

bruri-2 sartana-2

INAS EWS Transformer-Turnitin Check

 Jurnal

Document Details

Submission ID**trn:oid:::3618:128768947****Submission Date****Feb 20, 2026, 11:45 PM GMT+7****Download Date****Feb 20, 2026, 11:51 PM GMT+7****File Name****INAS EWS Transformer-Turnitin Check.docx****File Size****3.3 MB****15 Pages****7,180 Words****46,537 Characters**

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

-  **41 Not Cited or Quoted 6%**
Matches with neither in-text citation nor quotation marks
-  **1 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 4%  Internet sources
- 2%  Publications
- 3%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 41 Not Cited or Quoted 6%**
Matches with neither in-text citation nor quotation marks
- 1 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 4% Internet sources
- 2% Publications
- 3% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.mdpi.com	<1%
2	Internet	inass.org	<1%
3	Internet	www.techscience.com	<1%
4	Student papers	International Institute of Information Technology, Hyderabad on 2024-10-23	<1%
5	Student papers	Glyndwr University on 2025-10-20	<1%
6	Student papers	Tilburg University on 2025-12-01	<1%
7	Internet	bright-journal.org	<1%
8	Internet	www.technicalgeography.org	<1%
9	Publication	"The Second International Symposium on Generative AI and Education (ISGAIE'20...	<1%
10	Internet	www.coursehero.com	<1%

11	Internet	app.readytensor.ai	<1%
12	Internet	www.researchsquare.com	<1%
13	Publication	"Advanced Intelligent Computing Technology and Applications", Springer Science...	<1%
14	Student papers	Georgia Institute of Technology Main Campus on 2026-02-16	<1%
15	Student papers	Sheffield Hallam University on 2026-01-08	<1%
16	Student papers	Vellore Institute of Technology on 2026-02-07	<1%
17	Internet	itegam-jetia.org	<1%
18	Publication	"Responsible Artificial Intelligence", Springer Science and Business Media LLC, 2025	<1%
19	Publication	Abdul Razaque, Zhuldyz Kalpeyeva, Uskenbayeva Raissa Kabiyeveva, Ryskhan Zha...	<1%
20	Student papers	Tilburg University on 2025-12-01	<1%
21	Student papers	University of Johannesburg on 2026-02-16	<1%
22	Internet	educationaltechnologyjournal.springeropen.com	<1%
23	Internet	files01.core.ac.uk	<1%
24	Internet	oulurepo.oulu.fi	<1%

25	Internet	www.anzam.org	<1%
26	Internet	www.frontiersin.org	<1%
27	Internet	www.researchgate.net	<1%
28	Student papers	The University of Manchester on 2021-05-17	<1%
29	Publication	Bayan Alnasyan, Mohammed Basher, Madini Alassafi, Khozama Alnasyan. "Kanfo...	<1%
30	Student papers	City University on 2024-10-16	<1%
31	Publication	Hung-Yeh Lin, Teng-Sheng Moh, Bryce Westlake. "Gun Violence News Informatio...	<1%



Temporal Models for Early At-Risk Student Detection: A Daily-Sequence Framework with Multi-Horizon Evaluation

Bruri Trya Sartana^{1*} Safitri Juanita² Utomo Budiyo³ Mauridhi Hery Purnomo⁴

^{1,2,3}Faculty of Information Technology, Universitas Budi Luhur, Indonesia

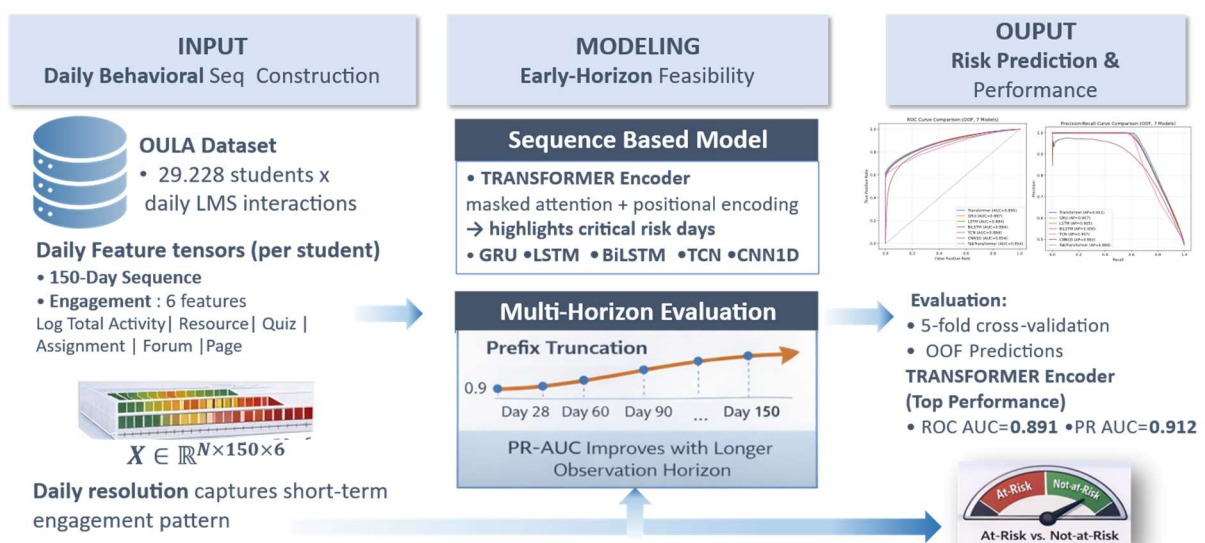
⁴Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Indonesia

