A Dual-Phase Attention CNN-LSTM for MOOC Student Dropout Prediction

Fatma Dhaoui[1], Kalthoum Rezgui[2] and Nadia Ben Azzouna[3]

***Abstract*— Persistently high dropout rates in Massive Open Online Courses (MOOCs), routinely exceeding 80–90%, underscore the need for accurate and interpretable early-warning systems. While existing attention-based sequential models achieve strong ranking performance, they remain limited in terms of F1-score due to class imbalance and insufficient modeling of phase-specific engagement patterns. To address these challenges, this paper introduces CNN-LSTM-Dual-ATT, a novel hybrid deep learning (DL) architecture that integrates a parallel CNN-LSTM backbone with two independent multi-head attention modules, explicitly designed to capture early and late engagement dynamics. These representations are combined through a learnable fusion gate. The model is evaluated on the KDDCup 2015 benchmark dataset across four temporal windows and three attention head configurations. Experimental results show that the proposed approach achieves a best F1-score of 92.91% and an AUC-PR of 94.06% at 30 days, outperforming the baselines. In addition, explainable AI (XAI) techniques, comprising SHAP and fusion gate analysis, are employed to capture temporal learning patterns. The results demonstrate that the learnable fusion gate (w_{early}, w_{late}) provides actionable, phase-specific insights at the individual student level.**

I. INTRODUCTION

Massive Open Online Courses (MOOCs) have significantly expanded access to education, enrolling millions of learners worldwide through platforms, such as Coursera, edX, and XuetangX. However, this large-scale accessibility is accompanied by persistently high dropout rates, often exceeding 80–90% [1], with many learners disengaging within the first weeks. Accurately identifying at-risk students early enough to enable timely intervention remains a critical challenge in learning analytics.

MOOC platforms generate fine-grained temporal interaction data, capturing diverse learner activities such as video engagement, assessments, and navigation behaviors. These sequential traces contain valuable signals for dropout prediction, but require models capable of capturing both short-term patterns and long-range temporal dependencies. While DL approaches, particularly CNN-LSTM and attention-based models, have improved predictive performance [2], existing methods typically apply a single attention mechanism over the entire sequence. This design overlooks the potential differences between early and late engagement phases and limits both predictive performance and interpretability, especially under severe class imbalance.

To address these limitations, this paper proposes a novel hybrid architecture, CNN-LSTM-Dual-ATT, which explicitly models phase-specific engagement dynamics by applying separate multi-head attention mechanisms to early and late learning phases, combined through a learnable fusion gate. The proposed approach aims to improve the precision–recall trade-off while providing more interpretable, student-level insights.

The remainder of this paper is organized as follows. Section II reviews related work on attention-based deep learning approaches for MOOC dropout prediction. Section III describes the proposed CNN-LSTM-Dual-ATT architecture, detailing its four main components. Section IV presents the dataset and preprocessing pipeline. Section V reports the experimental results and comparative performance analysis. Section VI provides the explainability analysis using SHAP and attention weight visualization. Section VII discusses the key findings and architectural insights, and Section VIII concludes the paper with directions for future work.

II. RELATED WORK

Recent advances in student dropout prediction have increasingly emphasized the role of attention mechanisms in enhancing both predictive performance and interpretability. By enabling models to focus on the most informative parts of learner interaction data, attention-based approaches not only improve accuracy but also provide insights into which behaviors drive predictions [3].

Several studies have integrated attention within hybrid DL architectures to jointly capture spatial and temporal patterns in learner data. For instance, CNN–BiLSTM models augmented with attention have shown strong performance by refining latent representations and linking behavioral sequences to dropout outcomes [4]. Similarly, multi-stage frameworks combining CNN feature extraction, attention layers, and Temporal Convolutional Networks (TCN) have demonstrated the ability to capture multi-scale temporal dependencies in learning interactions [5].

