Received: February 23, 2026. Revised: March 21, 2026. 682

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Temporal Models for Early At-risk Student Detection: A Daily-sequence Framework with Multi-horizon Evaluation

Bruri Trya Sartana[1]* Safitri Juanita[1] Utomo Budiyanto[1] Mauridhi Hery Purnomo[2]

[1]*Faculty of Information Technology, Universitas Budi Luhur, Indonesia*
[2]*Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Indonesia*
* Corresponding author's Email: brury@budiluhur.ac.id

Abstract: Early warning systems in online learning often rely on temporally aggregated representations such as weekly summaries, which obscure short-term engagement dynamics preceding academic risk. This study reframes early detection as a fine-grained temporal modeling problem by introducing a standardized daily-sequence framework constructed from fixed 150-day learner trajectories. Each student is represented using six normalized daily behavioral features. Six sequence architectures, including Transformer (TRF), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Temporal Convolutional Network (TCN), and One-Dimensional Convolutional Neural Network (CNN1D), are evaluated under identical stratified five-fold cross-validation. Receiver Operating Characteristic Area Under the Curve (ROC-AUC) and Precision–Recall Area Under the Curve (PR-AUC) serve as ranking metrics. At the full horizon, the Transformer achieves the strongest ranking performance (ROC-AUC: 0.891; PR-AUC: 0.912). Multi-horizon evaluation using truncated windows of 28, 60, 90, and 150 days shows monotonic performance gains as temporal evidence accumulates, highlighting the importance of preserving daily temporal structure for reliable early risk detection.

Keywords: Early warning system, Temporal models, Daily sequence modeling, Multi-horizon evaluation, Transformer.

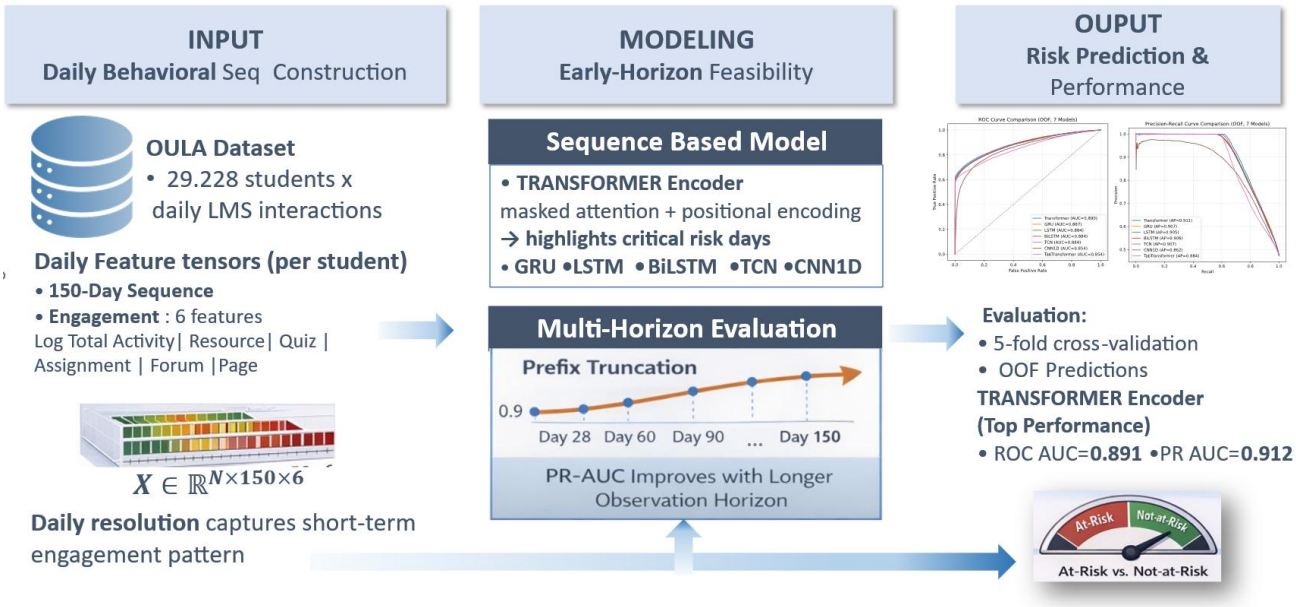

Figure. 1 Graphical Abstract

1. Introduction

Online learning platforms generate extensive fine-grained behavioral data that reflect student engagement and learning trajectories throughout the semester. Such data create opportunities to develop early-warning systems that detect struggling learners before their performance deteriorates significantly [1-4]. The shift toward online and hybrid learning has further amplified the need for accurate, interpretable predictive models that operate early in the academic year [3].

Existing predictive systems commonly rely on weekly or end-of-module aggregated features [6, 7]. Although these features are easy to compute, they often conceal important short-term variations in students' learning effort and interaction patterns. Several studies report that daily or ultra-fine engagement patterns are strongly associated with performance outcomes [8-10]. For instance, [8] demonstrated that micro-level clickstream patterns during video learning sessions predict assessment outcomes with greater reliability than weekly aggregates. While prior studies have emphasized the importance of temporal modeling, they have given limited attention to the distinction between representation-level and architecture-level improvements. It remains unclear whether performance gains arise primarily from sophisticated temporal architectures or from preserving fine-grained daily behavioral dynamics. This distinction is critical, as aggregated representations may compress engagement trajectories and obscure short-term fluctuations that precede withdrawal.

Temporal modeling has therefore become increasingly relevant. Deep sequence models such as LSTM, GRU, and Transformer architectures have been shown to model dependency structures and capture evolving engagement patterns over time [12-14]. Attention-based models, particularly Transformers, offer the additional advantage of modeling long-range dependencies and focusing on influential time steps [15, 16]. However, despite methodological advances, only a limited number of studies have evaluated daily-level sequences extending across the full duration of a course. Many adopt weekly summaries or segment sequences into limited windows, potentially reducing temporal sensitivity [17, 18].

To address this limitation, this study develops a daily-aligned sequence framework for early detection of at-risk students and evaluates it on a controlled cross-architecture benchmark. Student activity is standardized into fixed daily trajectories, allowing the effect of temporal representation to be examined under identical experimental conditions across multiple sequence models. A multi-horizon evaluation (28, 60, 90, and 150 days) further simulates progressive early-warning deployment and reveals how predictive signals accumulate over time. This design allows a direct comparison between sequence-based and aggregated representations for early risk prediction. Rather than proposing a new neural architecture, this study focuses on understanding the role of temporal representation in early warning systems through a controlled multi-model benchmark.

The main contributions of this study are summarized as follows:

Construction of aligned daily behavioral sequences from Virtual Learning Environment (VLE) logs to preserve engagement dynamics.

A controlled benchmark comparing multiple sequence architectures under identical experimental settings.

A multi-horizon early-warning protocol (28, 60, 90, and 150 days) to examine the accumulation of predictive signals over time.

Empirical evidence that fine-grained temporal representation improves early risk detection while maintaining consistent performance across architectures.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 describes the methodology, including daily-sequence construction and the multi-horizon protocol. Section 4 presents experimental results and statistical comparisons. Section 5 concludes the paper.

2. Related works

Research on learning analytics shows that temporal information is valuable for identifying struggling students. Studies focusing on group problem-solving logs [15], video clickstream interaction, and fine-grained digital engagement [16] [17] demonstrate that behavior evolves in time and that these dynamics influence learning performance. Weekly summaries often capture overall engagement levels but miss short-term fluctuations that can signal confusion, low motivation, or cognitive overload [17].

Deep learning models have been widely adopted for sequence prediction in educational contexts. LSTM networks have historically been effective, but GRU networks, due to their simpler gating structure, often converge faster while maintaining performance [18].

Transformers offer additional benefits, including parallelization and attention mechanisms that

highlight relevant time steps [26]. [10] applied a diagnostic Transformer for knowledge tracing, and [6, 27] demonstrated that clickstream sequences improve concept mastery estimation when modeled with hybrid deep networks. [32, 33] reported strong results for multi-category performance prediction using sequence models.