* Corresponding author's Email: brury@budiluhur.ac.id

Abstract: Prior early warning systems often rely on temporally aggregated representations, such as weekly summaries, to simplify modeling. Although practical, such aggregation compresses short-term behavioral fluctuations and obscures engagement timing patterns that frequently precede academic risk. This study reframes early warning as a fine-grained temporal modeling problem by introducing a daily-sequence framework constructed from complete 150-day learner trajectories. Each student is represented using six normalized daily behavioral features that preserve engagement dynamics over time. Six sequence-based architectures, including Transformer, GRU, LSTM, BiLSTM, TCN, and CNN1D, are evaluated under identical stratified five-fold cross-validation. Results indicate that all temporal models perform within a narrow range, suggesting that representation granularity contributes substantially to predictive effectiveness. The Transformer achieves the **strongest ranking performance (ROC-AUC: 0.891; PR-AUC: 0.912)**, while GRU and TCN demonstrate competitive sensitivity. To assess practical early-warning feasibility, truncated observation horizons (28, 60, 90, and 150 days) are further examined. Predictive performance increases progressively with longer horizons, confirming that risk signals accumulate over time. Notably, at the full 150-day horizon, daily sequence modeling surpasses aggregated tabular baselines, demonstrating the advantage of dynamic temporal representation over static feature aggregation.

Keywords: Early Warning System, Temporal Models, Daily Sequence Modeling, Multi-Horizon Evaluation, Transformer

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][2][3][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. [5].

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][9][10]. For instance, [11] 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][13][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].

This study addresses this gap by developing a daily-sequence temporal modeling framework based on complete 150-day behavioral trajectories and progressive observation horizons. The main contributions are fourfold.

1. Daily behavioral representation using six normalized features preserves fine-grained engagement dynamics that are typically obscured in aggregated summaries.
2. A controlled comparison of convolutional, recurrent, and attention-based temporal models isolates the effect of representation design from architectural bias.
3. Aggregated tabular baseline quantifies the effect of removing daily temporal structure and serves as a static reference point.
4. Multi-horizon evaluation (28, 60, 90, and 150 days) examines how predictive performance evolves as temporal evidence accumulates, supported by cross-validated metrics and statistical testing.

At the core of this study is the assumption that student risk is influenced not only by the amount of activity performed, but also by when and how engagement patterns evolve. Two students may exhibit similar levels of total interaction, yet differ substantially in risk due to differences in temporal consistency, early disengagement, or abrupt drops in activity. By preserving daily resolution and modeling behavioral sequences at varying temporal horizons, the proposed framework retains engagement dynamics that are typically obscured in aggregated representations.

2. Related Works

Research on learning analytics shows that temporal information is valuable for identifying struggling students. Studies focusing on group problem-solving logs [19], video clickstream interaction, and fine-grained digital engagement [20] [21] 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. [21].

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 [22]. Transformers offer additional benefits, including parallelization and attention mechanisms that highlight relevant time steps. [23]. [14] applied a diagnostic Transformer for knowledge tracing, and [24] [8] demonstrated that clickstream sequences improve concept mastery estimation when modeled with hybrid deep networks. [25][26] reported strong results for multi-category performance prediction using sequence models.

Table 3.1. Comparative analysis of temporal methodologies in educational Early Warning Systems

Study (Year)	Methodology	Data Type	Key Limitations	Relevance
Sun (2023) [31]	Lag Seq + ENA	Course Logs	Aggregate course-level focus; lacks individual-level daily trajectory modeling	Sequence over aggregates
Yürüm (2023) [11]	Supervised ML(Video)	Clickstream	Focuses on isolated video sessions	Seq data improves prediction
Shou (2024) [32]	Multidimensional Time-Series	E-learning	Information loss due to coarse granularity	Merging multi-activity streams
Junejo (2025) [12]	Sequence-based Neural Prediction	Academic Time Series	Grade-based time series focus; lacks alignment with behavioral risk signals"	Early-semester feasibility
Kalita (2025) [34]	LSTM + SHAP	OULAD Logs	Focus on global interpretability; lacks fine-grained temporal focus	Validates OULAD for sequential modeling
Uguina (2024) [35]	Real Time Analysis	Real Time Logs	Descriptive; lacks predictive risk modeling	Operational real-time monitoring
Yin (2023) [14]	Transformer strength in sequence modeling (non-EWS)	Response Seq	Knowledge tracing rather than engagement-based clickstream sequences	Transformer's seq outperforms

Other lines of research focus on dropout prediction using ensemble or classical ML models. Studies by [27][28][29][30] found that tabular models remain competitive when features are well-engineered. Across this literature, a consistent limitation emerges: few studies construct full daily sequences covering 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.

