Communications in Mathematics and Applications
Vol. 17, No. 1, pp. 207–222, 2026
ISSN 0975-8607 (online); 0976-5905 (print)
Published by RGN Publications
DOI: 10.26713/cma.v17i1.3339

http://www.rgnpublications.com

Research Article

Interpretable Sequence Modeling of Educational Behavior using Temporal Attention Networks

Rakan Saad Alotaibi*, Fahad Mazyed Alotaibi, Sameer Abdullah Nooh and Abdulaziz A. Alsulami

Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
***Corresponding author:** ralotaibi0331@stu.kau.edu.sa

Received: August 26, 2025 **Revised:** November 29, 2025 **Accepted:** December 19, 2025

Abstract. The increasing popularity of online learning platforms has led to vast amounts of sequential data regarding learner behavior. Available predictive models tend to focus on fixed features. They cannot pinpoint the dynamic temporal changes in learning activity, thereby reducing the effectiveness of predictive methods and making them more challenging to understand. In this work, the *Dynamic Temporal Attention Network* (DTAN) is proposed. This novel deep learning architecture learns e-learning behavior using time-aware attention and temporal convolution to enhance predictive accuracy and interpretability. TCN has already been combined with two significant attention modules. The *Attention-over-Time windows* (ATW) and the *Dynamic Contextual Attention Mechanism* (DCAM). With such components, the model can learn short and long-term dependencies on the learner's behavior and adaptively prioritize critical time slots. To train and evaluate the model, two large-scale datasets are used: EdNet, containing over 130 million question and answer interactions in the K-12 context, and OULAD, an exam-taking dataset in the university context. By outperforming state-of-the-art models, including *long short-term memory* (LSTM) and *gated recurrent units* (GRU), as well as standard TCNs, TAN achieves significant improvements across diverse classification tasks. It has strong early prediction skills and an interpretable visualization that focuses on weight, highlighting critical incidents in a growing learner journey. Such observations are essential in the process of providing individual instruction and necessary intervention. DTAN delivers an interpretable solution for sequential data modeling in education, significantly boosting efficiency, particularly in adaptive learning systems.

Keywords. Dynamic Temporal Attention Network (DTAN), Educational data mining, Learning analytics, Temporal Convolutional Networks (TCN), Attention mechanisms, Sequence modeling, Student performance prediction

Mathematics Subject Classification (2020). 68T07, 68T05, 62M45

1. Introduction

E-learning platforms produce detailed traces of sequential activity, such as watching videos, attempting quizzes, revisiting materials, and engaging in discussions, and this can be tracked over time (Romero and Ventura [16]). When modeled accordingly, such time-stamped interactions become an illustration of patterns of engagement, knowledge retention, and subsequent learning attainment. Traditional predictive pipelines in educational data mining have frequently operated on static feature sets or aggregate descriptions (i.e., they only observe a single temporal occurrence) (Alnasyan *et al*. [1]). The learning process is less dynamic than it seems, as students have varied learning rates, re-examine previous material, and behavior is nonstationary. Modeling of the temporal dependence of data, thus, demands models tailored explicitly to such data (Xie *et al*. [20]). More recent developments in deep sequence modeling, such as RNNs, LSTMs, and *Temporal Convolutional Networks* (TCNs), have increased accuracy in outcome prediction by training on ordered interaction logs (Piech *et al*. [15], and Xu *et al*. [21]). However, these architectures have trouble with long-range connections (e.g., vanished gradients or fixed memory spans) or with modeling all time points as equally informative, which dulls their performance in focusing on critical points of learning (Hochreiter and Schmidhuber [8]). Attention mechanisms overcome this shortcoming by dedicating greater weights to the most pertinent parts of a sequence and have already been relocated successfully, in the NLP domain, to time-series modeling, with beneficial effects on accuracy and interpretability (Vaswani *et al*. [19]). Nevertheless, attention remains underutilized in learning analytics, particularly when it comes to supporting multi-scale thinking, such as the need to take the intensity of individual bursts of activity and put it into a broader context, and to support explanations that can be used in instructions (Koedinger *et al*. [9]).

To fill this gap, we propose a sequence model, the *Dynamic Temporal Attention Network* (DTAN), that combines a TCN backbone with two time-aware attention modules. The former, *Attention-over-Time Windows* (ATW), focuses on local time windows to garner short-term cues (e.g., a streak of practice or a lapse). The second, a Dynamic Contextual Attention Mechanism (DCAM), re-weights longer-range context adaptively, ensuring that long-distance yet meaningful behaviors are not diluted. Because of this design, DTAN produces explainable attention maps over time without sacrificing good temporal feature extraction, and this has the benefit of an accurate, but still explainable, model to all interested educational stakeholders (Zhou and Kang [24]).

We evaluate DTAN on two public, real-world datasets: EdNet and OULAD, which represent complementary e-learning settings. Across tasks, we compare against established sequence baselines (LSTM, GRU, TCN) and assess both overall performance and early-prediction capability, which is central for timely intervention. Beyond accuracy-oriented metrics, we visualize attention distributions to surface actionable insights for instructors and course designers:

- This study presents DTAN, integrating TCNs with dynamic, multi-scale attention to model both short-term actions and long-term trends in learner behavior.
- We design an attention framework, ATW (local windows) and DCAM (context-adaptive, more extended range), that assigns adaptive importance to time steps while supporting interpretability.
- We conduct an empirical study on EdNet and OULAD, benchmarking DTAN against LSTM/GRU/TCN baselines and analyzing early-warning performance for intervention.

- We provide attention-based visualizations that make model decisions transparent and pedagogically meaningful, aligning with the need for explainable analytics in education

The rest of the paper is structured as follows: Section 2 reviews related work. Section 3 outlines the proposed methodology. Section 4 presents experimental results and discussion. Finally, Section 5 concludes the study and outlines directions for future research.