Other lines of research focus on dropout prediction using ensemble or classical ML models. Studies by [34-37] found that tabular models remain competitive when features are well-engineered. Across this body of literature, a consistent limitation emerges: few studies construct full daily sequences spanning the entire academic period. Instead, they rely on weekly summaries, partial sequences, or event-level features. Our study fills this methodological gap by evaluating GRU and Transformer models using daily behavioral sequences aligned over 150 days.

Recent studies have explored different strategies for early detection of student risk using learning analytics and educational data mining techniques. However, these studies differ substantially in their temporal representation of learner behavior and in the scope of their experimental evaluation. Table 1 summarizes several recent studies and highlights their temporal modeling strategies and evaluation designs.

Most existing studies rely on aggregated behavioral features, limited early windows, or single-model evaluation. In contrast, this study focuses on aligned daily engagement sequences and evaluates multiple sequence architectures within a controlled multi-horizon early-warning framework.

3. Methodology

The methodology preserves the temporal structure of student behavior by maintaining daily engagement signals over a standardized 150-day timeline rather than compressing interaction logs into static summaries. This design allows models to capture not only the intensity of engagement but also the timing of behavioral changes associated with academic risk. The workflow starts from raw CSV tables and produces out-of-fold predictions for cross-architecture evaluation. In addition to sequence models, an aggregated tabular baseline is included to assess the incremental value of explicit temporal modeling. The complete pipeline is illustrated in Fig. 2.

3.1 Data preparation

The experiments use the Open University Learning Analytics Dataset (OULAD), which contains Virtual Learning Environment (VLE) interaction logs collected by The Open University. The dataset used in this study includes activity records from 32,593 students across 22 courses and 32 module presentations, with over 10 million interaction events capturing students' engagement throughout each course presentation. All available course presentations are included to avoid sampling bias. To ensure a strict prediction-before-outcome setting, student activity is aligned to a unified daily timeline and truncated at 150 days from the start of each presentation, which occurs prior to final assessments and course completion periods.

Table 1. Comparison of recent studies on learning analytics and early warning systems, highlighting differences in temporal representation and evaluation scope

Study	Temporal Unit	Horizon	Benchmark	Key Difference
[19]	Weekly VLE logs	Weeks 1–4	Multiple ML	Uses weekly aggregation
[20]	Real-time activity	Continuous	Rule-based	Monitoring rather than prediction
[21]	Engagement sequence	Full course	Single LSTM	No multi-model benchmark
[22]	Aggregated features	End-of-course	Limited ML	No temporal alignment
[23]	Weekly engagement features	End-of-course	LSTM + SHAP	Uses weekly aggregation
[24]	Static academic & demographic features	Single prediction point	XGBoost	No temporal modeling
[25]	Aggregated interaction features	Course completion	Ensemble ML	Static feature aggregation
This Study	Daily behavioral sequences	Multi-horizon (28, 60, 90, 150 days)	TRF, GRU, LSTM, BiLSTM, TCN, CNN1D	Daily temporal modeling with controlled benchmark

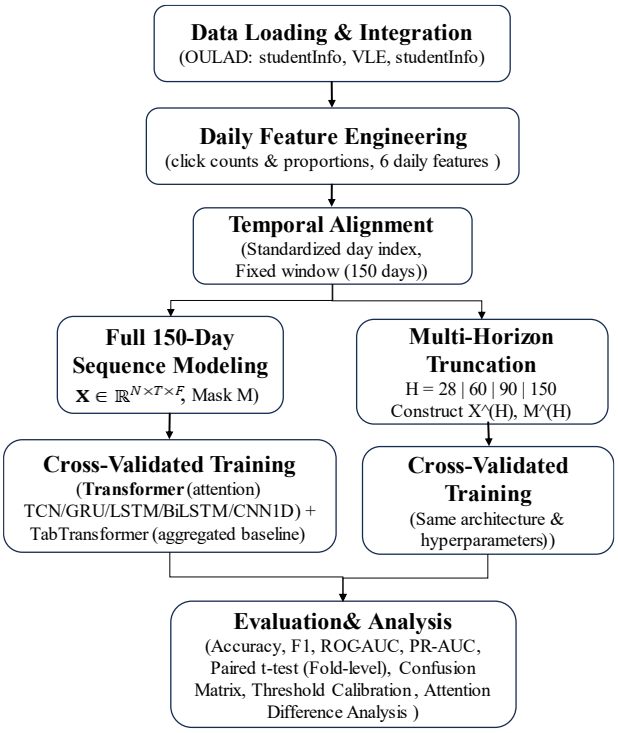

Figure. 2 Methodology Pipeline for Daily-Sequence Early Warning Framework

Temporal preprocessing begins by aligning student activity to a unified daily timeline. OULAD records interaction time as integer offsets relative to the module start. For each module presentation, the minimum offset is identified, and all records are shifted so that the first active day corresponds to index 1, producing a standardized day index (day_idx) that enables comparison across courses.

To prevent temporal leakage, the 150-day observation window was verified to occur strictly before outcome-determining events in all course presentations.

An empirical audit of the OULAD dataset shows that the maximum activity timeline per presentation ranges from 234 to 269 days (mean ≈ 255), while final assessment dates occur between day 187 and 261 (mean ≈ 222). Consequently, the 150-day truncation precedes both course completion and final grading events.

The target variable y represents student academic risk derived from the final_result field in the studentInfo table of the OULAD dataset. Students with outcomes classified as Fail or Withdrawn are labeled at-risk (y=1), while students with outcomes Pass or Distinction are labeled non-at-risk (y=0). This formulation follows common practice in early warning system research, where withdrawal and failure outcomes are jointly treated as indicators of academic risk. The label is determined only from the final course outcome and is not used during feature construction, ensuring that behavioral features remain temporally prior to the prediction target.

All input features are derived exclusively from daily VLE clickstream interactions. Assessment scores, submission timestamps, and other evaluation-related signals are excluded from the modeling pipeline, ensuring that predictions rely solely on behavioral information observed prior to the outcome.

3.2 Feature engineering

Feature engineering focuses on transforming raw clickstream logs into structured daily representations that capture both the intensity and distribution of student activity. Each day is summarized using six normalized behavioral features that reflect total engagement and the proportion of interactions across different resource categories. These features preserve short-term fluctuations that are often lost in aggregated representations and serve as the foundation for the daily sequences used by temporal models.

Each student-day pair is transformed into a vector of behavioral features derived from raw click counts. For each activity category c (resource, quiz, assignment, forum, and page), The daily click count is denoted as $x_{t,c}$ for day t. The total click volume on the day t is computed as Eq. (1).

$$C_t = \sum_c x_{\{t,c\}} \quad (1)$$

To mitigate the effects of extreme click bursts and stabilize training, the study applies a log transformation to the total daily clicks, as defined in Eq. (2).

$$\hat{C}_t = \log(1 + C_t) \quad (2)$$

In addition to total intensity, the model also needs to understand how student attention is distributed across different types of activities. Daily proportional features are therefore computed using the structure shown in Eq. (3).

$$p_{t,c} = \begin{cases} \frac{x_{t,c}}{C_t}, & C_t > 0, \\ 0, & C_t = 0. \end{cases} \quad (3)$$

The numerator x sub, t, c end subscript represents the raw count of clicks for the category on day t. In contrast, the denominator C_t normalizes this count relative to the total number of clicks that day. When $C_t = 0$, the proportion is defined as zero to avoid division by zero and to reflect complete inactivity.