In addition, attention mechanisms have been employed to explicitly model interactions among heterogeneous behavioral features. Approaches incorporating attention-based feature fusion or combining DL with ensemble methods, such as LightGBM, have improved representation learning and prediction accuracy, particularly in complex MOOC environments [6]. These models highlight the importance of

*This work was not supported by any organization

[1]Fatma Dhaoui is with SMART Lab, Higher Institute of Management of Tunis (ISGT), University of Tunis, Tunis, Tunisia fatmadhaoui0@gmail.com

[2]Kalthoum Rezgui is with SMART Lab, Higher Institute of Management of Tunis (ISGT), University of Tunis, Tunis, Tunisia kalthoum.rezgui@isamm.uma.tn

[3]Nadia Ben Azzouna is with SMART Lab, Higher Institute of Management of Tunis (ISGT), University of Tunis, Tunis, Tunisia nadia.benazzouna@ensi.rnu.tn

capturing dependencies not only across time but also between different types of learner activities.

More advanced approaches have further incorporated masked self-attention, Conditional Random Fields (CRF), or meta-embedding strategies to better model structured sequences and learner interaction dynamics [7]. Context-aware architectures leveraging attention-enhanced CNNs have also been proposed to capture interactions between learners and course content [8].

Despite these advances, several limitations persist. Most existing approaches rely on single-level or task-specific attention mechanisms, which may not effectively capture both local and global temporal dependencies in learner behavior. Furthermore, although attention-based models have been widely applied in related domains, such as knowledge tracing, their use for early dropout prediction remains relatively limited. Recent DL architectures show promise in modeling complex dependencies within learner interaction data; however, they are still insufficiently evaluated in early prediction scenarios [9]. Unlike these approaches, our model explicitly separates early and late engagement phases through two independent attention modules combined via a learnable fusion gate.

III. PROPOSED CNN-LSTM-DUAL-ATT MODEL

The proposed CNN-LSTM-Dual-ATT model for student dropout prediction is organized in four stages: a shared CNN-LSTM backbone, an adjusted positional encoding, dual-phase multi-head attention modules, and a learnable fusion gate followed by a shared classification head. The CNN branch applies 1D convolutions to extract local behavioral patterns, while the LSTM branch captures long-range temporal dependencies. Both branches process the same input sequence of length $\mathrm{T} \times N$, where T denotes the observation window length and N is the number of behavioral features.

After positional encoding, the temporal learning sequence of length T is split into two equal halves at position $\lfloor \frac{T}{2} \rfloor$.

The early phase covers time steps $[0, \tau - 1]$, corresponding to the initial portion of the observation window and capturing initial engagement signals. The late phase covers time steps $[\tau, T-1]$, corresponding to the subsequent portion of the observation window and reflecting behavioral changes closer to the prediction deadline, where τ denotes the temporal boundary separating the two phases ($\tau = T/2$ in the current implementation). Separate multi-head attention mechanisms are then applied independently to each phase, and the resulting representations are combined via the learnable fusion gate before the final classification.

- **Shared CNN-LSTM backbone:** The proposed architecture combines convolutional and recurrent learning components to capture complementary aspects of student learning behaviour from the same temporal activity sequence, enabling the network to learn both short-term interactions and long-term temporal dependencies relevant to dropout prediction. Given an input sequence $\mathbf{X} \in \mathbb{R}^{B \times T \times N}$, where B is the batch size, T is the observation window length, and N denotes the number of behavioral features (with $N = 7$ in our experiments), the CNN branch applies 1D convolutions to extract local behavioural patterns, producing a tensor of shape $[B,\ T,\ d_{\text{cnn}}]$:

$$\mathbf{C} = \text{CNN}(\mathbf{X}) \in \mathbb{R}^{B \times T \times d_{\text{cnn}}} \tag{1}$$

while the LSTM branch captures long-range temporal dependencies, producing a tensor of shape $[B,\ T,\ d_{\text{lstm}}]$:

$$\mathbf{R} = \text{LSTM}(\mathbf{X}) \in \mathbb{R}^{B \times T \times d_{\text{lstm}}} \tag{2}$$