2. Related Work

Sequential modeling for student behavior. Early work in educational data mining often aggregated logs into static features (e.g., total time-on-task, counts of attempts), which obscured the temporal structure of learning behavior (Peña-Ayala [13]). Sequence-aware modeling began with HMMs/BKT, which track latent mastery as learners progress through discrete steps (Corbett and Anderson [4]) while influential, these probabilistic models struggle to represent high-dimensional behaviors at scale. Deep sequence models—RNNs, GRUs, and LSTMs, advanced the field by learning correlations from ordered interaction streams (clicks, quiz histories, forum actions) and have been applied to outcomes such as dropout and performance prediction (Cho *et al*. [3], and Piech *et al*. [15]). However, recurrent architectures can suffer from vanishing gradients and efficiency limits when processing long sequences. *Temporal Convolutional Networks* (TCNs) address these issues via dilated causal convolutions, enabling parallelism and large receptive fields, and have shown strong results on educational time series; yet, by default, they tend to weight all time steps uniformly, under-emphasizing pivotal events (Xu *et al*. [21]). Attention for temporal salience and interpretability. Attention mechanisms reweight sequence elements by relevance, improving long-range reasoning and offering model interpretability. Initially developed in NLP, self-attention/Transformer layers have been adapted to time-series applications across domains (Lim and Zohren [11], Shickel *et al*. [17], and Vaswani *et al*. [19]). In learning analytics, attention has been used to highlight key moments in a learner's history, for example, behaviors predictive of dropout or completion—thereby improving transparency for instructors (He *et al*. [7], and Zhang *et al*. [22]). Models such as SAINT+ demonstrate how attention across questions and concepts can strengthen knowledge tracing (Shin *et al*. [18]). Still, many educational applications either fix attention scope or do not impose temporal structure, which can create context mismatches for inherently ordered learning processes (Shin *et al*. [18], Vaswani *et al*. [19], and Zhang *et al*. [22]).

Hybrid models of temporal and attention, and the remaining gaps. Leveraging attention and temporal backbones enables the extraction of local features and prioritization of the global context. Graph- and network-based methods have also been used to investigate the focus on surface salient structures and patterns of behavior in education (Alqahtani [2], Ghaoui *et al*. [6], and Mir *et al*. [12]). However, two deficiencies still prevail: (i) lack of explanatory tooling on short-term scope (i.e., concentrated practice) and long-term course (i.e., gradual disengagement); and (ii) limited explanatory tooling for stakeholders that remains faithful to the learned signal (He *et al*. [7], Shickel *et al*. [17], and Zhang *et al*. [22]). Our contribution fills these gaps by augmenting a TCN backbone with dynamic time-aware attention that operates on smaller windows and long-range context, in pursuit of a dual goal: enhancing predictive performance and interpretability in educational sequences. A brief comparison of DTAN with respect to aggregate, HMM/BKT, RNN/GRU/LSTM, TCN, and Transformer-based methods, based on long-range modeling, multi-scale emphasis, and interpretability, is provided in Table 1.

Table 1. Positioning of this study (DTAN) relative to prior work

Model family (references)	Long-range temporal modeling	Adaptive multi-scale temporal focus	Educational interpretability emphasis
Aggregates [9]	☒	☒	☒
HMM/BKT [4]	Δ (short horizons)	☒	Δ (model-based, low-dimensional)
RNN/GRU/LSTM [3, 15]	Δ (limited by recurrence)	☒	☒/Δ
TCN [21]	☑	☒	☒
Transformer/SAINT+ [18, 19, 22]	☑	Δ (global, often not time-aware)	Δ
Temporal attention in LA [7, 17, 22]	Δ	Δ (fixed scope)	☑

3. Proposed Methodology

The proposed methodology combines domain-specific language modeling and attention-enhanced deep learning to forecast job performance from e-learning feedback. The whole data flow is shown in Figure 1.

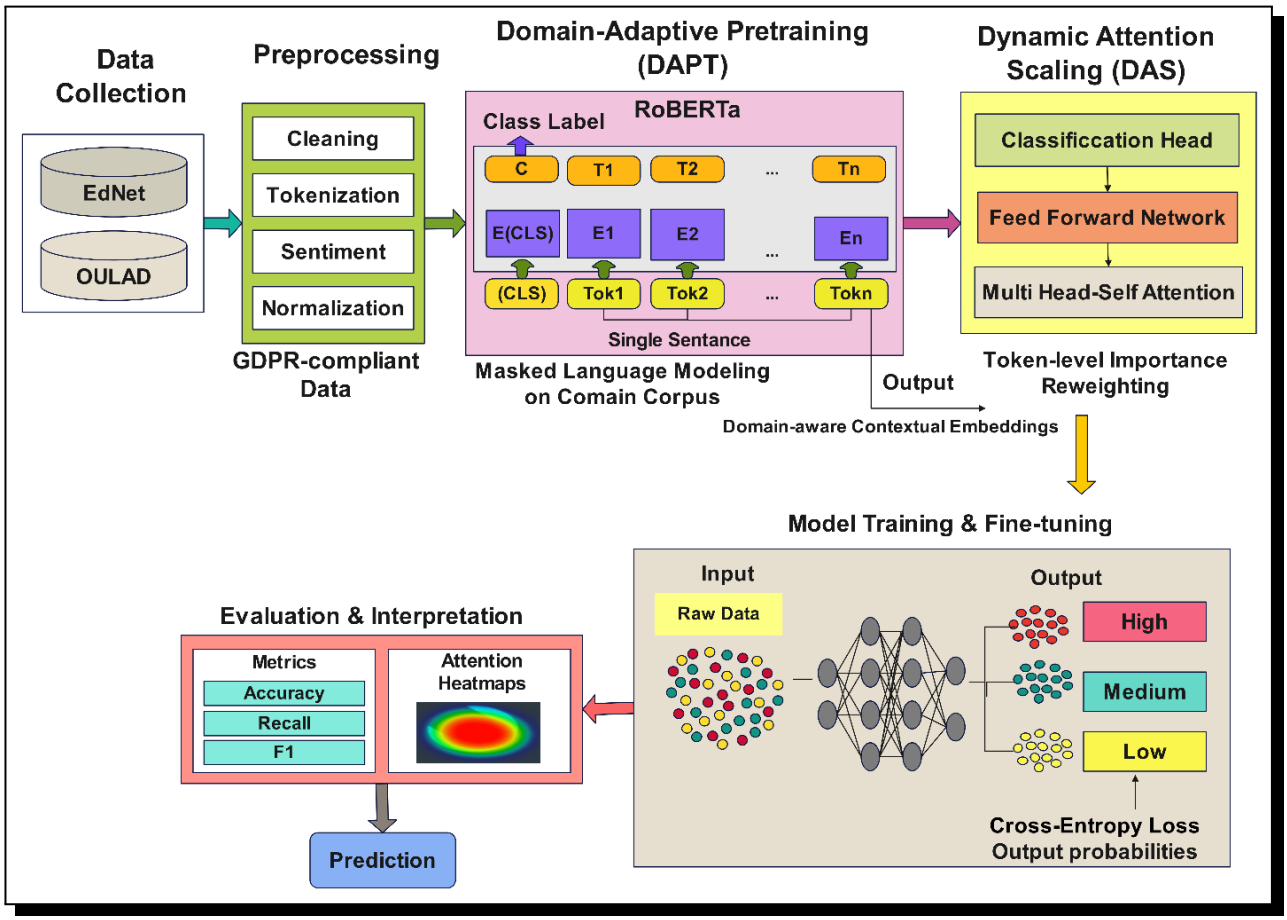


Figure 1. Proposed methodology for interpretable e-learning performance prediction using RoBERTa and dynamic attention scaling

