large language models to log analysis with interpretable domain knowledge, *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 1135–1144. https://doi.org/10.1145/3746252.3761189.

[5] Mousa M. (2025) AI-supported formative assessments: Enhancing student-centered learning and teacher perceptions, *Journal of Pedagogy and Education Science*, vol. 4, no. 2, pp. 127–141. https://doi.org/10.56741/jpes.v4i02.857.

[6] Wongvorachan T., Bulut O., Liu J. X., and Mazzullo E. (2024) A comparison of bias mitigation techniques for educational classification tasks using supervised machine learning, *Information*, vol. 15, no. 6, Art. no. 326. https://doi.org/10.3390/info15060326.

[7] Kumar G. B., Shabbeer S. M., Kumar Y. R., and Durga Prasad S. V. (2025) A predictive approach to student performance in online learning using data analytics, *Journal of Computational Analysis and Applications*, vol. 34, no. 6, pp. 1121–1135. https://doi.org/10.48047/jocaaa.2025.34.06.14.

[8] Xu J. and Hu W. (2024) An enhanced deep knowledge tracing model via multiband attention and quantized question embedding, *Applied Sciences*, vol. 14, no. 8, Art. no. 3425. https://doi.org/10.3390/app14083425.

[9] Zhan Z., Shen T., and Wang J. (2025) Predicting student performance based on knowledge characteristics and learning ability, *IEEE Access*, vol. 13, Art. no. 3538700. https://doi.org/10.1109/ACCESS.2025.3538700.

[10] Wang T., Li J., Wu H., Li C., Snoussi H., and Wu Y. (2022) ResLNet: deep residual LSTM network with longer input for action recognition, *Frontiers of Computer Science*, vol. 16, no. 6, p. 166334. https://doi.org/10.1007/s11704-021-0236-9.

[11] Balachandar V. and Venkatesh K. (2025) A multi-dimensional student performance prediction model (MSPP): An advanced framework for accurate academic classification and analysis, *MethodsX*, vol. 14, Art. no. 103148. https://doi.org/10.1016/j.mex.2025.103148.

[12] Campilongo G. B., Tonzar-Santos G., Verginio M. E. D. S., and Lellis-Santos C. (2025) Smartphone-assisted experimentation as a medium of understanding human biology through inquiry-based learning, *Education Sciences*, vol. 15, no. 8, Art. no. 1005. https://doi.org/10.3390/educsci15081005.

[13] Gong Y., Wang F., Zhang Y., and Geng J. (2025) TGEL-transformer: Fusing educational theories with deep learning for interpretable student performance prediction, *PLOS ONE*, vol. 20, no. 6, Art. no. e0327481. https://doi.org/10.1371/journal.pone.0327481.

[14] Wang S. and Luo B. (2024) Academic achievement prediction in higher education through interpretable modeling, *PLOS ONE*, vol. 19, no. 9, Art. no. e0309838. https://doi.org/10.1371/journal.pone.0309838.

[15] Luo Y., Zhu M., Chen T., and Zheng Z. (2025) Remaining useful life prediction for stratospheric airships based on a channel and temporal attention network, *Communications in Nonlinear Science and Numerical Simulation*, vol. 143, p. 108634. https://doi.org/10.1016/j.cnsns.2025.108634.

[16] Jin J., Zhu J., Zhou L., Chen C., Wu L., Lu M., and Li Z. (2024) A Complex-Valued Variant-Parameter Robust Zeroing Neural Network Model and its Applications, *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 2, pp. 1303–1321. https://doi.org/10.1109/TETCI.2024.3356163.

[17] Bautista López, J., and Ruiz-Velasco Sánchez, E. (2026) Remote Laboratory Based on FPGA Devices Using the E-Learning Approach, *Applied System Innovation*, vol. 9, no. 2, p. 37. https://doi.org/10.3390/asi9020037.

[18] Nikolic S., Suesse T. F., Grundy S., Haque R., Lyden S., Lal S., et al. (2025) Assessment integrity and validity in the teaching laboratory: Adapting to GenAI by developing an understanding of the verifiable learning objectives behind laboratory assessment selection, *European Journal of Engineering Education*, vol. 50, no. 4, pp. 673–701. https://doi.org/10.1080/03043797.2025.2456944.

[19] Zhang Y., Stöhr C., Jämsvi S. S., Kabo J., and Malmqvist J. (2025) Student Experiences of Hybrid and Online Engineering Labs in a Logic Control Course, *Computer Applications in Engineering Education*, vol. 33, no. 3, Art. no. e70032. https://doi.org/10.1002/cae.70032.

[20] Hashi A. O., Mohd Hashim S. Z., Mirjalili S., Kebande V. R., Al-Dhaqm A., Nasser M., and Samah A. B. A. (2025) A hybrid CNN–transformer framework optimized by grey wolf algorithm for accurate sign language recognition, *Scientific Reports*, vol. 15, no. 1, Art. no. 43550. https://doi.org/10.1038/s41598-025-27390-2.

[21] Almujally N. A., Rafique A. A., Al Mudawi N., Alazeb A., Alonazi M., Algarni A., Jalal A., and Liu H. (2024) Multi-modal remote perception learning for object sensory data, *Frontiers in Neurorobotics*, vol. 18, p. 1427786. https://doi.org/10.3389/fnbot.2024.1427786.

[22] National University Health System (2025) Remote Biology Laboratory Interaction Dataset, *Zenodo*. Available at: https://zenodo.org/records/18147010.

[23] Xiang J., Zhang J., Zheng R., Li X., and Li M. (2021) NIDM: network impulsive dynamics on multiplex biological network for disease-gene prediction, *Briefings in Bioinformatics*, vol. 22, no. 5, p. bbab080. https://doi.org/10.1093/bib/bbab080.

[24] Liu H., Bing P., Zhang M., Tian G., Ma J., Li H., Bao M., He K., He J., He B., and Yang J. (2023) MNNMDA: Predicting human microbe-disease association via a method to minimize matrix nuclear norm, *Computational and Structural Biotechnology Journal*, vol. 21, pp. 1414–1423. https://doi.org/10.1016/j.csbj.2022.12.053.

[25] Li X. Y., Xiang J., Wu F. X., and Li M. (2022) NetAUC: A network-based multi-biomarker identification method by AUC optimization, *Methods*, vol. 198, pp. 56–64. https://doi.org/10.1016/j.ymeth.2021.08.001.

[26] Li Z., Cai J., Chen Q., Chen L., Qing M., and Yang S. X. (2025) An LSTM Network With Neural Plasticity for Driver Fatigue Recognition on Real Roads, *IEEE Transactions on Industrial Electronics*, pp. 1–9. https://doi.org/10.1109/TIE.2025.3585046.

[27] Ding X., Xu Y., Li G., Yang K., Yuan J., and An J. (2025) Design and Performance Evaluation for BILCM-ID System With Improved Stopping Criterion, *IEEE Transactions on Vehicular Technology*, vol. 74, no. 4, pp. 6779–6784. https://doi.org/10.1109/TVT.2024.3519540.

[28] de la Cruz Huayanay A., Bazán J. L., and Russo C. M. (2025) Performance of evaluation metrics for classification in imbalanced data, *Computational Statistics*, vol. 40, no. 3, pp. 1447–1473. https://doi.org/10.1007/s00180-024-01539-5.