The two outputs are concatenated along the feature dimension to form a unified representation of shape $[B,\ T,\ d_{\text{model}}]$, where $d_{\text{model}} = d_{\text{cnn}} + d_{\text{lstm}}$:

$$\mathbf{Z} = [\mathbf{C} \parallel \mathbf{R}] \in \mathbb{R}^{B \times T \times d_{\text{model}}} \tag{3}$$

- **Adjusted positional encoding:** Positional information is incorporated into the temporal representations to preserve the order of learner activities within the sequence. This mechanism is designed to emphasize short-term behavioral changes and local temporal patterns, which are particularly important for modeling the phase-specific learning dynamics captured by the proposed Dual-ATT framework.
- **Dual-phase temporal splitting and MHSA:** The temporal learning sequence of shape $[B,\ \mathrm{T},\ d_{\text{model}}]$ is divided into two equal halves:

$$\mathbf{X}_{\text{early}} = \mathbf{X}_{[0\,:\,\lfloor T/2 \rfloor]}, \quad \mathbf{X}_{\text{late}} = \mathbf{X}_{[\lfloor T/2 \rfloor\,:\,T]} \tag{4}$$

Separate multi-head attention mechanisms with h heads are then applied to each phase independently:

$$\mathbf{H}_{\text{early}} = \text{MHSA}(\mathbf{X}_{\text{early}},\, h)\,, \quad \mathbf{H}_{\text{late}} = \text{MHSA}(\mathbf{X}_{\text{late}},\, h) \tag{5}$$

followed by mean pooling to produce $\mathbf{H}_{\text{early}}$ and $\mathbf{H}_{\text{late}}$, each of shape $[B,\ d_{\text{model}}]$:

$$\mathbf{H}_{\text{early}} = \frac{1}{T/2} \sum_{t=0}^{T/2-1} \mathbf{H}_{\text{early}}^{(t)}, \quad \mathbf{H}_{\text{late}} = \frac{1}{T/2} \sum_{t=T/2}^{T-1} \mathbf{H}_{\text{late}}^{(t)} \tag{6}$$

This design allows the model to better distinguish between early engagement signals and later behavioral changes, both of which may contribute differently to dropout prediction.
- **Learnable fusion gate and classification head:** To combine information from the two learning phases, the model concatenates $\mathbf{H}_{\text{early}}$ and $\mathbf{H}_{\text{late}}$ into a vector of shape $[B,\ 2 \cdot d_{\text{model}}]$, then computes a gate scalar:

$$z = \sigma(\mathbf{W} \cdot [\mathbf{H}_{\text{early}};\, \mathbf{H}_{\text{late}}] + b) \tag{7}$$

where $w_{\text{early}} = z$ and $w_{\text{late}} = 1 - z$ are the learned per-student weights. The fused representation:

$$\mathbf{H}_{\text{fused}} = z \cdot \mathbf{H}_{\text{early}} + (1 - z) \cdot \mathbf{H}_{\text{late}} \tag{8}$$