This study uses two public datasets: EdNet[1], with about 131 million question-answer interactions from a mobile learning app, and OULAD (Kuzilek *et al*. [10]), with activity logs, assessments, and demographics. Perform cleaning, tokenization, and normalization, then derive temporal patterns and sentiment features from textual feedback. Apply domain adaptive pretraining with RoBERTa using masked language modeling on the training corpus to learn domain-specific representations. Add a Dynamic Attention Scaling layer to reweight token importance, emphasizing performance-indicative patterns and down weighting nonessential input. Fine-tune the E RoBERTa model on labeled sequences of learner outcomes, with a Softmax head predicting Low, Medium, or High performance. Evaluate using accuracy, precision, recall, and F1 score, and visualize attention heatmaps plus token importance scores to support interpretability for instructors.

3.1 Datasets

To evaluate the proposed DTAN, we used two large-scale, publicly available datasets from diverse educational settings: EdNet[1], and the OULAD (Kuzilek *et al*. [10]). EdNet, released by KAIST with Santa, is the largest public dataset for K-12 mathematics education. The corpus contains more than 131 million interaction records from about 784,000 students using a mobile learning platform. Each record stores the timestamp, question identifier, correctness, response time, and knowledge concept tags. For modeling, interactions are grouped by user and segmented into sessions with a maximum sequence length of 100. Sequences are filtered to remove null or outlier entries, and labels are derived from recent streaks of correctness and time-weighted accuracy. Binary correctness also serves as a proxy for moment to moment learning success. OULAD captures engagement for 32,593 students across 22 online courses, including more than 10 million virtual learning environment clicks, demographic attributes, enrollments, and performance outcomes. Each engagement event is time stamped and categorized by activity type, such as quizzes, forums, or resource views. Weekly summaries form chronological learning trajectories. Final academic status labels, Pass, Fail, Withdraw, or Distinction, define a multiclass classification task. Records with missing data are removed, and engagement counts are standardized using z-score normalization.

3.2 Feature Engineering and Labeling

Effective sequential modeling of learning behavior requires meaningful feature extraction from raw interaction logs. The EdNet and OULAD datasets were processed to generate temporally ordered sequences enriched with behavioral, contextual, and performance-based indicators.

3.2.1 Feature Construction

Represent the learning history of student s as

$$S_s = [x_1, x_2, \ldots, x_T] \tag{1}$$

with feature vector $x_t \in \mathbb{R}^d$ at time t.

3.2.2 EdNet Features

One hot question identifier, binary correctness $c_t \in \{0,1\}$, elapsed time $\Delta_t = \tau_t - \tau_{t-1}$ where τ_t denotes the timestamp, categorical knowledge tag, and rolling accuracy over a window w:

$$a_t = \frac{1}{w} \sum_{i=t-w+1}^{t} c_i \,. \tag{2}$$

[1]D. Shin and S. Lee, *EdNet*, Github, (2020), URL: https://github.com/riiid/ednet.

3.2.3 OULAD Features

Total clicks κ_t, categorical activity type, cumulative assessment score σ_t, and standardized engagement

$$e_t = \frac{\kappa_t - \mu_\kappa}{\sigma_\kappa} \tag{3}$$

with μ_κ computed across students.Each sequence is padded or truncated to a fixed length L for batch training.

3.2.4 Label Construction

EdNet short horizon mastery label from the final k interactions:

$$y = \begin{cases} 1, & \text{if } \sum_{i=T-k}^{T} c_i \geq \tau, \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

with mastery threshold τ. The OULAD final outcome label

$$y \in [0,1,2,3] \tag{5}$$

with 0 = Fail, 1 = Pass, 2 = Distinction, 3 = Withdraw. For binary analyses, collapse to Pass and Fail.Use an eighty percent training split, ten percent validation split, and ten percent test split with stratification.

3.3 Temporal Convolutional Network Layer

TCNs are an efficient alternative to recurrent architectures for modeling sequential data. Unlike RNNs or LSTMs, TCNs allow parallel computation and long-range dependency modeling through causal, dilated convolutions and residual connections. These features make TCNs particularly effective for processing learning behavior sequences where both recency and temporal context are essential. Given an input sequence of feature vectors using eq. (6):

$$X = [x_1, x_2, \ldots, x_T], \quad x_t \in \mathbb{R}^d. \tag{6}$$

The TCN processes this sequence using a stack of 1D convolutional layers with causal structure, meaning that the output at time step t only depends on current and past inputs, i.e., $[x_1, x_2, \ldots, x_t]$. Let $F^{(l)}$ represent the output of the lth TCN layer, and k be the filter size. Then, for each layer by using eq. (7):

$$F_t^{(l)} = \sum_{i=0}^{k-1} W_i^{(l)} \cdot F_{t-d\cdot i}^{(l-1)} + b^{(l)}, \tag{7}$$

where d is the dilation factor, which increases exponentially at each layer ($d = 2^{l-1}$), $W_i^{(l)}$ are the learned convolutional weights, $b^{(l)}$ is the bias term.

The receptive field of the TCN grows exponentially with depth, allowing the model to capture long-range dependencies by using eq. (8):

$$\text{Receptive Field} = (k-1)\sum_{l=0}^{L-1} 2^l + 1. \tag{8}$$

To enhance stability during training, residual connections are applied by using eq. (9):

$$F^{(l)} = \text{ReLU}(F^{(l)} + X). \tag{9}$$

This ensures the model can learn identity mappings and reduces vanishing gradient issues in deep TCN stacks.

Algorithm 1 presents the Temporal Convolutional Block for extracting hierarchical temporal features using dilated convolutions and residual connections.

Algorithm 1. Temporal Convolutional Block

Input: Sequence $X = [x_1, x_2, \ldots, x_T]$

Output: Encoded sequence $F = [f_1, f_2, \ldots, f_T]$

For each layer l in $[1, \ldots, L]$ do

Compute dilated convolution with filter $W^{(l)}$ and dilation $d = 2^{(l-1)}$

Apply ReLU activation and dropout

Add residual connection if dimensions match

Set $F^{(l)}$ = Output of layer l

End For

Return Final output $F = F^{(L)}$

Figure 2 illustrates a three-layer TCN. Each layer performs a 1D dilated causal convolution, exponentially increasing dilation rates $(1, 2, 4)$.