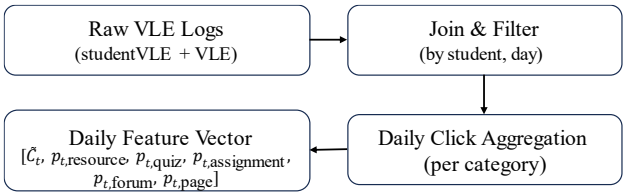

Figure. 3 Daily feature construction from OULAD clickstream logs

Combining these elements, each day t is represented as a six-dimensional feature vector in Eq. (4).

$$\mathbf{x}_t = [\tilde{C}_t, p_{t,\text{resource}}, p_{t,\text{quiz}}, p_{t,\text{assignment}}, p_{t,\text{forum}}, p_{t,\text{page}}] \quad (4)$$

This representation captures both the intensity of engagement (through $\tilde{C}_t$) and the relative distribution of effort across different activity types. Fig. 3 illustrates the construction of these daily feature vectors from raw logs, showing how individual clicks are aggregated by category and transformed into normalized daily descriptors.

3.3 Daily sequence tensor construction

After daily feature extraction, each student's behavioral trajectory is organized into a fixed-length temporal sequence. For student i, the daily feature vectors from day 1 to day 150 are concatenated as follows:

$$\mathbf{s}_i = [\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \dots, \mathbf{x}_{i,150}] \quad (5)$$

All student sequences are then stacked into a three-dimensional tensor (see Eq. (6)).

$$X \in \mathbb{R}^{N\times T\times F} \quad (6)$$

where Ndenotes the number of students, $T = 150$ is the number of days in the sequence, and $F = 6$ is the number of features per day. In parallel, a binary mask tensor (Eq. (7)).

$$M \in \{0,1\}^{N\times T} \quad (7)$$

The mask distinguishes valid temporal observations from padded positions. Observed days are assigned a value of 1, while padded days beyond the student's actual activity period are assigned 0. This masking mechanism ensures that temporal models process only valid behavioral signals and prevents artificial patterns introduced by padding from influencing learning.

3.4 Multi-horizon truncation for early warning simulation

To evaluate the operational feasibility of early risk detection, truncated observation horizons are introduced to simulate real-world early-warning deployment. Rather than relying solely on complete 150-day trajectories, predictive models are trained and evaluated using progressively shortened behavioral prefixes. Let the set of evaluation horizons be defined as: H∈{28,60,90,150}

where H denotes the maximum number of observable days available for prediction. Given the original daily sequence tensor $X \in \mathbb{R}^{N\times T\times F}$, with $T = 150$and $F = 6$, a horizon-specific tensor $X^{(H)}$is constructed as:

$$X^{(H)}_{i,t,:} = \begin{cases} X_{i,t,:}, & t \le H, \\ 0, & t > H, \end{cases} \quad (8)$$

Similarly, the corresponding validity mask $M \in \{0,1\}^{N\times T}$is truncated to obtain:

$$M^{(H)}_{i,t} = \begin{cases} M_{i,t}, & t \le H, \\ 0, & t > H. \end{cases} \quad (9)$$

Eqs (8) and (9) restrict the effective input sequence to the first Hobserved days. Because this operation modifies the available temporal evidence for each student, the effective number of valid time steps must be computed explicitly. The effective sequence length under the horizon H is therefore defined as:

$$T_i^{(H)} = \sum_{t=1}^{T} M^{(H)}_{i,t} \quad (10)$$

Here, N denotes the total number of students, Trepresents the fixed maximum sequence length of 150 days, and F indicates the number of daily behavioral features. The tensor element $X_{i,t,:}$corresponds to the feature vector of the student i at day t, while $M_{i,t}$ is a binary indicator specifying whether the day t contains a valid activity for the student i. The variable Hrepresents the truncation horizon used for early-warning simulation. The tensors $X^{(H)}$ and $M^{(H)}$ denote the horizon-specific truncated inputs and validity masks, respectively.

The quantity $T_i^{(H)}$represents the total number of valid observed days remaining after truncation. This value is used to normalize masked pooling operations

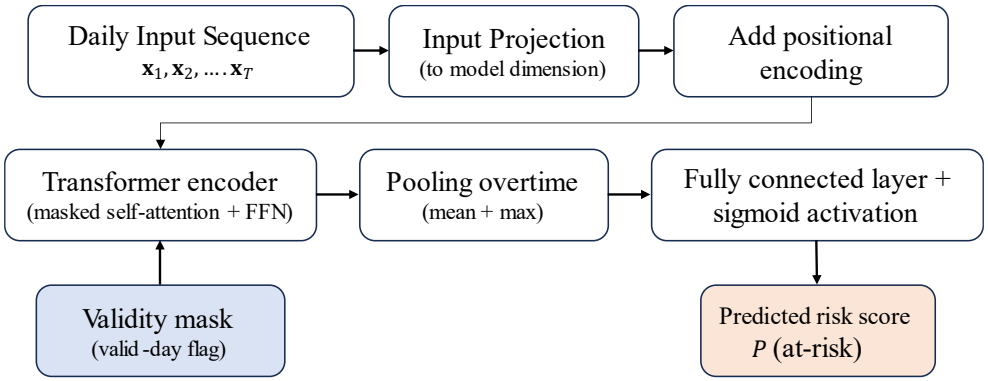

Figure. 4 Transformer encoder architecture using masked self-attention and temporal pooling to classify student risk from daily sequences

and ensure that statistical summaries are computed only over observed days, thereby preventing bias from zero-padded segments beyond the selected horizon.

To ensure fair comparison across horizons, cross-validation splits and training hyperparameters remain identical for all values of H . Consequently, performance differences across horizons reflect changes in available temporal evidence rather than variations in model configuration.

3.5 Aggregated baseline representation

To assess whether temporal ordering provides benefits beyond static summaries, an aggregated feature representation is constructed from the same daily signals. For each student, summary statistics such as mean engagement, maximum activity, total active days, and category-wise aggregates are computed and used to train a TabTransformer model under the same cross-validation splits. This design enables fair comparison between sequence-based and aggregated representations.

3.6 Temporal modeling architectures

Six sequence-based architectures were selected to represent different temporal inductive biases in modeling daily behavioral dynamics. The Transformer encoder captures long-range dependencies through masked self-attention, while the Temporal Convolutional Network (TCN) models short-term patterns using dilated convolutions. Recurrent architectures (GRU, LSTM, and BiLSTM) serve as established sequential baselines to mitigate vanishing-gradient issues, and CNN1D provides a simpler convolutional alternative for local pattern extraction. All models are evaluated using an identical stratified five-fold cross-validation to ensure that performance differences reflect architectural characteristics rather than effects of data partitioning.

3.7 Transformer encoder

The Transformer models temporal dependencies using self-attention rather than recurrence, enabling direct modeling of long-range behavioral patterns in student activity sequences. [26]. This capability is useful for modeling student behavior because certain engagement patterns, such as sudden drops in activity or extended periods of inactivity, can influence outcomes even when they occur early in the course. As shown in Fig. 4, daily features are projected with positional encoding and processed by stacked attention layers before temporal pooling produces a fixed-length representation for classification.

Positional Encoding

Because the Transformer has no inherent notion of order, positional encoding is added to encode the temporal position of the day. t, as seen in Eq. (11).

$$PE_{(t,2k)} = \sin\left(\frac{t}{10000^{\frac{2k}{d_{model}}}}\right),$$
$$PE_{(t,2k+1)} = \cos\left(\frac{t}{10000^{2k/d_{model}}}\right) \quad (11)$$

This assigns a unique sinusoidal signature to every day in the sequence. The use of sin and cos at different frequencies ensures that the model can recognize both relative and absolute positions, which is essential for reasoning over long sequences, such as the 150-day sequence used here.

Self-Attention

Self-attention computes interactions among all days using the scaled dot-product mechanism defined in Eq. (12).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}} + A_{mask}\right)V \quad (12)$$

Table 2. Sequence Model Architecture Configuration

Model	Key Parameters
Transformer	d_model=128, heads=4, layers=2, FF=256, dropout=0.2
GRU	hidden=128, layers=1, dropout=0.2
LSTM	hidden=128, layers=1, dropout=0.2
BiLSTM	hidden=128, layers=1, bidirectional
TCN	channels=128, kernel=3, dilations=[1,2,4,8], dropout=0.2
CNN1D	3 conv layers, channels=128, kernel=3, dropout=0.2

Here:Q, K, and $V$are learned projections of the daily vectors. A_{mask}prevents attention from flowing into padded days. The denominator $\sqrt{d_k}$ stabilizes gradients.