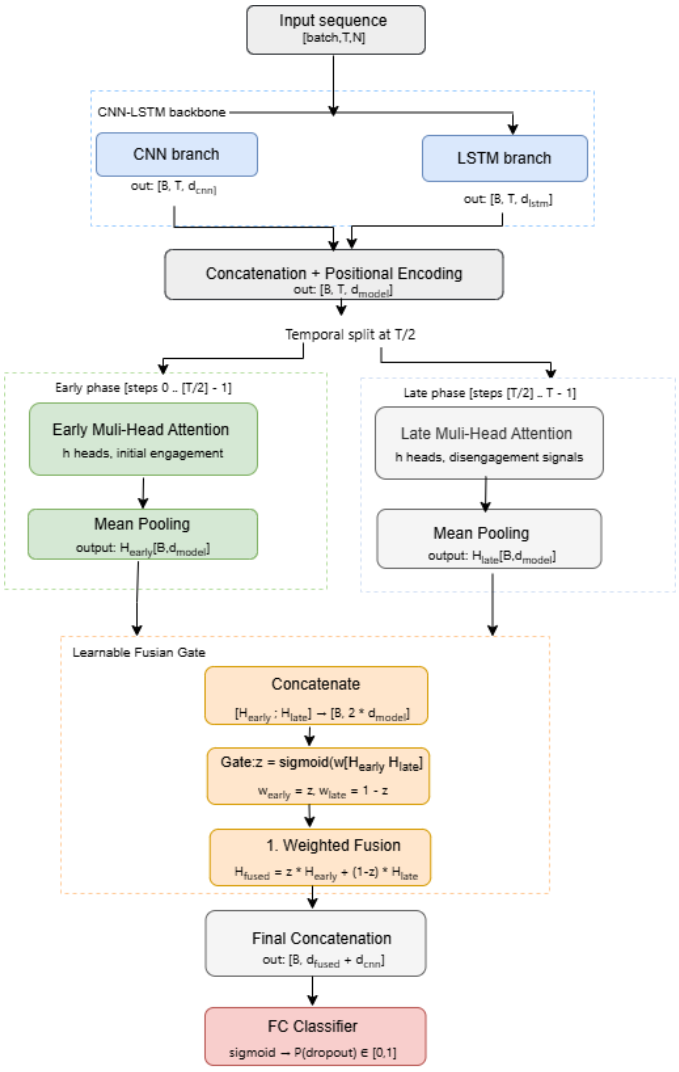

Fig. 1. CNN-LSTM with dual-phase multi-head attention architecture.

is then integrated with the CNN features and passed to the final fully connected classifier, which applies a sigmoid activation to estimate $P(\text{dropout}) \in [0, 1]$. The training process also incorporates class-imbalance handling via `BCEWithLogitsLoss` with `pos_weight`, improving the robustness of predictions under the 79% dropout rate observed in the dataset.

The detailed architecture of the proposed model, including tensor shapes at each stage, is summarized in Figure 1.

IV. DATASET

The KDD Cup 2015 dataset is derived from XuetangX, one of the largest MOOC platforms in China. It consists of anonymized, event-level interaction logs comprising approximately 120,542 activity records from 72,395 registered learners enrolled in 39 courses. Each course spans 30 days, and the data cover the period from October 27, 2013, to August 1, 2014.

Dropout is defined as a binary label, where a student is considered to have dropped out if no course interaction occurs within ten days after the prediction reference point. The training set exhibits a dropout rate of about 79.3%, making it a highly imbalanced classification problem.

To evaluate temporal performance, four observation windows are considered: Week 1 ($T = 7$), Week 2 ($T = 14$), Week 3 ($T = 21$), and Week 4 ($T = 30$).

The enrolment interaction logs are first aggregated into daily feature vectors of dimension $N = 7$. A sliding window of length T is then applied to generate overlapping temporal sequences. Unlike standard approaches, short enrolments (with length $< T$) are retained by applying zero-padding to reach a fixed sequence length. To ensure robustness, a binary mask is introduced to distinguish real observations from padded positions, allowing the attention mechanism to ignore non-informative inputs during computation.

V. EXPERIMENTAL RESULTS

A. Experimental setup

All experiments are implemented using PyTorch and executed on GPU (Google Colab with CUDA). A unified experimental protocol is adopted to ensure a fair comparison across all models. Specifically, all methods are trained and evaluated using the same 80/20 stratified data split (random_state = 42), the same set of seven behavioral features, and StandardScaler normalization fitted exclusively on the training data to prevent data leakage.

For evaluation, the proposed CNN-LSTM-Dual-ATT model uses an optimized decision threshold of 0.35, while all baseline models follow the standard threshold of 0.5 commonly used in the literature. Each model is trained over 10 independent runs to ensure statistical robustness. Performance is assessed using F1-score and AUC-PR, with AUC-PR considered the primary metric due to its threshold-independent nature and its ability to better reflect discriminative performance under class imbalance.

To ensure a comprehensive evaluation of the proposed architecture, a total of 12 configurations are tested, corresponding to the combination of 4 temporal window sizes (7, 14, 21, 30) and 3 attention head settings (8, 16, 32). This design allows the analysis of both short-term and long-term behavioral dependencies, as well as the impact of the attention capacity on model performance. Training is conducted using the Adam optimizer, combined with a ReduceLROnPlateau learning rate scheduler and early stopping to prevent overfitting. Detailed hyperparameter configurations are reported in Table I.