To situate this study within existing research, Table 3.1 summarizes recent works in learning analytics and temporal modeling, highlighting their methodological choices and temporal representations. The comparison shows that prior approaches predominantly rely on aggregated or coarse-grained temporal units, such as weekly summaries, fixed windows, or interaction-level features, and often lack explicit alignment with calendar days or full-sequence modeling across the course duration. Building on these observations, this study reframes early risk detection as a fine-grained temporal modeling problem by representing student engagement as complete daily behavioral sequences. Rather than introducing a novel architecture, the contribution lies in shifting the temporal resolution and problem formulation, enabling the examination of short-term behavioral dynamics over extended learning periods using GRU and Transformer models with validity masking and multi-threshold evaluation.

3. Methodology

The methodology is designed to preserve the temporal structure of student behavior while enabling a controlled comparison across multiple modeling paradigms. Instead of compressing interaction logs into static summaries, the pipeline maintains daily behavioral resolution across a fixed 150-day timeline. This design allows the models to learn not only how much students engage, but also when meaningful behavioral changes occur.

The workflow begins with raw CSV tables and culminates in out-of-fold predictions for cross-architecture evaluation. Unlike the earlier version of this study, the updated framework includes both sequence-based models and an aggregated tabular baseline to assess the incremental value of explicit temporal modeling. Fig. 3.1 illustrates the end-to-end pipeline, beginning with data integration, followed by daily feature construction, sequence tensorization, model training under stratified five-fold cross-validation, and post-hoc evaluation.

3.1 Data Preparation

Temporal preprocessing begins with aligning all student activity to a consistent daily timeline. OULAD encodes interaction time using integer offsets relative to the module start. For each module presentation, the minimum date offset is identified, and all rows are shifted such that the first active day corresponds to index 1. The result is a standardized

day index (day_idx) that supports comparison across different courses and time spans.

Because module durations vary in OULAD, a fixed temporal window is needed for modeling. Based on the empirical distribution of presentation lengths, this study sets a maximum sequence length of 150 days, which covers most course timelines, including the 4-, 6-, 8-, and 15-week instructional cycles reported by The Open University. This choice ensures that the temporal model receives sequences long enough to capture behavioral progression, while avoiding excessive sparsity beyond the maximum observed useful range.

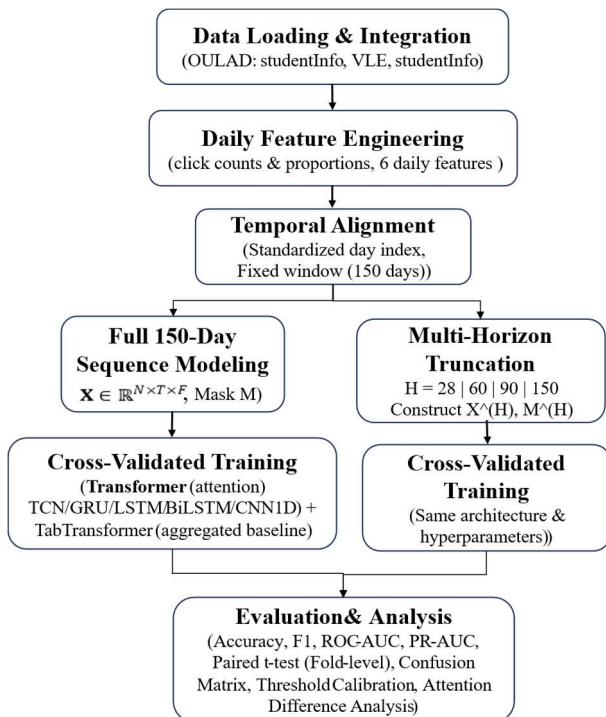


Figure 3.1 Methodology Pipeline for Daily-Sequence Early Warning Framework

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 $x_{t,c}$ represents the raw count of clicks for the category c on day t , while 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.

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

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

This representation captures both the intensity of engagement (through $\hat{C}_t$) and the relative distribution of effort across different activity types. Fig. 3.2 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.

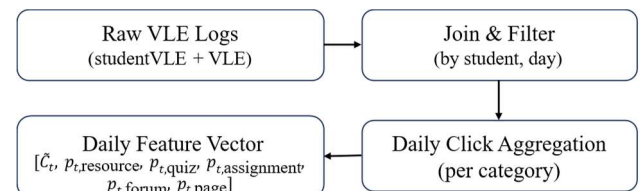


Figure 3.2 Daily feature construction from OULAD clickstream logs

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 N denotes 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: $H \in \{28, 60, 90, 150\}$

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:

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

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

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

Equations (8) and (9) restrict the effective input sequence to the first H observed 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 horizon H is therefore defined as:

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

Here, N denotes the total number of students, T represents 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 H represents 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 and to ensure that statistical summaries are computed only over observed days, thereby preventing bias introduced by 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 evaluate whether temporal ordering provides measurable benefits beyond static summaries, an aggregated feature representation is constructed from the same daily signals.

For each student, summary statistics, including mean engagement, maximum daily activity, total active days, and category-wise aggregates, are computed. These features are used to train a TabTransformer model under the same cross-validation splits. This design ensures that both sequence-based and aggregated models operate on the same population, enabling a fair comparison at the representation level.

3.6 Temporal Modeling Architectures

Six sequence-based architectures were selected to represent a broad spectrum of temporal inductive biases, enabling a structured evaluation of how

different modeling principles capture daily behavioral dynamics.

The Transformer Encoder was included for its ability to model long-range dependencies through masked self-attention, allowing the model to focus on critical risk days within the 150-day trajectory.