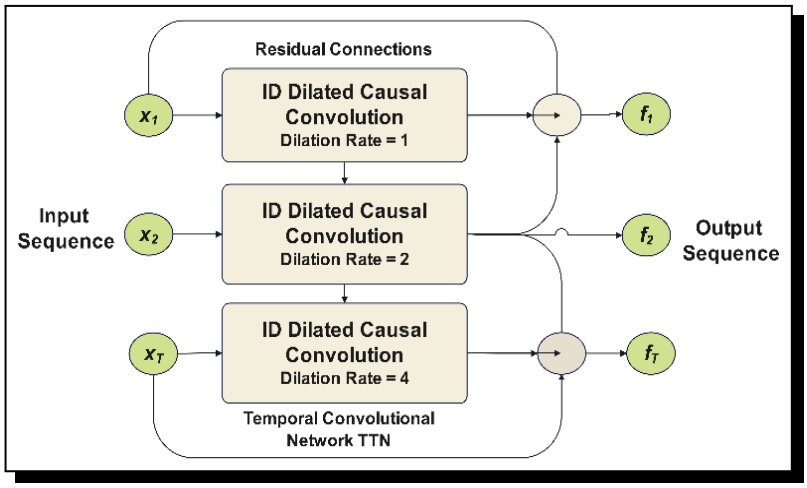


Figure 2. TCN layer architecture

Residual skip connections attach each layer to its input. The input sequence (e.g., student activity vectors at a time) is entered on the left into the network and flows out on the right. The outputs of each node were coded as information weighted over time in previous steps, thus enabling the model to learn temporal dependencies of both short and long range.

TCNs are especially amenable to the modeling of e-learning sequences since they do not require recursion, easily capture the existence of long histories, and retain sequence length information. Their parallelizable model makes them faster to train, and when applied to behavioral data modeling, high predictive capacity is maintained.

3.4 Attention-over-Time Windows (ATW) Module

The ATW module is designed to improve the interpretability and accuracy of temporal modeling by allowing the network to focus selectively on the most relevant segments of the input sequence. Unlike standard attention mechanisms that operate over all time steps uniformly, the ATW module divides the sequence into fixed-length temporal windows. It applies attention weights

within and across these segments. This structure reflects how learners often show bursts of meaningful activity (or disengagement) during specific periods. Let $X = [x_1, x_2, \ldots, x_T]$ be the encoded sequence output from the TCN layer, where $x_t \in \mathbb{R}^d$. The sequence is partitioned into M overlapping or non-overlapping windows $[W_1, W_2, \ldots, W_M]$, each of length w. Each window W_i contains vectors $[x_{(i-1)w+1}, \ldots, x_{iw}]$. For each window, we compute a context vector c_i using a soft attention mechanism by using eqs. (10) and (11):

$$\alpha_{i,t} = \exp(e_{i,t}) \sum_{j=1}^{w} \exp(e_{i,j}), \tag{10}$$

where $e_{i,t} = v^\top \tanh(W_{x_t} + b)$,

$$c_i = \sum_{t=1}^{w} \alpha_{i,t} \cdot x_t, \tag{11}$$

where

- v, W, and b are learnable parameters,
- $\alpha_{i,t}$ is the attention weight assigned to the position t within window i,
- c_i is the aggregated context vector for that window.

Once all window-level vectors $[c_1, \ldots, c_M]$ are computed, they are merged (e.g., via concatenation or averaging) and passed to the next stage of the model (e.g., classification head or higher-level attention). Algorithm 1 presents ATW for computing context-aware representations by applying attention within fixed-length temporal segments.

Algorithm 2. Attention-over-Time Windows

Input: Encoded sequence $X = [x_1, x_2, \ldots, x_T]$, window size w

Output: Context-aware representation $C = [c_1, \ldots, c_M]$

1. Partition X into M time windows: $W_1, W_2, \ldots, W_M$
2. For each window W_i do
   For each position t in the window do
   Compute $e_{it} = v^t \cdot \tanh(W \cdot x_t + b)$
   End For
   Compute attention weights $\alpha_{it} = \text{softmax}(e_{it})$ over $t = 1$ to w
   Compute window-level context vector:
   $c_i = \sum(\alpha_{it} \cdot x_t)$ for $t = 1$ to w
   End For
   Aggregate $[c_1, \ldots, c_M] \to C$

Return C

Figure 3 shows the ATW module that divides the encoded behavioral sequence into several temporal windows and passes local attention to compute the context-attentive representation of each window. This module is brought into play to allow the model to identify any time-localized changes in behavior requirement, sudden peaks in effort or engagement disengagement that might be a precursor to performance change. The combination of its work with the TCN is the backbone of the multi-resolution attention structure of the proposed DTAN architecture.

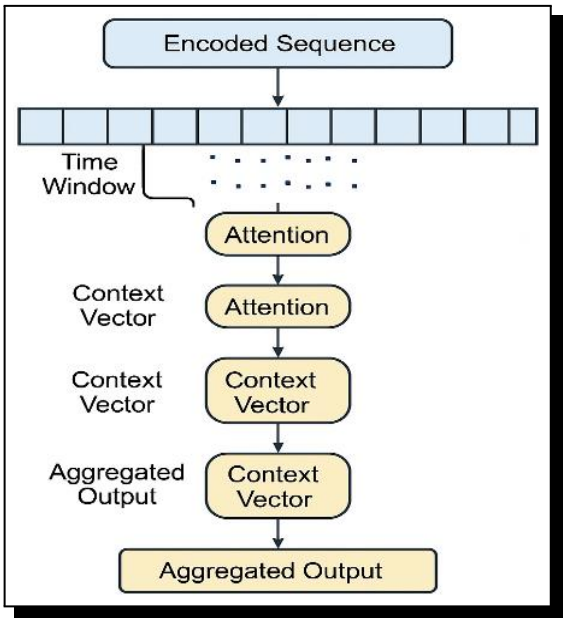


Figure 3. Attention-over-Time Windows (ATW) structure

3.5 Dynamic Contextual Attention Mechanism (DCAM)

The DCAM builds on the outputs of earlier layers by modeling global context-aware attention across the entire learner behavior sequence. While the ATW module captures local temporal salience, DCAM dynamically adjusts attention based on the global relevance of behaviors, learning dependencies across windows, and beyond. This mechanism enables the model to weigh what actions were necessary within local contexts and when and why those actions matter in broader behavioral patterns. For example, a burst of activity in early weeks may hold different predictive importance compared to one near the end of a course. Let $C = [c_1, c_2, \ldots, c_M]$ be the sequence of context vectors generated by the ATW module, where M is the number of temporal windows, and $c_i \in \mathbb{R}^d$. The DCAM computes a context-sensitive attention score for each c_i relative to the full sequence using scaled dot-product attention by using eqs. (12), (13) and (14):