To further enhance interpretability, post-hoc explainability analysis is conducted for the proposed Dual-Attention model using two complementary XAI techniques. First, SHAP KernelExplainer is applied to quantify the global contribution of input features to model predictions. A background set of 20 training samples and 5 test samples is used, with inputs reshaped into a flattened representation. This allows fine-grained attribution of each time-step and feature to the predicted dropout probability. Second, attention weights from both early and late attention modules are analyzed to capture

phase-specific temporal importance, highlighting the most influential periods in the learning sequence.

B. Comparison models

To rigorously position the proposed architectures within the state of the art, we conduct a comprehensive benchmarking study against well-established baseline models, including CNN-LSTM [10], ConRec [10], and LSTM-ATT. These baselines were selected due to their strong relevance in modeling temporal dynamics and sequential patterns in similar tasks.

- CNN-LSTM [10]: The CNN-LSTM model integrates convolutional neural networks (CNNs) with LSTM networks to predict student dropout in MOOCs. It leverages 43-dimensional behavioral features derived from students' learning activity logs. The CNN component automatically extracts time-dependent patterns from these logs, while the LSTM component models the temporal dependencies in students' learning behaviors.
- ConRec [11]: The model is composed of an input layer, an output layer, and six hidden layers. The first and third hidden layers are convolutional layers that perform feature extraction, while the second and fourth are pooling layers used to reduce dimensionality. The fifth layer is a fully connected layer, and the sixth is a recurrent layer designed to capture temporal patterns in the time series data. The output layer produces predictions indicating whether a student is likely to drop out of the course based on the extracted and learned features.
- LSTM-ATT: It is a type of recurrent neural network (RNN) specifically tailored for time series data, combining LSTM layers with an attention component. The model processes input sequences containing N features at each time step through the LSTM, allowing it to learn and capture temporal dependencies in the data. The attention mechanism, then, assigns different weights to the outputs at each time step, emphasizing the most relevant parts of the sequence for making the final prediction. The resulting attention-weighted output is flattened and passed through a fully connected layer with a sigmoid activation function.

C. Effect of attention head count

Table II reports the performance of the CNN-LSTM-Dual-Att model across different temporal window sizes (T) and attention head counts (h). The results highlight a clear and consistent trend: the temporal window size is the dominant factor influencing model performance. Specifically, increasing T from 7 to 30 leads to substantial improvements across all evaluation metrics, including F1-score, AUC-PR, recall, and precision, while significantly reducing the error rate. This suggests that capturing longer temporal dependencies is critical for the task, as larger windows provide richer contextual information that enhances the model's discriminative capability.

In contrast, the impact of the number of attention heads (h) is comparatively limited and non-monotonic. While moderate improvements are observed when increasing h from 8 to 16, further increasing it to 32 does not consistently yield better performance and may even lead to marginal degradation in some cases. This behavior indicates diminishing returns from increasing attention head diversity, possibly due to redundancy between heads or over-parameterization. Moreover, the interaction between T and h reveals that the effectiveness of multi-head attention is contingent on the availability of sufficient temporal context. When T is small, increasing h does not improve performance, suggesting that multiple attention heads cannot compensate for the lack of informative input. Conversely, when T is large (e.g., T = 30), a moderate number of heads (h = 16) achieves the best balance, enabling the model to capture diverse temporal patterns without introducing unnecessary complexity. In sum, the best performance is achieved with the configuration (T = 30, h = 16), which provides the highest F1-score while maintaining a strong balance between precision and recall. These findings emphasize that temporal context plays a primary role in model effectiveness, whereas the number of attention heads serves as a secondary hyperparameter that requires careful tuning to avoid diminishing returns.