The Temporal Convolutional Network (TCN) was chosen to assess dilated convolution as a stable mechanism for capturing short-term behavioral shifts. GRU and LSTM were incorporated as established recurrent baselines designed to address vanishing gradients in sequential learning, while BiLSTM was added to examine whether bidirectional processing improves representation when full sequences are available. CNN1D was included as a simpler convolutional alternative to evaluate the performance limits of local pattern extraction without gating or attention mechanisms.

All models were trained under identical stratified five-fold cross-validation conditions to ensure that performance differences reflect architectural characteristics rather than data partitioning effects.

3.7 Transformer Encoder

Unlike other sequence models, the Transformer does not rely on recurrence. It processes all days in parallel through a self-attention mechanism that models long-range relationships directly. [23]. 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. Fig. 3.5 shows the Transformer architecture used in this work; the daily feature vectors are first projected into a model dimension and enriched with positional encoding to preserve chronological order. The sequence is then processed by stacked multi-head self-attention and feedforward layers, followed by temporal pooling to obtain a fixed-length representation for classification. The diagram highlights how attention weights are distributed across valid days, reflecting the model's capacity to emphasize informative temporal segments in the learning trajectory.

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.

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

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 relative and absolute positions, which is essential for reasoning over long sequences such as the 150 days used here.

Self-Attention

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

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

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).

$$\mathbf{h}_{mean} = \frac{1}{T} \sum_{t=1}^T \mathbf{o}_t, \mathbf{h}_{max} = \max_{1 \leq t \leq T} \mathbf{o}_t. \quad (13)$$

The pooled outputs are concatenated and passed to a sigmoid classifier.

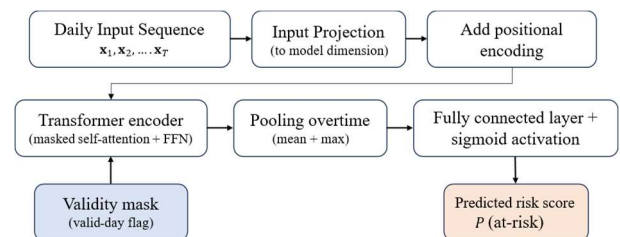


Fig. 3.5. Transformer encoder architecture using masked self-attention and temporal pooling to classify student risk from daily sequences

3.8 Model Training Procedure

All temporal and aggregated models are trained under identical experimental conditions to ensure controlled comparison. For each observation horizon $H \in \{28, 60, 90, 150\}$, truncated sequences are partitioned using stratified five-fold cross-validation, preserving the class distributions of at-risk and non-at-risk students across folds. Mini-

batch gradient descent with the Adam optimizer is employed, using a fixed batch size and consistent learning rate across architectures and horizons.

Model selection is performed based on the highest validation F1 score within each fold. Hyperparameters, including the learning rate, batch size, and number of training epochs, remain constant across models and horizons to isolate the effects of temporal evidence length and architectural design from tuning differences. This unified protocol enables fair comparison both across temporal architectures and across progressive observation horizons.

3.9 Statistical Significance Testing

To assess whether performance differences between observation horizons are statistically meaningful, paired statistical testing is 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 k under horizon H , where $k = 1, \dots, K$ and $K = 5$. The difference between two horizons is defined as:

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

The paired t-statistic is computed as:

$$t = \frac{\bar{d}}{s_d / \sqrt{K}} \quad (15)$$

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.

3.10 Evaluation Metrics

Multiple metrics are used to evaluate both predictive discrimination and operational usefulness. Accuracy reflects overall correctness but may be insufficient for moderately imbalanced early warning scenarios. Therefore, greater emphasis is placed on Precision, Recall, and the F1-score, which capture the trade-off between identifying at-risk students and limiting false alarms at a specific decision threshold. In addition, ROC-AUC and PR-AUC are reported to assess ranking performance across the full range of thresholds.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (17)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (18)$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (19)$$

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 4.1 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 large enough to support stable model training and robust out-of-fold evaluation, particularly for threshold-sensitive operational metrics such as recall and false-negative patterns.

Table 4.1. 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

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.

Table 4.2. 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

*Note: Acc = Accuracy; Prec = Precision

In addition to sequence-based models, an aggregated TabTransformer (TabTRF) baseline was evaluated using static summary features derived from the same student population. This baseline removes temporal ordering and serves to examine whether predictive improvements stem from advanced architectures or from preserving daily temporal structure itself. Table 4.2 and Table 4.3 summarize the out-of-fold performance across all evaluated models.

Table 4.3. 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

The results reveal two important patterns. First, the top-performing sequence models operate within a very narrow performance range. Differences in F1

scores among the strongest architectures are minimal, suggesting that the daily temporal representation itself contributes substantially to predictive performance. This supports the argument that preserving fine-grained engagement trajectories mitigates the information loss commonly introduced by weekly aggregation. Second, different architectures exhibit complementary strengths. The Transformer demonstrates the strongest ranking performance, achieving the highest Precision (0.936) and leading ROC-AUC (0.890) and PR-AUC (0.911). This indicates effective discrimination and reduced false positives when generating high-confidence alerts.

For balanced classification performance, the TCN achieves the highest F1 (0.793) and Recall (0.759), suggesting stronger coverage of at-risk learners. The GRU remains a competitive and stable baseline, offering a balanced trade-off between sensitivity and precision. In contrast, the CNN1D model performs lower overall, which may reflect limitations in capturing long-range temporal dependencies across the 150-day sequence.