$$e_i = \frac{q^\top \tanh(W_c c_i + b)}{\sqrt{d}}, \tag{12}$$

$$\alpha_i = \exp(e_i) \sum_{j=1}^{M} \exp(e_j), \tag{13}$$

$$z = \sum_{i=1}^{M} \alpha_i \, c_i, \tag{14}$$

where

- $q \in \mathbb{R}^d$ is a learnable global query vector,
- $W_c \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^d$ are trainable weights,
- α_i is the normalized attention score for window i,
- z is the final aggregated global context vector.

The learner behavioral trajectory is, in turn, summarized into a context vector z as input into the classification head, where performance prediction is made. An interpretation can be made about the contributions of the learned attention weights a_i of the windows that were recorded

in all the timelines that were utilized in the final prediction of the model. This enables teachers and curriculum designers to identify critical learning sessions, such as sudden downturns in interaction levels or recurring topics. The structure of the CAM enables bridging localized focus (via ATW) and temporal encoding (via TCN), through the combination of window-level vectors to produce a global behavior signature. This renders the last step of prediction relatively stable and explicable.

4. Results and Discussion

4.1 Overall performance

Evaluation considers multiclass prediction on EdNet and binary Pass versus Fail prediction on OULAD. DTAN attains the highest scores on both datasets. On EdNet, accuracy, macro precision, macro recall, and macro F1 improve overall baselines as summarized in Table 2. On OULAD, accuracy, precision, recall, F1, and AUC also improve, as shown in Table 3. Class wise precision recall analysis on EdNet confirms robustness under imbalance, with the best average PR AUC in Table 4. The macro F1 comparison in Figure 4(a) and the precision-recall curves in Figure 4(b) visualize these gains.

Table 2. Multiclass classification metrics — EdNet

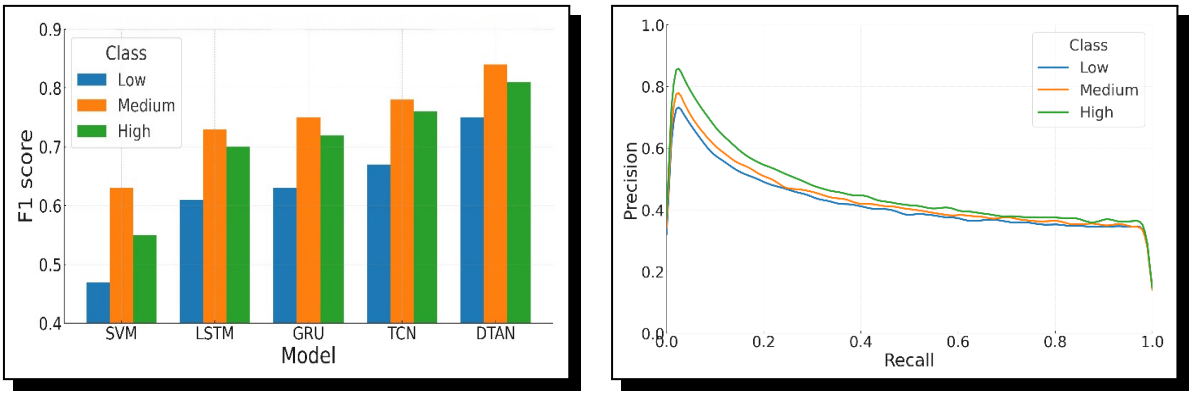
Model	Accuracy	Precision (Macro)	Recall (Macro)	F1 Score (Macro)	F1 Low	F1 Medium	F1 High
SVM	0.64	0.58	0.59	0.58	0.47	0.63	0.55
LSTM	0.72	0.68	0.69	0.68	0.61	0.73	0.70
GRU	0.74	0.70	0.70	0.70	0.63	0.75	0.72
TCN	0.77	0.73	0.74	0.74	0.67	0.78	0.76
DTAN	0.83	0.79	0.80	0.80	0.75	0.84	0.81

Table 3. Binary classification metrics — OULAD (Pass versus Fail)

Model	Accuracy	Precision	Recall	F1 Score	AUC
SVM	0.79	0.81	0.74	0.77	0.83
LSTM	0.85	0.86	0.83	0.84	0.89
GRU	0.86	0.87	0.84	0.85	0.90
TCN	0.88	0.89	0.87	0.88	0.92
DTAN	0.91	0.92	0.89	0.90	0.95

Table 4. Precision, recall, AUC by class — EdNet

Model	PR AUC Low	PR AUC Medium	PR AUC High	Average PR AUC
SVM	0.56	0.62	0.58	0.59
LSTM	0.64	0.70	0.68	0.67
GRU	0.67	0.72	0.70	0.70
TCN	0.71	0.77	0.74	0.74
DTAN	0.78	0.82	0.79	0.80



(a) Macro F1 across models on EdNet **(b)** Precision recall curves for EdNet

Figure 4. EdNetMacro F1 and Precision recall curves

Figure 4(a) reports macro F1 across models (SVM, LSTM, GRU, TCN, DTAN) for Low, Medium, and High classes; DTAN attains the highest scores across all classes. Figure 4(b) shows precision recall curves for DTAN by class, indicating higher precision at matched recall and stable behavior across the full recall range.

For OULAD, the multiclass confusion matrix in Figure 5 shows strong diagonal counts and low off diagonals.