D. Comparative performance

Table III presents a performance comparison between the proposed Dual-ATT model and three baseline models across four temporal windows, in terms of AUC-PR and F1-score. For the proposed model, each reported result corresponds to the best-performing configuration identified in Table II. The proposed model consistently achieves the highest F1-score across all windows, demonstrating strong and stable classification performance. At Week 4, it obtains F1 = 92.91% and AUC-PR = 94.06%, improving over LSTM-ATT by +11.51 F1 and +1.96 AUC-PR, and over ConRec by +13.22 F1 and +0.30 AUC-PR.

While ConRec achieves competitive AUC-PR (93.76%), its relatively lower F1-score (79.69%) suggests a potential trade-off between ranking quality and classification threshold performance.

The performance gap in F1 generally increases with longer observation windows, indicating that the attention-based architecture benefits from richer temporal dependencies. At Week 1, the proposed model already achieves strong performance (F1 = 88.42%), supporting its applicability for early-stage prediction.

VI. EXPLAINABILITY ANALYSIS

To further understand the decision process of our dual-phase multi-head attention model, we analyze two complementary interpretability tools: the dual-phase attention weight visualization and the SHAP feature importance analysis. The attention weight plot (Figure 2) reveals a temporal asymmetry in the model's behavior. During the early phase (steps 0–6), attention weights remain perfectly uniform at approximately 0.1429, indicating that the model distributes its focus equally across all early time steps, treating this

TABLE I

EXPERIMENTAL SETUP AND TRAINING CONFIGURATION.

Component	**Configuration**
Data Split	80/20 stratified split (random_state = 42)
Input Features	7 behavioral features
Normalization	StandardScaler fitted on training data only
Thresholds	Proposed model: 0.35; Baselines: 0.5 (default literature setting)
Training Runs	10 independent runs per model
Evaluation Metrics	F1-score and AUC-PR (in %)
Optimizer	Adam (lr = 1e-3, weight decay = 1e-4)
Learning Rate Scheduler	ReduceLROnPlateau (patience = 2, factor = 0.5)
Batch Size	32
Epochs	Maximum 50 epochs
Early Stopping	Patience = 5 (validation loss)

TABLE II

CNN-LSTM-DUAL-ATT PERFORMANCE ACROSS ALL 12 CONFIGURATIONS.

T	**h**	**F1**	**AUC-PR**	**Recall**	**Precision**	**Error**
30	8	0.9284	0.9376	0.9404	0.9167	0.1169
30	16	**0.9291**	0.9371	**0.9508**	0.9084	0.1169
30	32	0.9265	**0.9406**	0.9301	**0.9229**	0.1190
21	8	0.9154	0.9286	0.9382	0.8937	0.1332
21	16	0.9131	0.9295	0.9295	0.8973	0.1359
21	32	0.9090	0.9277	0.9203	0.8990	0.1415
14	8	0.8992	0.9070	0.9111	0.8875	0.1560
14	16	0.9015	0.9079	0.9156	0.8878	0.1528
14	32	0.9022	0.9075	0.9164	0.8884	0.1517
7	8	0.8842	0.8707	0.9029	0.8664	0.1808
7	16	0.8833	0.8705	0.9021	0.8652	0.1824
7	32	0.8807	0.8706	0.8935	0.8682	0.1852

period as a stable, homogeneous baseline context with no single moment standing out as particularly discriminative. In contrast, the late phase (steps 7–13) exhibits a different dynamic: attention weights initially dip at the phase boundary (steps 7–9), suggesting a transitional period of lower informativeness, before exhibiting a slight monotonic increase toward step 13, where the maximum weight of approximately 0.1434 is reached. This pattern suggests that the final time steps may contain more discriminative information for dropout prediction.

The SHAP summary plot (Figure 3) supports and further refines this interpretation. In line with the late-phase emphasis observed in attention weights, the most influential features are located at time step 12, particularly time_12_navigate, which shows the strongest negative SHAP values when its magnitude is high, and time_12_wiki, which yields the strongest positive SHAP values under similarly high values. This contrast highlights a meaningful behavioral divergence: high engagement with wiki resources at the final time step appears to be associated with a higher likelihood of dropout, whereas intensive navigation activity at the same moment shifts the prediction away from dropout. Importantly, this opposing effect within a single time step reveals a level of feature-specific directionality that attention weights alone cannot provide, since attention primarily reflects temporal importance rather than the sign and influence of individual features. The relatively uniform attention observed in earlier steps is also reflected in the SHAP results, where features from time steps 0–6 have minimal impact on the prediction, aside from slightly positive contributions from early page-closing actions at steps 0–1.