In this study, the attention mechanism allows the model to highlight specific days that signal early risk, such as abrupt drops in activity or consistently low engagement in resource or quiz categories.

Sequence Pooling

After the encoder layers, two pooling operations are applied (Eq. (13)). The pooled outputs are concatenated and passed to a sigmoid classifier.

$$\mathrm{h}_{mean} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{o}_t, \mathrm{h}_{max} = \max_{1\le t\le T} \mathrm{o}_t \qquad (13)$$

3.8 Implementation details

Models are implemented in PyTorch and trained under a controlled protocol to ensure fair architectural comparison. All models share identical optimization settings: AdamW (learning rate 10-3, weight decay 10-4), batch size 256, and early stopping (patience 5), based on validation ROC-AUC. Binary cross-entropy with logits is used as the loss function, and stratified five-fold cross-validation preserves the class distribution across folds. The architectural configurations are summarized in Table 2.

All sequence models operate on tensors of shape $T\times F$, where T=150 denotes the standardized temporal window, and F=6 represents the daily behavioral features. A validity mask excludes padded days from masked mean pooling before classification. Hyperparameters are intentionally kept consistent across architectures to ensure controlled comparison. The aggregated baseline employs a TabTransformer that treats each numerical feature as a token, projects them into a 64-dimensional embedding space, and processes them through a two-layer Transformer encoder. The resulting representations are mean-pooled and passed to a classification head (64–128–1) with a dropout rate of 0.2.

3.9 Model training procedure

All models are trained under a unified experimental pipeline to ensure consistent comparison across architectures. Stratified five-fold cross-validation is used for evaluation with shuffle enabled and a fixed random seed (42). The same fold assignments are applied to all models to guarantee identical training–validation partitions.

To evaluate early prediction capability, the daily behavioral sequences are truncated at predefined observation horizons H ∈ {28, 60, 90, 150. For each horizon, sequences are truncated before model training so that only behavioral evidence available up to day H is used.

Model evaluation uses multiple classification metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. Model selection within each fold is based on the highest validation F1-score. Hyperparameters remain fixed across architectures and horizons to isolate the effect of temporal evidence length and architectural design rather than hyperparameter tuning.

3.10 Experimental setup and reproducibility

All experiments follow a fully specified and reproducible pipeline with fixed random seeds (42). The dataset is constructed from OULAD tables (studentInfo, studentVLE, vle, studentAssessment, assessments), where interaction logs are merged with activity types and organized into student–presentation instances.

Activity types are mapped into six categories (resource, quiz, assign, forum, page, other), and daily interactions are aggregated into a standardized 150-day timeline.

Table 3. Dataset characteristics and label distribution

N Students	T Days	Non-Risk Count	At-Risk Count	Positive Ratio
29,228	150	15,382	13,846	0.474

Table 4. Comparison of classification metrics across models

Model	Acc	Prec	Recall	F1-Score
TRF	0.828	0.936	0.685	0.791
BiLSTM	0.822	0.897	0.706	0.790
LSTM	0.820	0.892	0.704	0.787
GRU	0.819	0.871	0.727	0.792
TCN	0.812	0.830	0.759	0.793
CNN1D	0.792	0.834	0.701	0.762
TabTRF	0.809	0.986	0.605	0.750

** Acc = Accuracy; Prec = Precision*

Each day is represented by six features capturing interaction intensity and category distribution, forming a tensor of size N×150×6

The binary target is derived from final_result, where Fail and Withdrawn are labeled as at-risk. The aggregated baseline uses summary statistics from the same 150-day window, ensuring no temporal leakage.

All models are evaluated under identical data splits, cross-validation folds, and training conditions. Feature normalization is applied only within each training fold. This controlled setup ensures that performance differences reflect model architecture and temporal representation. Full preprocessing and model configurations will be made available to support reproducibility.

3.11 Statistical significance testing

To assess whether performance differences across observation horizons are statistically meaningful, paired statistical tests are conducted at the fold level. For each pair of horizons H_a and H_b Performance metrics obtained from the same cross-validation folds are compared using a paired t-test.

Let $m_k^{(H)}$denote the evaluation metric (e.g., PR-AUC) obtained in fold $kunder$ horizon H, where $k = 1, \dots, K$and $K = 5$. The difference between two horizons is defined as:

$$d_k = m_k^{(H_a)} - m_k^{(H_b)} \tag{14}$$

The paired t-statistic is computed as:

$$t = \frac{\bar{d}}{s_d/\sqrt{K}} \tag{15}$$

Where $\bar{d}$ is the mean of fold-level differences, s_d is the standard deviation of the differences and $K$is the number of folds

Statistical significance is evaluated at the 0.05 level. This procedure determines whether performance improvements across increasing temporal horizons are attributable to systematic temporal accumulation rather than random variation across folds.

3.12 Evaluation metrics

Multiple metrics are used to assess both discrimination and practical usefulness. While accuracy shows overall correctness, greater emphasis is placed on Precision, Recall, and F1-score to capture the trade-off between detecting at-risk students and limiting false alarms. ROC-AUC and PR-AUC are also reported to evaluate ranking performance across thresholds.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{16}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{17}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{18}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{19}$$

4. Results and discussion

This section reports the empirical performance of the proposed daily sequence framework on the OULAD dataset. The discussion follows the methodological flow in Section 3. It begins with a descriptive view of the daily representation, then expands the evaluation from a multiple-model comparison to a controlled benchmark across multiple sequence architectures. Finally, it introduces a representation-level comparison against an aggregated tabular baseline to assess whether preserving daily temporal structure yields measurable advantages for early warning. In addition, progressive multi-horizon evaluation is performed to assess how predictive performance evolves as temporal evidence accumulates.

4.1 Descriptive view daily sequence dataset

After all preparation steps, the final dataset consists of N=29,228 learners, each represented by a sequence of T=150 days and F=6 daily features, yielding a main input tensor with shape (29,228,150,6).

Table 3 shows that the binary label distribution is moderately imbalanced, with 15,382 learners in the non-risk class and 13,846 learners in the at-risk class. This distribution reflects a realistic early-warning setting in which a substantial fraction of students failed or withdrew, while the majority completed the course. Importantly, the size of the positive class is sufficient to support stable model training and robust out-of-fold evaluation, particularly for threshold-sensitive operational metrics such as recall and false-negative rates.

4.2 Cross-architecture comparison (Full 150-Day Only)

The first stage of benchmarking evaluates whether the proposed daily representation provides

stable predictive signals across different temporal inductive biases. Six sequence-based architectures were trained under identical five-fold stratified cross-validation conditions: Transformer (TRF), Bidirectional Long Short-Term Memory (BiLSTM), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Temporal Convolutional Network (TCN), and One-Dimensional Convolutional Neural Network (CNN1D). This design isolates architectural differences by keeping data splits, label distributions, and training protocols constant, enabling direct comparison across temporal modeling strategies.

In addition to sequence-based models, an aggregated TabTransformer (TabTRF) baseline was evaluated using static summary features derived from the same students. This baseline removes temporal ordering to assess whether performance gains arise from model architecture or from preserving daily temporal structure. Tables 4 and 5 summarize the out-of-fold (OOF) performance across models.

The results reveal two key patterns. First, the strongest sequence models perform within a narrow range of F1 scores, indicating that the daily temporal representation itself contributes substantially to predictive performance. This finding supports the importance of preserving fine-grained engagement trajectories rather than compressing interactions into weekly summaries.