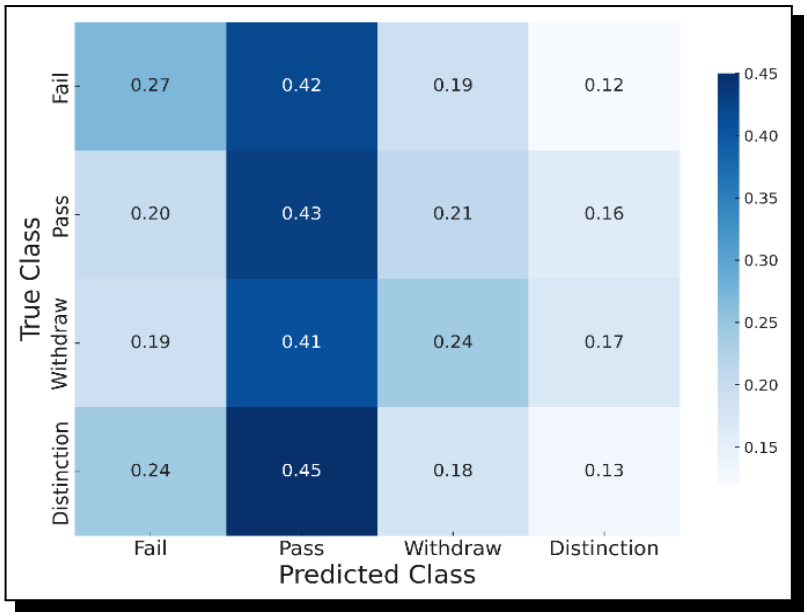


Figure 5. Confusion matrix for DTAN on OULAD (multiclass)

4.2 Early Prediction and Temporal Dynamics

Time-sensitive evaluation examines performance from partial sequences. On EdNet, accuracy grows from early steps to the full length, with a consistent margin over baselines at each checkpoint in Table 5. On OULAD, macro F1 rises across weeks with the largest gains in the first half of the course in Table 6.

The corresponding trend lines appear in Figure 7. Accuracy from partial sequences at steps 5, 10, 20, 30, 40, and 50 for SVM, LSTM, GRU, TCN, and DTAN. Curves rise with additional interactions; DTAN leads at every step, increasing from 0.65 at step 5 to 0.83 at step 50.

Table 5. Accuracy at incremental time steps — EdNet (sequence length 50)

Time step	SVM	LSTM	GRU	TCN	DTAN
5	0.51	0.58	0.59	0.61	0.65
10	0.53	0.62	0.63	0.67	0.71
20	0.56	0.66	0.68	0.71	0.75
30	0.59	0.69	0.71	0.74	0.78
40	0.61	0.71	0.73	0.76	0.80
50	0.62	0.72	0.74	0.77	0.83

Table 6. Macro F1 by week — OULAD (sequence length 10)

Week	SVM	LSTM	GRU	TCN	DTAN
1	0.48	0.56	0.57	0.59	0.63
2	0.51	0.59	0.60	0.63	0.67
4	0.54	0.63	0.64	0.67	0.70
6	0.56	0.66	0.67	0.70	0.74
8	0.58	0.68	0.69	0.72	0.76
10	0.60	0.70	0.71	0.74	0.78

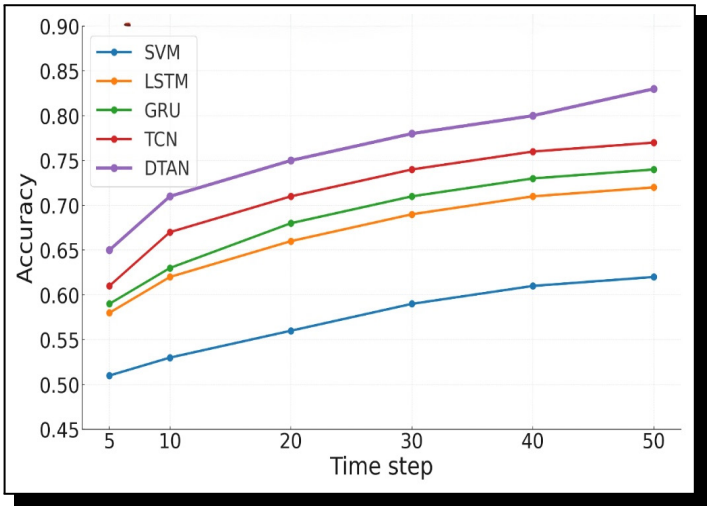


Figure 6. Accuracy versus time step on EdNet

Figure 7 presents weekly macro F1 at weeks 1, 2, 4, 6, 8, and 10 for SVM, LSTM, GRU, TCN, and DTAN. Performance improves with longer histories; DTAN remains highest throughout, from 0.63 in week 1 to 0.78 in week 10.

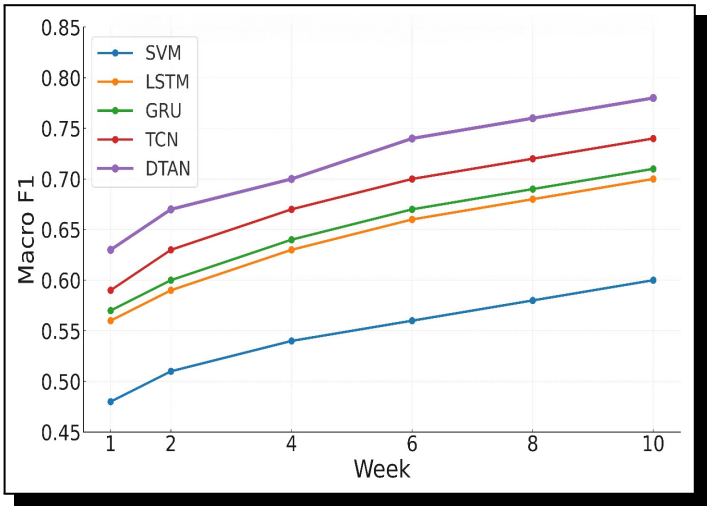
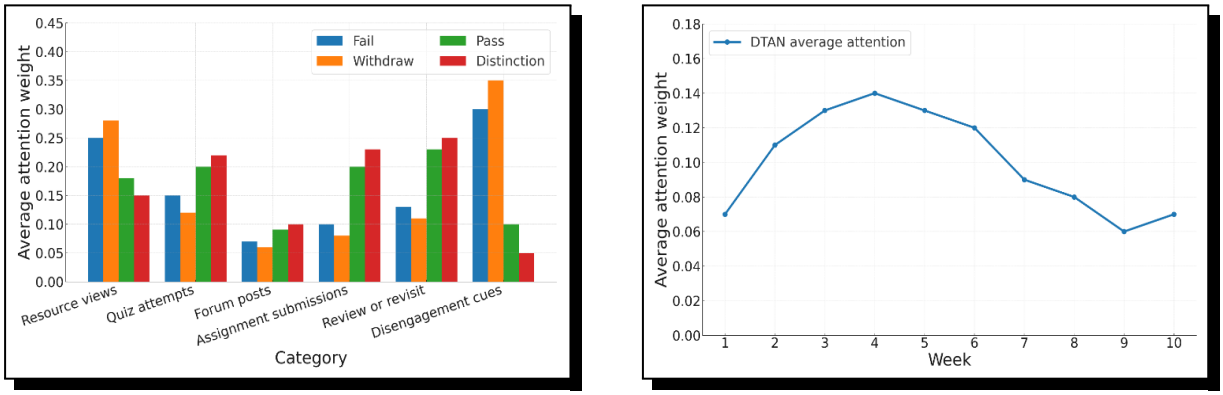


Figure 7. Macro F1 progression on OULAD

4.3 Interpretability via Attention

Attention analysis reveals both content salience and temporal salience. Semantic attention distributions differentiate successful and at-risk patterns, as described in Figure 8(a). Temporal attention shifts emphasize formative weeks with late spikes aligned with risk in Figure 8(b). These views align with the dual attention design and support targeted actions.