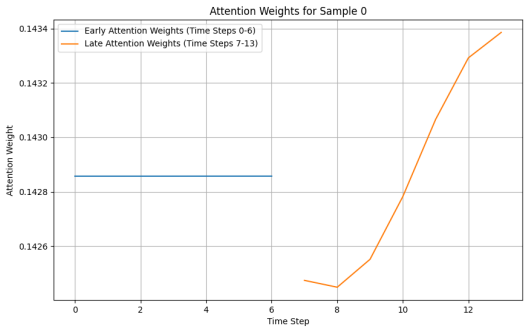

Fig. 2. Early vs late attention weights for sample 0.

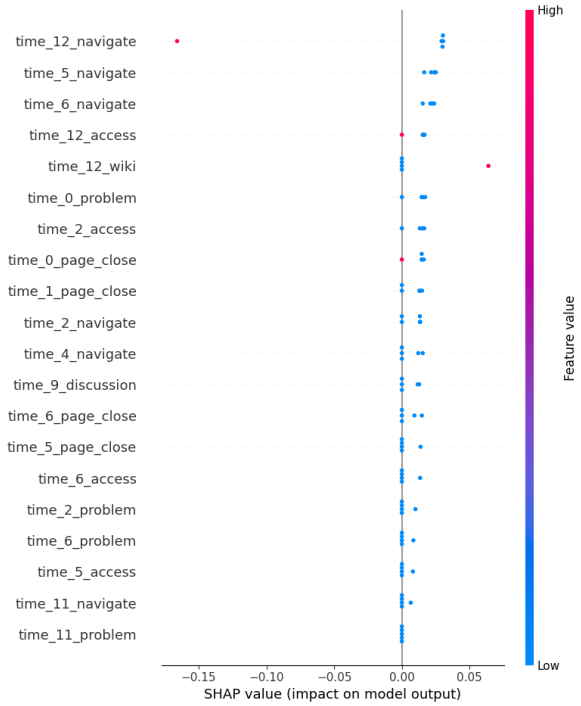

Fig. 3. SHAP summary of feature importance across time steps.

TABLE III

PERFORMANCE COMPARISON (AUC-PR AND F1, %) ACROSS FOUR TEMPORAL WINDOWS (W1–W4).

Model	**W1**		**W2**		**W3**		**W4**	
	AUC-PR	F1	AUC-PR	F1	AUC-PR	F1	AUC-PR	F1
CNN-LSTM	89.30	76.20	90.00	77.90	90.40	79.60	91.00	80.70
LSTM-ATT	90.22	77.36	91.10	79.99	91.55	80.98	92.10	81.40
ConRec	**93.06**	74.85	**93.41**	77.15	**93.59**	78.28	93.76	79.69
Ours (Dual-ATT)	87.07	**88.42**	90.75	**90.22**	92.86	**91.54**	**94.06**	**92.91**

VII. DISCUSSION

The stronger F1-score performance of the CNN-LSTM-Dual-ATT model relative to the baseline models can be attributed to several complementary architectural components. First, the parallel dual-branch design combines two fundamentally different inductive biases simultaneously: the CNN branch detects locally recurring and translation-invariant patterns, while the LSTM-attention branch captures long-range sequential dependencies. Second, sinusoidal positional encoding removes a fundamental ambiguity of vanilla attention, enabling differential weighting of early versus late events. Third, phase-specific temporal decomposition encodes a domain-informed inductive bias: early and late dropout patterns carry structurally different predictive information. Fourth, the learnable fusion gate adapts the early/late contribution ratio to each individual student's trajectory. Fifth, explicit class-imbalance handling via BCEWithLogitsLoss with pos_weight and a calibrated threshold of 0.35 directly addresses the precision–recall imbalance induced by the 79% dropout rate. Indeed, the model consistently leads on precision across all configurations. At T = 30 and h = 32, precision reaches 0.9229, the highest in the study. This precision advantage makes CNN-LSTM-Dual-ATT the preferred choice for precision-constrained deployment contexts where intervention resources are limited and false positives are costly, such as instructor-led personalized tutoring.