The aggregated TabTransformer baseline exhibits a markedly different profile. While achieving extremely high precision (0.986), it shows substantially lower recall (0.605) and reduced F1 (0.750). This pattern indicates that static aggregation tends to identify only the most evident risk cases while missing a considerable portion of at-risk students whose behavioral shifts are detectable only through temporal dynamics. This comparison highlights the added value of preserving daily temporal structure compared to static aggregation.

4.3 Attention Difference Analysis

To further investigate how the Transformer differentiates between at-risk and non-risk learners, the attention weights from the last encoder layer were aggregated across the entire population in each class (13,846 at-risk and 15,382 non-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 the average attention received by the non-risk group. Fig 4.1 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.

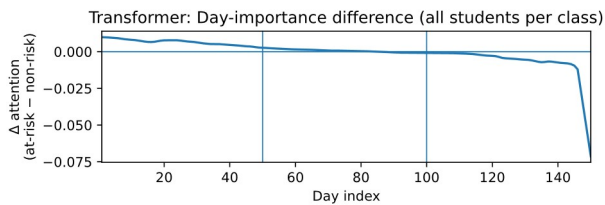


Figure 4.1. Smoothed trajectory of Δ attention (at-risk minus non-risk) across the 150-day sequence, illustrating the temporal direction and magnitude of attention differences between classes

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.

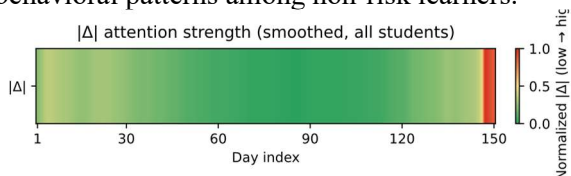


Figure 4.2. Normalized absolute attention difference $|\Delta|$ between at-risk and non-risk students

Fig. 4.2 illustrates the normalized absolute attention difference $|\Delta|$ 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 reflects how strongly the Transformer distinguishes between the two groups in terms of attention allocation.

The results show that most mid-semester days exhibit relatively low $|\Delta|$ values, indicating comparable attention patterns across classes during this period. In contrast, the magnitude of $|\Delta|$ 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, Fig. 4.1 and Fig. 4.2 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 Table 4.2 and Table 4.3 provide a summary of scalar performance metrics at the default classification threshold, a more comprehensive understanding of model behavior requires examining performance across the full decision spectrum. To this end, ROC and PR curves were generated using out-of-fold prediction scores from all evaluated models. Fig. 4.3 presents the combined ROC curves for the six sequence-based architectures and the aggregated TabTransformer baseline.

The ROC analysis reveals that sequence-based models consistently maintain stronger discrimination across the entire threshold range. The Transformer achieves the highest ROC-AUC (0.890), closely followed by GRU (0.887), TCN (0.884), LSTM (0.884), and BiLSTM (0.884). The CNN1D and TabTransformer models exhibit lower ROC-AUC values (0.854 and 0.854 respectively), indicating weaker global ranking ability. Importantly, the ROC curves of the sequence models remain visibly separated from the diagonal reference line across most operating regions, suggesting robust separability between at-risk and non-risk learners.

However, in early warning systems where class imbalance is present and intervention resources are limited, Precision-Recall behavior is often more informative. Fig. 4.4 illustrates the combined PR curves.

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

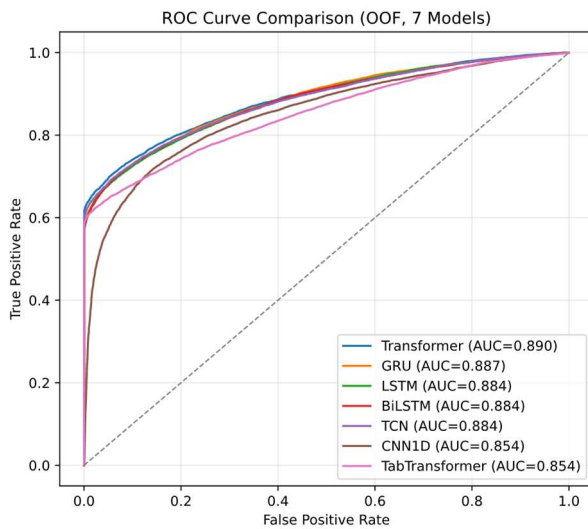


Figure 4.3. Combined ROC curve comparison across sequence models and aggregated TabTransformer baseline (OOF predictions)

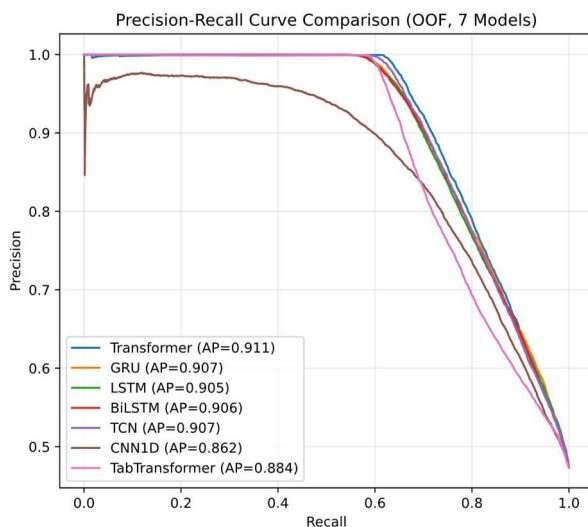


Figure 4.4. Combined Precision-Recall curve comparison across sequence models and aggregated TabTransformer baseline (OOF predictions)