(a) Semantic attention distribution across outcome groups on OULAD

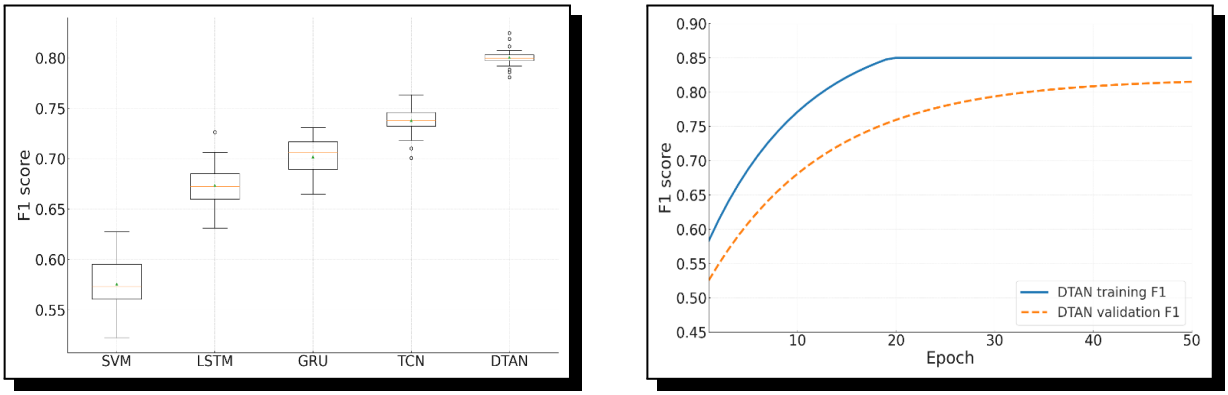
(b) Temporal attention shift over learning weeks for DTAN

Figure 8. Attention analysis on OULAD

Figure 8(a) presents semantic attention distribution across outcome groups—Fail, Withdraw, Pass, Distinction—over activity categories: Resource views, Quiz attempts, Forum posts, Assignment submissions, Review or revisit, Disengagement cues. Figure 8(b) Temporal attention shift for DTAN across learning weeks, with higher weights in early to mid-course weeks and reduced emphasis late in the course. Average attention weights are normalized within the group.

4.4 Efficiency and Training Behavior

Inference latency remains practical. Average per-sample times, with DTAN close to TCN and faster than recurrent baselines. Training behavior shows rapid convergence, balanced metric gains, and reduced variability. Learning curves, metric radar, and variability plots appear in Figure 9(a), and Figure 9(b), respectively.



(a) F1 variability across batches on EdNet

(b) Learning curves on EdNet

Figure 9. Training dynamics on EdNet

Figure 9(a) Batch level F1 variability by model; boxes show interquartile range, horizontal line the median, and marker the mean. DTAN exhibits the highest median and the tightest dispersion. Figure 9(b) DTAN learning curves over 50 epochs; training and validation F1 rise rapidly and plateau by ≈ 20 epochs with a small generalization gap.

4.5 Comparison With Reports in the Literature

A summary across prior studies appears in Table 7, which places DTAN at or above reported results on related tasks and datasets.

Table 7. Comparison with reported results in prior studies

Reference	Dataset	Task	Accuracy	F1-Score	AUC
[14]	ASSISTments	Knowledge Tracing	0.75	0.72	—
[23]	ASSISTments	Knowledge Tracing	0.77	0.74	—
[18]	EdNet	Question-Level Prediction	0.79	0.76	0.83
[5]	Simulated + Real	Knowledge Tracing	0.71	0.68	—
Proposed DTAN	EdNet & OULAD	Performance Classification	0.83	0.80	0.88

This study has significant implications for the creation of adaptive learning systems. The DTAN model presented indicated an excellent rate of predictive accuracy, temporal sensitivity, and interpretation. These attributes are required in real-time personalization. The DTAN can be utilized as a dynamic element within the content delivery to offer the trainer and the learning system some flexibility to modify their response to the behavior exhibited. As another example, when the model identifies the decline of a specific relevant activity to attention, certain interventions, such as motivational messages, individual feedback, or content rewriting, might occur automatically at particular points in time. With attention visualizations, educators can visualize when and why learners struggle, allowing them to move from a reactive to a proactive approach. Moreover, the fact that the DTAN is accurate at the early stage in predicting (based on time-step-level analyses) allows one to have early warning systems that the DTAN predicts at-risk learners sooner than the regular assessments, thus enabling a wider intervention window.

5. Conclusion and Future Work

The authors of DTAN utilize time convolution with attention to context to describe e learning behavior. Experiments on EdNet and OULAD demonstrate better accuracy, macro F1 score, and AUC compared to the RNN, GRU, TCN, and Transformer-based models. Early-stage prediction and time step level forecasting help to intervene in time. The attention visualizations can be used to showcase attention-capturing behaviors with instructor-facing explanations. Limitations remain. In high-stakes settings, attention weights should be complemented because they indicate correlation rather than causation. With widely distributed gaps in logs, performance can be degraded. There is no rigorous extrapolation to a multilingual, multiethnic, and informal environment. Real-time deployment on resource-constrained devices may require attention layers, which in turn introduce latency.

Future work will extend DTAN to multimodal inputs that combine textual reflections, video engagement, quiz outcomes, and sensor or biometric signals. Causal inference and counterfactual analysis will strengthen explanation and decision support. Transfer learning

and meta learning will enable adaptation across subjects, platforms, and learner populations, with fairness and leakage audits. Online studies within learning management systems will assess the impact on personalization and instructor action. Efficient variants through pruning and distillation will target edge deployment, and a complete release of code and configurations will support reproducibility.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] B. A. Alnasyan, M. Basheri and M. Alassafi, A comprehensive comparative analysis of deep learning models for student performance prediction in virtual learning environments: Leveraging the OULA dataset and advanced resampling techniques, *IEEE Access* **13** (2025), 75953 – 75972, DOI: 10.1109/ACCESS.2025.3564719.