Beyond predictive performance, the XAI analysis provides meaningful insights into the model's decision process. The dual-phase attention weight visualization reveals a clear temporal asymmetry: while early-phase attention remains uniformly distributed, late-phase attention increases monotonically toward the final time steps, suggesting that recent behavioral signals tend to carry more discriminative information for dropout prediction. This is further supported by the SHAP analysis, which identifies time_12_navigate and time_12_wiki as the two most influential features, with opposing effects: high navigation activity at the final step reduces dropout probability, whereas intensive wiki engagement increases it. Crucially, this directional information which attention weights alone cannot provide is only accessible through SHAP. Together, these findings validate the architectural choice of separating early and late phases: each phase captures structurally different predictive signals, and the learnable fusion gate successfully adapts their relative contribution at the individual student level.

VIII. CONCLUSIONS

This paper proposed and evaluated CNN-LSTM-Dual-ATT, a novel hybrid deep learning architecture for MOOC student dropout prediction. The model applies two independent multi-head attention modules to early and late engagement phases within a shared CNN-LSTM backbone, combines their outputs via a learnable per-student fusion gate, and is evaluated on the KDDCup 2015 benchmark across 12 configurations. At the 30-day window, the model achieves F1 = 92.91% and AUC-PR = 94.06%, improving over the CNN-LSTM baseline by +12.21 F1 points, over LSTM-ATT by +11.51 points, and over ConRec by +13.22 points. Besides, the XAI analysis demonstrates that the learnable fusion gate provides actionable per-student and phase-specific explanations. Future work includes cross-platform validation, incorporation of forum NLP features, and prospective A/B testing of the early-warning system in live MOOC deployments.

REFERENCES

[1] Wang, W., Zhao, Y., Wu, Y.J., Goh, M., 2023. Factors of dropout from moocs: a bibliometric review. Library Hi Tech 41, 432–453.

[2] Lim, B., Arık, S.Ö., Loeff, N., Pfister, T., 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. International journal of forecasting 37, 1748–1764.

[3] Ghosh, A., Heffernan, N., Lan, A.S., 2020. Context-aware attentive knowledge tracing, in: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 2330–2339.

[4] Wang, S., Xiao, Y., Meng, F., 2024. Deep learning-based method for predicting student dropouts in moocs, in: 2024 7th International Conference on Machine Learning and Natural Language Processing (MLNLP), IEEE. pp. 1–6.

[5] FLiu, H., Zhang, W., 2022. A hybrid deep learning model for moocs dropout prediction, in: 2022 4th International Conference on Computer Science and Technologies in Education (CSTE), IEEE. pp. 178–183.

[6] Liu, H., Chen, X., Zhao, F., 2023. Learning behavior feature fused deep learning network model for mooc dropout prediction. Education and Information Technologies , 1–22.

[7] Yin, S., Lei, L., Wang, H., Chen, W., 2020. Power of attention in mooc dropout prediction. IEEE Access 8, 202993–203002.

[8] Feng, W., Tang, J., Liu, T.X., 2019. Understanding dropouts in moocs, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 517–524.

[9] Prenkaj, B., Velardi, P., Distante, D., & Faralli, S. (2020, October). A reproducibility study of deep and surface machine learning methods for human-related trajectory prediction. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 2169-2172).

[10] Tang, X., Zhang, H., Zhang, N., Yan, H., Tang, F., Zhang, W., et al., 2022. Dropout rate prediction of massive open online courses based on convolutional neural networks and long short-term memory network. Mobile Information Systems 2022.

[11] Wang, W., Yu, H., Miao, C., 2017. Deep model for dropout prediction in moocs, in: Proceedings of the 2nd international conference on crowd science and engineering, pp. 26–32.