Second, different architectures exhibit complementary strengths. The Transformer achieves the highest ranking performance, with the best Precision (0.936), ROC-AUC (0.890), and PR-AUC (0.911). In contrast, the TCN provides the best-balanced classification, with the highest F1 (0.793) and Recall (0.759), while GRU remains a stable baseline with balanced precision and recall.

The aggregated TabTransformer baseline shows a markedly different behavior. Although it achieves extremely high precision (0.986), its recall (0.605) and F1 (0.750) are substantially lower, suggesting that static aggregation tends to capture only obvious risk cases while missing students whose behavioral changes emerge gradually over time.

Table 5. Comparison of ranking metrics across models

Model	ROC-AUC	PR-AUC
TRF	0.890	0.911
BiLSTM	0.884	0.906
LSTM	0.884	0.905
GRU	0.887	0.907
TCN	0.884	0.907
CNN1D	0.854	0.862
TabTRF	0.854	0.884

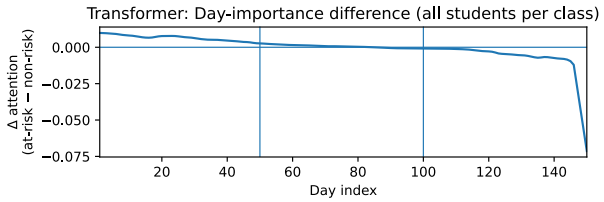

Figure. 5 Smoothed trajectory of Δ attention (at-risk minus non-risk) across the 150-day sequence, illustrating the temporal direction and magnitude of attention differences between classes

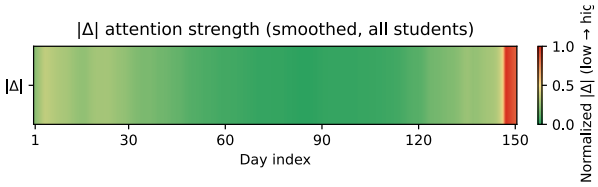

Figure. 6 Normalized absolute attention difference |Δ| between at-risk and non-risk students

4.3 Attention difference analysis

To further investigate how the Transformer differentiates between at-risk and non-at-risk learners, the attention weights from the last encoder layer were aggregated across the entire population for each class (13,846 at-risk and 15,382 non-at-risk students). For each day index, the difference in attention was computed as Δ, defined as the average attention received by the at-risk group minus that received by the non-risk group. Fig. 5 presents the smoothed temporal trajectory of Δ across the 150-day sequence. Positive Δ values indicate that a particular day receives relatively greater attention for at-risk learners, whereas negative values indicate stronger attention to non-at-risk learners.

The results reveal a consistent pattern. During the early phase of the semester (approximately days 1–40), Δ remains slightly positive, suggesting that the Transformer assigns relatively greater importance to early behavioral signals in students who eventually become at-risk. In the mid-semester period (approximately days 50–100), Δ approaches zero, indicating comparable attention allocation across both groups. Toward the later stages of the semester (after approximately day 120), Δ becomes increasingly negative, suggesting that the model allocates relatively greater attention to late-stage behavioral patterns among non-risk learners.

Fig. 6 illustrates the normalized absolute attention difference |Δ| between at-risk and non-risk students across the full 150-day sequence. Unlike the signed Δ trajectory, this visualization focuses exclusively on the magnitude of class-level differentiation, independent of direction. Each day shows how

strongly the Transformer distinguishes between the two groups in attention allocation.

The results show that most mid-semester days exhibit relatively low |Δ| values, indicating comparable attention patterns across classes during this period. In contrast, the magnitude of |Δ| increases toward the later stage of the semester, with the strongest differentiation observed near the final days. This pattern suggests that late-phase behavioral signals play a disproportionately important role in the model's discriminative process.

Importantly, this visualization complements the signed Δ trajectory by highlighting where temporal differentiation is strongest, regardless of which class receives greater attention. Together, these results demonstrate that the Transformer's decision process is temporally structured and phase-sensitive, reinforcing the value of daily-resolution modeling in early risk detection.

Together, Figs. 5.1 and 5.2 demonstrate that the Transformer's decision mechanism is temporally structured and phase-sensitive. Early engagement dynamics and late-stage behavioral stability carry the most discriminative signals, reinforcing the importance of preserving daily temporal resolution rather than relying on aggregated representations.

4.4 ROC–PR curve analysis (Full 150-Day)

While Tables 4.2 and 4.3 report scalar metrics at the default threshold, model behavior was further examined across the full decision spectrum using ROC and PR curves derived from out-of-fold predictions. This analysis is not intended as a model selection exercise, but rather as a controlled comparison of temporal modeling behaviors under identical experimental conditions. By evaluating all architectures using the same data representation, preprocessing pipeline, and validation protocol, the results provide a consistent reference point for understanding how different temporal designs influence ranking performance.

Fig. 7 presents the combined ROC curves for the six sequence architectures and the aggregated TabTRF baseline.

Sequence models consistently show stronger discrimination across thresholds. TRF achieves the highest ROC-AUC (0.890), followed by GRU (0.887), TCN (0.884), LSTM (0.884), and BiLSTM (0.884). CNN1D and TabTRF exhibit lower ROC-AUC values (both 0.854), indicating weaker ranking ability. The ROC curves of the sequence models remain clearly separated from the diagonal reference line, suggesting robust separability between at-risk and non-risk learners.

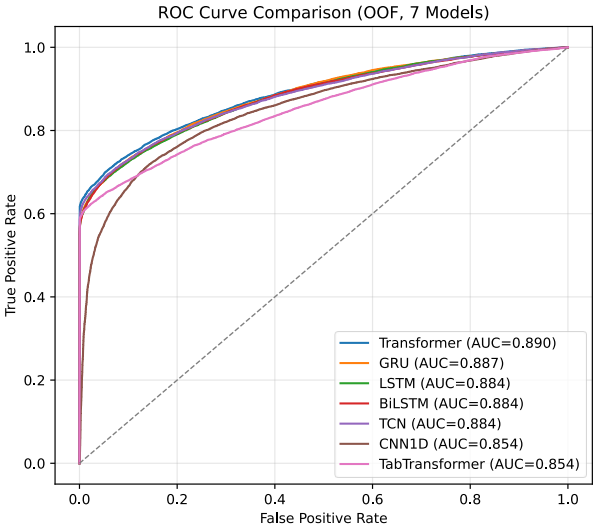

Figure. 7 Combined ROC curve comparison across sequence models and aggregated TabTransformer baseline (OOF predictions)

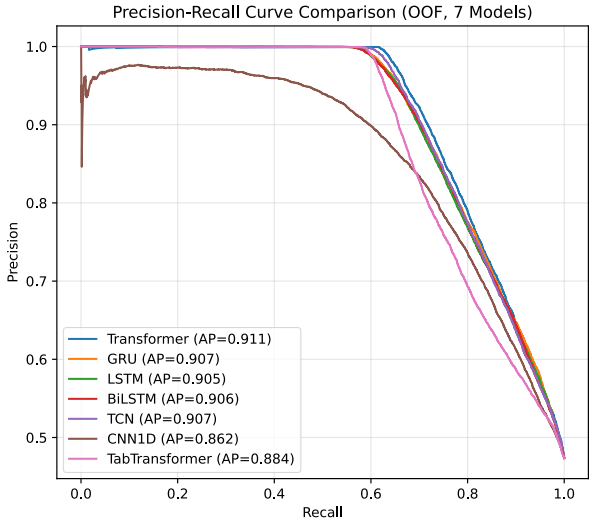

Figure. 8 Combined Precision–Recall curve comparison across sequence models and aggregated TabTransformer baseline (OOF predictions)

However, in early warning systems where class imbalance is present, and intervention resources are limited, Precision–Recall behavior is often more informative. Fig. 8 illustrates the combined PR curves.

The PR analysis provides additional insight into model behavior under realistic risk distributions. The Transformer achieves the highest PR-AUC (0.911), followed closely by the GRU (0.907) and the TCN (0.907).