Taken together, Fig. 4.3 and Fig. 4.4 reinforce two important findings. First, daily sequence modeling consistently improves ranking discrimination compared to aggregated representations. Second, the advantage of sequence models becomes particularly visible when recall increases, suggesting that temporal dynamics contribute meaningfully to identifying learners whose risk patterns evolve progressively. The following section further examines how this

ranking behavior evolves as the observation horizon is progressively truncated.

4.5 Multi-Horizon Performance Evaluation

To assess early-warning feasibility with limited temporal evidence, model performance was evaluated using progressively truncated observation horizons $H \in \{28, 60, 90, 150\}$, 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.

Table 4.4 and Table 4.5 report the mean cross-validated performance across horizons for the Transformer model. Fold-level metrics are aggregated using the same five stratified splits.

Table 4.4. 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

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.

Table 4.5 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

Fig. 4.5 visualizes the relationship between PR-AUC and the observation horizon for the Transformer and GRU models. As the horizon increases, both models demonstrate systematic improvement consistent with progressive temporal evidence accumulation. The aggregated baseline is

computed once using full 150-day summary features and therefore does not vary with the truncation horizon, resulting in a constant performance level across horizon settings. This contrast highlights the sensitivity of sequence-based models to temporal accumulation effects.

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.

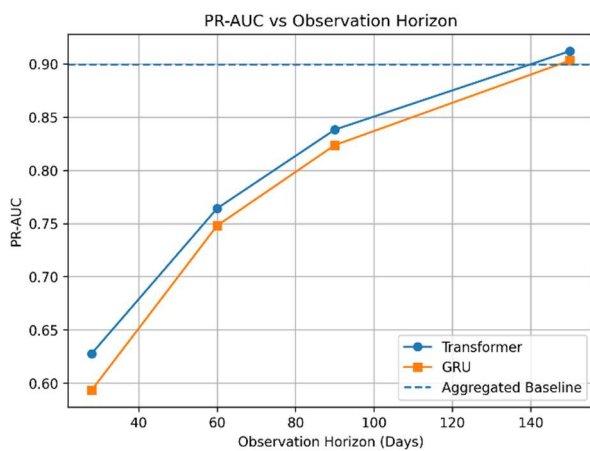


Figure 4.5 PR-AUC progression across observation horizons.

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.

4.6 Error Structure and Operational Implications

To further examine deployment implications, confusion matrices were analyzed at the default classification threshold of 0.5 using the full observation horizon ($H = 150$), where models achieve their strongest discrimination. The results are summarized in Table 4.6.

The comparison reveals distinct operational profiles. The Transformer produces the fewest false positives (645), reflecting its precision-oriented behavior and reduced risk of unnecessary interventions. In contrast, the TCN achieves the lowest false-negative count among sequence

models (3,339), indicating stronger sensitivity in identifying at-risk learners. The GRU demonstrates a balanced trade-off between precision and recall, maintaining moderate false-positive and false-negative counts.

The aggregated TabTransformer baseline exhibits a markedly different structure. Although it yields an extremely low number of false positives (123), its false-negative count increases substantially (5,471). This pattern suggests that static aggregation primarily captures highly pronounced risk cases while overlooking students whose behavioral decline develops gradually over time.

From an institutional perspective, these differences imply alternative intervention strategies. Precision-oriented models generate smaller, high-confidence alert lists, suitable when intervention resources are limited. Recall-oriented models expand risk coverage but increase monitoring overhead. The ability of sequence-based models to maintain both controlled false-positive rates and improved recall supports the practical value of preserving daily temporal structure.

Table 4.6 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

4.7 Threshold Calibration and Deployment Scenarios

While the previous sections evaluated model performance at the default classification threshold of 0.5, early warning systems in real educational settings often require calibration based on institutional capacity and intervention priorities. A fixed threshold may not align with operational constraints such as limited counseling resources or the need to minimize false alarms. Therefore, threshold sensitivity analysis was conducted using the Temporal Convolutional Network (TCN) model, which demonstrated the strongest balanced performance in terms of F1 and Recall. Fig. 4.6 presents the precision-recall curves of the TCN as a function of the classification threshold.