[2] S. Alqahtani, Leveraging techniques of epistemic network analysis to discover behaviors of student learning reflections in online learning environments, *Engineering, Technology & Applied Science Research* **14**(3) (2024), 14191 – 14199, DOI: 10.48084/etasr.7274.

[3] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, 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 (editors), Association for Computational Linguistics, Doha, Qatar, pp. 1724 – 1734 (2014), DOI: 10.3115/v1/D14-1179.

[4] A. T. Corbett and J. R. Anderson, Knowledge tracing: Modeling the acquisition of procedural knowledge, *User Modeling and User-Adapted Interaction* **4** (1994), 253 – 278, DOI: 10.1007/BF01099821.

[5] T. Gervet, K. Koedinger, J. Schneider and T. Mitchell, When is deep learning the best approach to knowledge tracing?, *Journal of Educational Data Mining* **12** (2020), 31 – 54, DOI: 10.5281/zenodo.4143614.

[6] S. Ghaoui, S. M. Hemam and T. Djouad, An MDA-based approach for the design and automatic computation of collaboration indicators in e-learning systems, *Engineering, Technology & Applied Science Research* **15** (2025), 23235 – 23245, DOI: 10.48084/etasr.10607.

[7] L. He, X. Li, P. Wang, J. Tang and T. Wang, MAN: Memory-augmented attentive networks for deep learning-based knowledge tracing, *ACM Transactions on Information Systems* **42**(1) (2023), 1 – 22, DOI: 10.1145/3589340.

[8] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation* **9**(8) (1997), 1735 – 1780, DOI: 10.1162/neco.1997.9.8.1735.

[9] K. R. Koedinger, S. D'Mello, E. A. McLaughlin, Z. A. Pardos and C. P. Rosé, Data mining and education, *WIREs Cognitive Science* **6**(4) (2015), 333 – 353, DOI: 10.1002/wcs.1350.

[10] J. Kuzilek, M. Hlosta and Z. Zdrahal, Open university learning analytics dataset, *Scientific Data* **4** (2017), Article number: 170171, DOI: 10.1038/sdata.2017.171.

[11] B. Lim and S. Zohren, Time-series forecasting with deep learning: A survey, *Philosophical Transactions of the Royal Society A* **379**(2194) (2021), 20200209, DOI: 10.1098/rsta.2020.0209.

[12] A. A. Mir, M. F. Zuhairi, S. Musa, F. Alanazi, A. Namoun and A. Alrehaili, Enhanced variational graph convolutional networks with multi-scale convolutions and attention mechanisms for dynamic network analysis, *Engineering, Technology & Applied Science Research* **15**(1) (2025), 19838 – 19847, DOI: 10.48084/etasr.9443.

[13] A. Peña-Ayala, Educational data mining: A survey and a data mining-based analysis of recent works, *Expert Systems with Applications* **41** (2014), 1432 – 1462, DOI: 10.1016/j.eswa.2013.08.042.

[14] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas and J. Sohl-Dickstein, Deep knowledge tracing, in: *Proceedings of the 29th International Conference on Neural Information Processing Systems* (NIPS'15), Association for Computing Machinery, Vol. 1, pp. 505 – 513, (2015), URL: https://dl.acm.org/doi/10.5555/2969239.2969296.

[15] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas and J. Sohl-Dickstein, Deep knowledge tracing, in: *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett (editors), Vol. 28, Curran Associates, Inc. (2015).

[16] C. Romero and S. Ventura, Educational data mining and learning analytics: An updated survey, *WIREs Data Mining and Knowledge Discovery* **10**(3) (2020), e1355, DOI: 10.1002/widm.1355.

[17] B. Shickel, P. J. Tighe, A. Bihorac and P. Rashidi, Deep EHR: A survey of recent advances in deep learning techniques for Electronic Health Record (EHR) analysis, *IEEE Journal of Biomedical and Health Informatics* **22**(5) (2017), 1589 – 1604, DOI: 10.1109/JBHI.2017.2767063.

[18] D. Shin, Y. Shim, H. Yu, S. Lee, B. Kim and Y. Choi, Saint+: Integrating temporal features for ednet correctness prediction, in: *Proceedings of the 11th International Learning Analytics and Knowledge Conference* (LAK21), pp. 490 – 496, (2021), DOI: 10.1145/3448139.3448188.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems* (NIPS'17, Long Beach, California, USA), pp. 6000 – 6010, Curran Associates Inc., Red Hook, NY (2017), DOI: 10.5555/3295222.3295349.

[20] T. Xie, Q. Zheng and W. Zhang, Mining temporal characteristics of behaviors from interval events in e-learning, *Information Sciences* **447** (2018), 169 – 185, DOI: 10.1016/j.ins.2018.03.018.

[21] C. Xu, P. Zhao, Y. Liu, J. Xu, V. S. S. Sheng, Z. Cui, X. Zhou and H. Xiong, Recurrent convolutional neural network for sequential recommendation, in: Proceedings of the World Wide Web Conference (WWW'19, San Francisco, CA), pp. 3398 – 3404, Association for Computing Machinery, New York (2019), DOI: 10.1145/3308558.3313408.

[22] X. Zhang, J. Zhang, N. Lin and X. Yang, Sequential self-attentive model for knowledge tracing, in: *Artificial Neural Networks and Machine Learning* (ICANN 2021), I. Farkaš, P. Masulli, S. Otte and S. Wermter (editors), Lecture Notes in Computer Science series, Vol. 12891, pp. 318 – 330, Springer, Cham. (2021), DOI: 10.1007/978-3-030-86362-3_26.

[23] J. Zhang, X. Shi, I. King and D.-Y. Yeung, Dynamic key-value memory networks for knowledge tracing, in: *Proceedings of the 26th International Conference on World Wide Web* (WWW'17), pp. 765 – 774, (2017), DOI: 10.1145/3038912.3052580.

[24] Y. Zhou and J. Kang, Enriching Multimodal Data: A temporal approach to contextualize joint attention in collaborative problem-solving, *Journal of Learning Analytics* **10**(3) (2023), 87 – 101, DOI: 10.18608/jla.2023.7989.