Although the TabTransformer achieves extremely high precision at specific thresholds, its curve drops more rapidly as recall increases, reflecting its lower sensitivity (Recall = 0.605 at threshold 0.5). In contrast, sequence-based models

maintain a more stable precision level across broader recall ranges, indicating stronger coverage of at-risk learners without drastic increases in false alarms.

Overall, Figs. 7 and 8 reinforce two important findings. First, daily sequence modeling consistently improves ranking discrimination compared to aggregated representations. Second, the advantage of sequence models is particularly pronounced when recall increases, suggesting that temporal dynamics meaningfully contribute to identifying learners whose risk patterns evolve.

4.5 Multi-horizon performance evaluation

To assess early-warning feasibility with limited temporal evidence, model performance was evaluated using progressively truncated observation horizons H ∈ {28, 60, 90, 150}, as defined in Section 3.3.1. For each horizon, truncated tensors $X(H)$ and masks $M(H)$ were constructed while preserving identical cross-validation splits and training hyperparameters across all configurations. This ensures that observed performance differences reflect changes in available temporal evidence rather than architectural or optimization variations.

Tables 6 and 7 report the mean cross-validated performance across horizons for the Transformer model. Fold-level metrics are aggregated using the same five stratified splits.

A clear monotonic improvement pattern is observed. For the Transformer, PR-AUC increases from 0.628 at $H = 28$ to 0.764 at $H = 60$, 0.838 at $H = 90$, and 0.912 at $H = 150$. Similar trends are observed for GRU, although with slightly lower ranking performance. These results indicate that discriminative temporal signals accumulate progressively as more behavioral observations become available.

Fig. 9 illustrates the relationship between PR-AUC and the observation horizon for the Transformer and GRU models. Performance improves as the horizon increases, reflecting the accumulation of temporal evidence. In contrast, the aggregated baseline is computed once from the full 150-day summary and therefore remains constant across horizons, highlighting the sensitivity of sequence models to temporal information.

To evaluate whether improvements between horizons are statistically meaningful, paired t-tests were conducted at the fold level. For example, comparing the PR-AUC between $H = 90$ and $H = 150$ yields a statistically significant difference ($p < 0.001$), confirming that additional temporal evidence provides systematic performance gains rather than random fold-level variation.

Table 6. Mean classification performance across observation horizons for the Transformer model

Cutoff days	Acc	Prec	Recall	F1-Score
28	0.598	0.559	0.728	0.631
60	0.667	0.654	0.640	0.645
90	0.733	0.757	0.654	0.699
150	0.831	0.928	0.699	0.797

Table 7. Mean ranking performance (ROC-AUC and PR-AUC) across observation horizons for the Transformer model

Cutoff days	ROC-AUC	PR-AUC
28	0.656	0.628
60	0.742	0.764
90	0.810	0.838
150	0.891	0.912

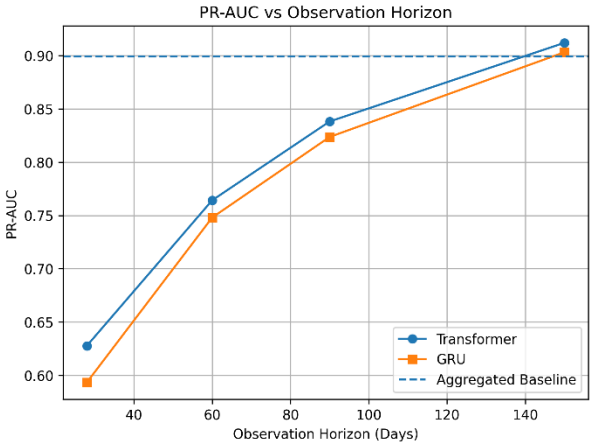

Figure. 9 PR-AUC progression across observation horizons

Overall, these findings demonstrate that early behavioral signals are detectable within the first month; however, predictive stability and ranking quality improve substantially beyond mid-semester. The progressive performance trajectory reinforces the importance of modeling temporally structured behavioral evolution rather than relying solely on aggregated representations.

These results should be interpreted within the context of a controlled experimental design, in which all models are evaluated using identical data splits and feature representations. Rather than aiming to outperform specific external studies, the objective is to isolate the effect of temporal granularity on early risk detection. This positioning also explains why direct numerical comparisons with prior studies should be interpreted with caution, given differences in preprocessing, label definitions, and evaluation settings.

Table 8. Pairwise statistical comparison of PR-AUC between sequence architectures

Model Comparison	p-value
Transformer vs GRU	0.000418
Transformer vs TCN	0.000868
GRU vs TCN	0.333

Table 9. Confusion matrices for all models

Model	TN	FP	FN	TP
TRF	14,737	645	4,368	9,478
TabTRF (Aggregated)	15,259	123	5,471	8,375
BiLSTM	14,261	1,121	4,074	9,772
LSTM	14,206	1,176	4,096	9,750
GRU	13,888	1,494	3,783	10,063
CNN1D	13,451	1,931	4,145	9,701
TCN	13,224	2,158	3,339	10,507

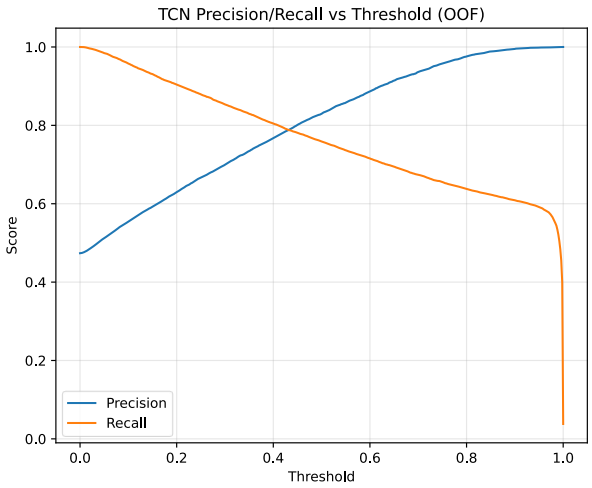

Figure. 10 Precision and Recall as a function of decision threshold for the TCN model (OOF predictions)

Table 10. Comparison with recent OULAD-based student outcome prediction studies (2024–2025)

Study	Prediction Task	Model	Temporal Setting	Result	Comparison Setting
[23]	4-class outcome prediction (Fail / Pass / Distinction / Withdraw) and early risk detection	Multidimensional time-series deep learning (MTAPSP)	Full course timeline	74% accuracy / 73% F1 (4-class); 99.08% accuracy and F1 for binary early-risk detection	Different preprocessing and label definition
[32]	Binary academic success prediction	LSTM + SHAP (LSTM-SHAP)	Full course timeline	92.88% accuracy; F1 = 92.79%	Subset population (disabled students)
[33]	Multiple prediction tasks: at-risk, high-performer, and multiclass performance prediction	BiLSTM + CNN + attention (MultiFAR)	Full course timeline	Accuracy range: 80.31–97.12%, depending on task	Multi-task setting, not directly comparable
[34]	3-class student outcome prediction (Distinction / Pass / Fail)	RF, LR, SVM, LDA	Full course timeline	Random Forest : 91% accuracy	Aggregated features, no temporal sequence
[35]	Binary and multiclass performance prediction	GRU compared with LR, NB, KNN, RF, XGBoost, and LSTM-XGBoost	Weekly and monthly interaction windows	GRU achieves 90.13% accuracy	Full-course prediction, no early horizon
This study	Early at-risk student detection	Transformer (daily behavioral sequences)	Fixed 150-day observation window; evaluated at 28, 60, 90, and 150 days	ROC-AUC = 0.890; PR-AUC = 0.911; F1 = 0.793	Controlled benchmark (same data, same split, same pipeline)

4.6 Comparison with recent OULAD-Based studies

The multi-horizon results show that performance improves as more behavioral data becomes available, indicating that daily sequences capture meaningful signals for early risk detection.