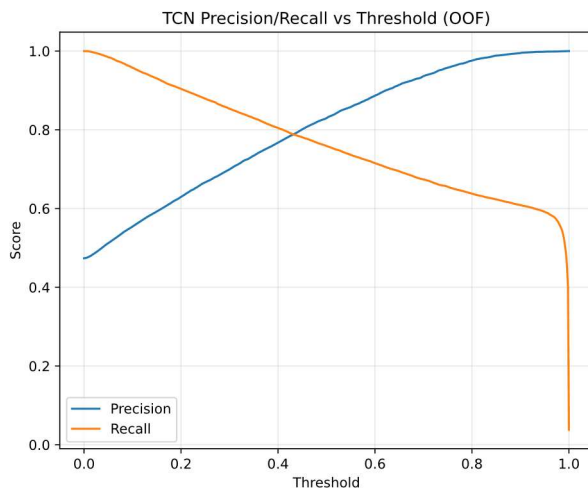


Figure 4.6. Precision and Recall as a function of decision threshold for the TCN model (OOF predictions)

Fig. 4.6 shows a clear trade-off between sensitivity and precision. At lower thresholds, recall increases substantially, allowing broader identification of at-risk students, but at the cost of reduced precision and a higher number of false positives. As the threshold increases, precision improves while recall declines, resulting in a smaller but more reliable alert list. Around the default threshold of 0.5, the TCN achieves a balanced configuration ($F1 = 0.793$; Recall = 0.759), indicating effective coverage without excessive false alarms.

To further examine operational implications, false positive rate (FPR) and false negative rate (FNR) were analyzed across threshold values. Fig 4.7 shows the curves reveal that lowering the threshold reduces the false negative rate, thereby increasing the number of at-risk students correctly identified. However, this improvement comes at the cost of a rising false positive rate. Conversely, increasing the threshold reduces false positives but increases the proportion of at-risk students missed. This trade-off highlights the importance of context-specific calibration rather than reliance on a universal cutoff.

From a deployment perspective, institutions with limited intervention resources may prefer a higher threshold to generate smaller, high-confidence alert lists. In contrast, institutions prioritizing early intervention and broader student coverage may adopt a lower threshold to reduce missed cases. The TCN's flexibility across threshold settings demonstrates that daily sequence modeling not only improves predictive performance but also provides adaptable risk-

monitoring mechanisms aligned with practical institutional needs.

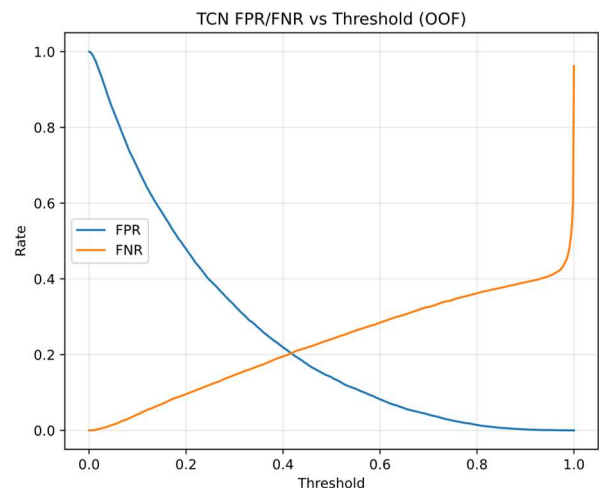


Figure 4.7. FPR and FNR across decision thresholds for the TCN model.

5. Conclusion

This study demonstrates that preserving daily temporal resolution substantially strengthens early detection of at-risk students in online learning environments. By representing each learner as a complete 150-day behavioral trajectory with six normalized daily features, the proposed framework captures engagement dynamics that are inherently compressed or obscured in aggregated representations.

Experimental evaluation across six sequence-based architectures, including Transformer, GRU, LSTM, BiLSTM, TCN, and CNN1D, shows that all temporal models perform within a consistently narrow range under identical validation conditions. This stability indicates that predictive gains arise primarily from the daily sequence representation rather than from architectural complexity alone. Among the evaluated models, the Transformer achieves the strongest global ranking performance, while the TCN and GRU provide higher sensitivity and more balanced classification behavior, making them suitable for broader risk coverage scenarios.

Multi-horizon analysis further reveals that predictive performance increases progressively as longer behavioral histories are observed, with statistically significant improvements between early and full observation settings. This pattern confirms that temporal evidence accumulates over time and strengthens risk discrimination as engagement trajectories unfold.

In contrast, the aggregated TabTransformer baseline exhibits substantially lower recall despite extremely high precision. This discrepancy confirms that static aggregation fails to capture

gradual behavioral shifts that evolve over time and are critical for early-stage intervention.

Taken together, these findings provide clear empirical evidence that modeling fine-grained temporal dynamics offers measurable advantages over aggregated approaches. Daily sequence modeling, therefore, represents a more reliable and operationally flexible foundation for next-generation early warning systems in digital education contexts.

Despite these contributions, **several limitations should be noted. The evaluation is conducted on a** single benchmark dataset; external validation using institutional LMS data would strengthen generalizability. The study focuses on binary risk prediction without measuring actual intervention outcomes, and the attention-based analysis is descriptive rather than causal. In addition, hyperparameters are intentionally kept consistent across models to isolate representation effects, thereby limiting architecture-specific optimization. Future work may extend this framework through cross-institutional validation and integration with adaptive intervention strategies.

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] R. Boudjehem and Y. Lafifi, "An early warning system to predict dropouts inside e-learning environments," *Educ. Inf. Technol.*, vol. 29, pp. 16365–16385, Feb. 2024, doi: 10.1007/s10639-024-12498-1.
- [3] M. Saqr, J. Nouri, H. Vartiainen, and J. Malmberg, "What makes an online problem-based group successful? A learning analytics study using social network analysis," *BMC Med. Educ.*, vol. 20, no. 1, p. 80, 2020, doi: 10.1186/s12909-020-01997-7.
- [4] K. Wang, "Academic Early Warning Model for Students Based on Big Data Analysis," *Int. J. Emerg. Technol. Learn.*, 2023, doi: 10.3991/ijet.v18i12.41087.
- [5] 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, p. 104468, 2022, doi: 10.1016/j.compedu.2022.104468.
- [6] H. Alhakbani, "Open learning analytics: A systematic review of OULAD benchmark studies," in *Proceedings of an International Conference on Learning Analytics*, 2022. doi: <https://doi.org/10.1145/3529399.3529413>.
- [7] K. A. Laksitowening, "Learning analytics and business intelligence for academic early warning," *Khazanah Inform.*, 2024, doi: 10.23917/khif.v10i2.4142.
- [8] M. Saqr, S. López-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, p. 104787, 2023, doi: 10.1016/j.compedu.2023.104787.
- [9] J. Y. Wu, "Temporal modeling of before-class engagement and performance," *Comput. Educ.*, 2024, doi: 10.1016/j.compedu.2023.104975.
- [10] Á. Hernández-García, C. Cuenca-Enrique, L. Del-Río-Carazo, and S. Iglesias-Pradas, "Exploring the relationship between LMS interactions and academic performance: A Learning Cycle approach," *Comput. Human Behav.*, vol. 155, p. 108183, 2024, doi: 10.1016/j.chb.2024.108183.
- [11] O. R. Yürüm, T. Taşkaya-Temizel, and S. Yıldırım, "The use of video clickstream data to predict university students' test performance," *Educ. Inf. Technol.*, vol. 28, pp. 5209–5240, 2023, doi: 10.1007/s10639-022-11403-y.
- [12] 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, p. 16251, 2025, doi: 10.1038/s41598-025-00256-3.
- [13] N. U. R. Junejo *et al.*, "SAPPNet: students' academic performance prediction during COVID-19 using neural network," *Sci. Rep.*, vol. 14, no. 1, p. 24605, 2024, doi: 10.1038/s41598-024-75242-2.
- [14] Y. Yin *et al.*, "Tracing Knowledge Instead of Patterns: Stable Knowledge Tracing with Diagnostic Transformer," in *Proceedings of the ACM Web Conference 2023*, in WWW '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 855–864. doi: 10.1145/3543507.3583255.
- [15] H. Wang, "Transformer-based deep learning for adaptive pedagogy under uncertain student preferences," *Sci. Rep.*, vol. 15, 2025, doi: 10.1038/s41598-025-25996-0.

- [16] 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, Mar. 2025, doi: 10.1007/s44163-025-00252-6.
- [17] 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.
- [18] Q. Huang, "Enhancing Academic Performance Prediction With Temporal Graph Networks for Massive Open Online Courses," *J. Big Data*, 2024, doi: 10.1186/s40537-024-00918-5.
- [19] 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, p. 104581, Jul. 2022, doi: 10.1016/j.compedu.2022.104581.
- [20] 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.
- [21] 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. Volume 13-2022, 2023, doi: 10.3389/fpsyg.2022.1096337.
- [22] K. Cho *et al.*, "Learning Phrase Representations using {RNN} Encoder{--}Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP})*, A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. doi: 10.3115/v1/D14-1179.
- [23] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.
- [24] 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, 2022, doi: 10.3390/educsci13010017.
- [25] 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.
- [26] 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, p. 100066, 2022, doi: 10.1016/j.caeai.2022.100066.
- [27] 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.
- [28] D. Bañeres, M. Rodríguez, A.-E. Guerrero, and P. Guasch, "An Early Warning System to Identify and Intervene Online Dropout Learners," *Int. J. Educ. Technol. High. Educ.*, vol. 20, p. 3, 2023, doi: 10.1186/s41239-022-00371-5.
- [29] N. Barz, M. Benick, L. Dörrenbächer-Ulrich, and F. Perels, "Fostering self-regulated learning using synchronous or asynchronous digital learning environments: a latent profile analysis of pre-service teachers' individual differences," *Front. Educ.*, vol. 9, Dec. 2024, doi: 10.3389/feduc.2024.1445182.
- [30] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technol. Soc.*, vol. 76, p. 102474, 2024, doi: 10.1016/j.techsoc.2024.102474.
- [31] J. C.-Y. Sun, Y.-H. Chen, and K.-P. Yeh, "Temporal learning analytics to explore traces of self-regulated learning behaviors in an asynchronous online course," *Front. Psychol.*, vol. 14, p. 1144303, 2023, doi: 10.3389/fpsyg.2023.1144303.
- [32] Z. Shou, "Predicting Student Performance in Online Learning: A Multidimensional Time-Series Data Analysis Approach," *Appl. Sci.*, 2024, doi: 10.3390/app14062522.
- [33] A. Z. Kusuma, M. Hatim, and Y. Uzer, "Need Analysis of English Language Used for Students in Hospitality Industry Department," *Esteem J. English Educ. Study Program.*, 2023, doi: 10.31851/esteem.v6i1.10168.

- [34] 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, p. 65, 2025, doi: 10.1007/s13278-025-01484-1.
- [35] L. Uguina-Gadella, I. Estévez-Ayres, J. A. Fisteus, C. Alario-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. 1754–1764, 2024, doi: 10.1109/TLT.2023.3331433.
- [36] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994, doi: 10.1109/72.279181.