Table 10 shows that most OULAD-based studies rely on aggregated features and evaluate performance using substantial portions of course data, often leading to high accuracy.

In contrast, this study models fine-grained daily behavior and evaluates performance across multiple horizons (28, 60, 90, 150 days) within a fixed pre-outcome window, enabling earlier and temporally consistent risk detection.

4.7 Architecture-level statistical comparison

Pairwise statistical tests were conducted across cross-validation folds to determine whether the observed performance differences among the leading architectures are statistically significant.

As shown in Table 8, the Transformer significantly outperforms both the GRU and the TCN in PR-AUC ($p < 0.01$). In contrast, the difference between GRU and TCN is not statistically significant ($p = 0.333$), indicating comparable predictive performance between these two architectures.

4.8 Error structure and operational implications

To further examine deployment implications, confusion matrices were analyzed at the default classification threshold of 0.5 over the full observation horizon (H = 150), at which models achieve their strongest discrimination. The results are summarized in Table 9.

The comparison reveals distinct operational profiles. The Transformer produces the fewest false positives (645), reflecting a precision-oriented approach. At the same time, the TCN yields the lowest false-negative count among sequence models (3,339), indicating greater sensitivity to at-risk learners. GRU strikes a balance between precision and recall.

In contrast, the aggregated TabTransformer baseline generates very few false positives (123) but substantially more false negatives (5,471), suggesting that static aggregation detects only the most evident risk cases while missing gradual behavioral decline. These results highlight the operational advantage of preserving daily temporal structure for early-warning systems.

4.9 Threshold calibration and deployment scenarios

While previous evaluations use the default classification threshold of 0.5, real-world early-warning systems often require threshold calibration based on institutional capacity and intervention priorities. A fixed threshold may not align with constraints such as limited counseling resources or the need to control false alarms.

Threshold sensitivity was analyzed using the TCN, which shows the best balance between F1 and Recall. Lower thresholds increase recall but reduce precision, while higher thresholds produce fewer but more reliable alerts. At the default threshold (0.5), the model achieves balanced performance (F1 = 0.793; Recall = 0.759).

Operational implications are shown in Fig. 11 through the relationship between the false positive rate (FPR) and false negative rate (FNR). Lower thresholds reduce missed cases (lower FNR) but increase false alarms, whereas higher thresholds show the opposite effect.

These results indicate that threshold selection should reflect institutional priorities. Resource-constrained institutions may adopt higher thresholds to focus on high-confidence alerts, while institutions emphasizing broad early intervention may prefer lower thresholds to maximize coverage.

5. Conclusion

This study examines whether preserving daily temporal structure improves early detection of at-risk students in online learning. Each learner is represented as a standardized 150-day behavioral sequence derived from six normalized daily engagement features extracted from OULAD clickstream logs, retaining interaction dynamics often lost in aggregated summaries. Under identical experimental conditions, sequence-based models consistently outperform the aggregated TabTransformer baseline in recall-oriented metrics. The Transformer achieves the highest ranking performance (ROC-AUC = 0.890; PR-AUC = 0.911), while the Temporal Convolutional Network (TCN) provides the most balanced classification results with the highest F1-score (0.793) and recall (0.759).

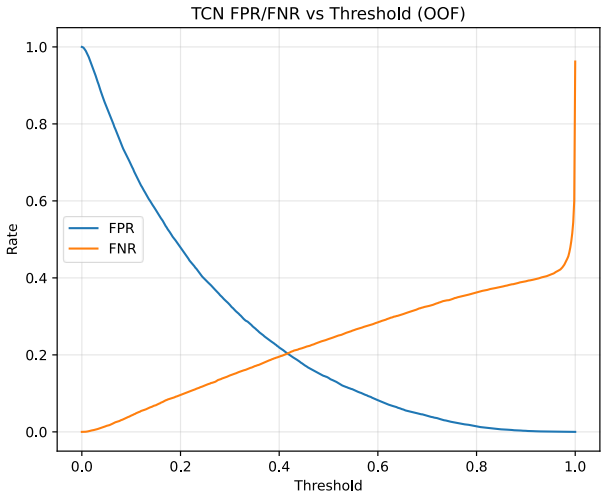

Figure. 11 FPR and FNR across decision thresholds for the TCN model

In contrast, the aggregated baseline yields extremely high precision (0.986) but substantially lower recall (0.605), indicating that static aggregation tends to identify only the most obvious risk cases while missing students whose behavioral decline develops gradually over time.

The multi-horizon evaluation further shows that predictive performance improves as longer behavioral histories become available. Across the 28-, 60-, 90-, and 150-day observation windows, model performance increases steadily, suggesting that meaningful engagement signals accumulate progressively during the learning process. This setting provides a practical approximation of real-world early-warning deployment, where predictions must be generated before final course outcomes are known.

Overall, the findings indicate that the design of temporal representations plays a critical role in early-warning performance. Rather than proposing a new neural architecture, this study shows that standardized daily alignment, combined with controlled cross-architecture benchmarking, offers a reproducible framework for evaluating temporal learning models on educational data. The results suggest that preserving daily behavioral trajectories can improve risk detection while maintaining manageable false-positive rates for institutional intervention systems.

This study focuses on the OULAD dataset, and future work will extend the framework to additional institutional datasets to examine its generalizability across different learning environments. Future studies may also explore architecture-specific hyperparameter optimization to further investigate performance variability.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Bruri Trya Sartana led the conceptualization, methodology development, software implementation, data preprocessing, formal analysis, visualization, and manuscript drafting. Safitri Juanita contributed to data curation, validation, and manuscript review. Utomo Budiyanto contributed to methodological refinement, manuscript review and editing, and project administration. Mauridhi Hery Purnomo supervised the research, provided strategic guidance, and critically reviewed the manuscript prior to submission.

References

[1] B. Flanagan, R. Majumdar and H. Ogata, "Early-warning prediction of student performance and engagement in open book assessment by reading behavior analysis", *Int. J. Educ. Technol. High. Educ.*, Vol. 19, No. 1, 2022, doi: 10.1186/s41239-022-00348-4.

[2] K. Wang, "Academic Early Warning Model for Students Based on Big Data Analysis", *Int. J. Emerg. Technol. Learn.*, Vol. 18, No. 12, pp. 16-31, 2023, doi: 10.3991/ijet.v18i12.41087.

[3] J. E. Raffaghelli, M. E. Rodríguez, A. E. Guerrero-Roldán and D. Bañeres, "Applying the UTAUT model to explain the students' acceptance of an early warning system in Higher Education", *Comput. Educ.*, Vol. 182, Art. 104468, 2022, doi: 10.1016/j.compedu.2022.104468.

[4] H. A. Alhakbani and F. M. Alnassar, "Open Learning Analytics: A Systematic Review of Benchmark Studies using Open University Learning Analytics Dataset (OULAD)", In: *Proc. of 2022 7th International Conference on Machine Learning Technologies*, pp. 81-86, 2022, doi: 10.1145/3529399.3529413.

[5] K. A. Laksitowening, T. Fahrudin, R. Insani and U. Umar, "Incorporating Learning Analytics and Business Intelligence into Higher Education E-Learning", *Khazanah Inform. J. Ilmu Komput. dan Inform.*, Vol. 10, No. 2, pp. 127-134, 2025, doi: 10.23917/khif.v10i2.4142.

[6] M. Saqr, S. L.-Pernas, S. Helske and S. Hrastinski, "The longitudinal association between engagement and achievement varies by time, students' profiles, and achievement state: A full program study", *Comput. Educ.*, Vol. 199, Art. 104787, 2023, doi: 10.1016/j.compedu.2023.104787.

[7] Á. Hernández-García, C. Cuenca-Enrique, L. D.-R.-Carazo and S. I.-Pradas, "Exploring the relationship between LMS interactions and academic performance: A Learning Cycle approach", *Comput. Human Behav.*, Vol. 155, Art. 108183, 2024, doi: 10.1016/j.chb.2024.108183.

[8] O. R. Yürüm, T. T.-Temizel and S. Yıldırım, "The use of video clickstream data to predict university students' test performance: A comprehensive educational data mining approach", *Educ. Inf. Technol.*, Vol. 28, No. 5, pp. 5209-5240, 2023, doi: 10.1007/s10639-022-11403-y.

[9] N. U. R. Junejo, M. W. Nawaz, Q. Huang, X. Dong, C. Wang and G. Zheng, "Accurate multi-

category student performance forecasting at early stages of online education using neural networks", *Sci. Rep.*, Vol. 15, No. 1, Art. 16251, 2025, doi: 10.1038/s41598-025-00256-3.

[10] Y. Yin, *et al.*, "Tracing Knowledge Instead of Patterns: Stable Knowledge Tracing with Diagnostic Transformer", In: *Proc. of ACM Web Conference 2023*, pp. 855-864, 2023, doi: 10.1145/3543507.3583255.

[11] H. Wang, "Transformer-based deep learning for adaptive pedagogy under uncertain student preferences", *Sci. Rep.*, Vol. 15, No. 1, Art. 42173, 2025, doi: 10.1038/s41598-025-25996-0.

[12] C. Sun, S. Huang, B. Sun and S. Chu, "Personalized learning path planning for higher education based on deep generative models and quantum machine learning: a multimodal learning analysis method integrating transformer, adversarial training and quantum state classification", *Discov. Artif. Intell.*, Vol. 5, No. 1, Art. 29, 2025, doi: 10.1007/s44163-025-00252-6.

[13] N. Sghir, A. Adadi and M. Lahmer, "Recent advances in Predictive Learning Analytics: A decade systematic review (2012-2022)", *Educ. Inf. Technol.*, Vol. 28, No. 7, pp. 8299-8333, 2023, doi: 10.1007/s10639-022-11536-0.

[14] Q. Huang and J. Chen, "Enhancing Academic Performance Prediction With Temporal Graph Networks for Massive Open Online Courses", *J. Big Data*, Vol. 11, Art. 52, 2024, doi: 10.1186/s40537-024-00918-5.

[15] M. Saqr and S. López-Pernas, "How CSCL roles emerge, persist, transition, and evolve over time: A four-year longitudinal study", *Comput. Educ.*, Vol. 189, Art. 104581, 2022, doi: 10.1016/j.compedu.2022.104581.

[16] C. J. Arizmendi, *et al.*, "Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work", *Behav. Res. Methods*, Vol. 55, No. 6, pp. 3026-3054, 2023, doi: 10.3758/s13428-022-01939-9.

[17] J. C.-Y. Sun, Y. Liu, X. Lin and X. Hu, "Temporal learning analytics to explore traces of self-regulated learning behaviors and their associations with learning performance, cognitive load, and student engagement in an asynchronous online course", *Front. Psychol.*, Vol. 13, Art. 1096337, 2023, doi: 10.3389/fpsyg.2022.1096337.

[18] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation", In: *Proc. of 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724-1734, 2014, doi: 10.3115/v1/D14-1179.

[19] Hamsiah, N. Adiyati and R. Subekti, "Early Prediction of At-Risk Students Using Minimal Data: A Machine Learning Framework for Higher Education", *Digitus J. Comput. Sci. Appl.*, Vol. 3, No. 2, pp. 105-116, 2025, doi: 10.61978/digitus.v3i2.953.

[20] L. U.-Gadella, I. E.-Ayres, J. A. Fisteus, C. A.-Hoyos and C. D. Kloos, "Analysis and Prediction of Students' Performance in a Computer-Based Course Through Real-Time Events", *IEEE Trans. Learn. Technol.*, Vol. 17, pp. 1794-1804, 2024, doi: 10.1109/TLT.2023.3331433.

[21] E. Kalita, H. El Aouifi, A. Kukkar, S. Hussain, T. Ali and S. Gaftandzhieva, "LSTM-SHAP based academic performance prediction for disabled learners in virtual learning environments: a statistical analysis approach", *Soc. Netw. Anal. Min.*, Vol. 15, No. 1, Art. 65, 2025, doi: 10.1007/s13278-025-01484-1.

[22] S. A. Oyedotun, O. P. Ejenarhome and G. P. Oise, "Learning Analytics and Predictive Modeling: Enhancing Student Success through Data-Driven Insights", *J. Sci. Res. Rev.*, Vol. 2, No. 3, pp. 42-51, 2025, doi: 10.70882/josrar.2025.v2i3.77.

[23] N. U. R. Junejo, M. W. Nawaz, Q. Huang, X. Dong, C. Wang and G. Zheng, "Accurate multi-category student performance forecasting at early stages of online education using neural networks", *Sci. Rep.*, Vol. 15, No. 1, Art. 16251, 2025, doi: 10.1038/s41598-025-00256-3.

[24] B. C.-Mendívil, A. A.-González, N. J. Ríos-Vázquez and M. d. P. Lizardi-Duarte, "Predicting Student Dropout from Day One: XGBoost-Based Early Warning System Using Pre-Enrollment Data", *Appl. Sci.*, Vol. 15, No. 16, Art. 9202, 2025, doi: 10.3390/app15169202.

[25] A. M. Rabelo and L. E. Zárate, "A model for predicting dropout of higher education students", *Data Sci. Manag.*, Vol. 8, No. 1, pp. 72-85, 2025, doi: 10.1016/j.dsm.2024.07.001.

[26] A. Vaswani, *et al.*, "Attention is all you need", *Adv. Neural Inf. Process. Syst.*, Vol. 30, pp. 5998-6008, 2017, doi: 10.1201/9781003561460-19.

[27] Y. Liu, S. Fan, S. Xu, A. Sajjanhar, S. Yeom and Y. Wei, "Predicting Student Performance Using Clickstream Data and Machine Learning", *Educ. Sci.*, Vol. 13, No. 1, Art. 17, 2023, doi: 10.3390/educsci13010017.

[28] K. S. Selim and S. S. Rezk, "On predicting school dropouts in Egypt: A machine learning approach", *Educ. Inf. Technol.*, Vol. 28, No. 7, pp. 9235-9266, 2023, doi: 10.1007/s10639-022-11571-x.

[29] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization", *Comput. Educ. Artif. Intell.*, Vol. 3, Art. 100066, 2022, doi: 10.1016/j.caeai.2022.100066.

[30] J. Pecuchova and M. Drlik, "Predicting Students at Risk of Early Dropping Out from Course Using Ensemble Classification Methods", *Procedia Comput. Sci.*, Vol. 225, pp. 3223-3232, 2023, doi: 10.1016/j.procs.2023.10.316.

[31] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education", *Technol. Soc.*, Vol. 76, Art. 102474, 2024, doi: 10.1016/j.techsoc.2024.102474.

[32] D. J. Kalita, S. Roy and P. Barman, "LSTM-SHAP based academic performance prediction for online learners using OULAD", *Soc. Netw. Anal. Min.*, Vol. 15, Art. 32, 2025.

[33] M. Fazil and A. Albahlal, "Machine Learning Approaches for Early Prediction of Student Performance Using Learning Management System Interaction Data", *Appl. Sci.*, 2025.

[34] F. E. EL Habti, M. Hiri, M. Chrayah, A. Bouzidi and N. Aknin, "Enhancing Student Performance Prediction in e-Learning Ecosystems Using Machine Learning Techniques", *Int. J. Inf. Educ. Technol.*, Vol. 15, No. 2, pp. 301-311, 2025, doi: 10.18178/ijiet.2025.15.2.2243.

[35] Z. Khoudi, N. Hafidi, M. Nachaoui and S. Lyaqini, "Leveraging machine learning and clickstream data to improve student performance prediction in virtual learning environments", *Inf. Discov. Deliv.*, Vol. 54, No. 4, 2025, doi: 10.1108/IDD-08-2024-0120.