

Deep Learning for Student Outcomes Temporal Deep Learning for Student Outcome Prediction: Comparing Multi-Task and Single-Task Approaches on the OULAD Dataset

Haili Qin

Shandong Xiehe University

Chuanxia Zhang

tvpj6643@outlook.com

Shandong Xiehe University

Linkai qi

Shanghai University of Engineering Sciences

Article

Keywords: Learning analytics, student outcome prediction, multi-task learning, deep learning, LSTM, educational psychology, behavioral engagement

Posted Date: July 10th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9014768/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

We test whether multi-task learning improves prediction of pass/fail and withdrawal from VLE logs—it does not. Using 14-day interaction sequences from 26,717 Open University students, we compare single-task and multi-task BiLSTM architectures against gradient boosting baselines across 10 runs. All deep learning variants achieved statistically equivalent F1 scores: 0.600 ± 0.004 (MTL without cross-attention), 0.598 ± 0.003 (MTL with cross-attention), and 0.597 ± 0.006 (single-task), with pairwise 95% CIs falling within $[-0.008, 0.003]$ (TOST $p < 0.001$, $\delta = 0.02$). BiLSTM outperformed LightGBM by 2.2% ($t = 7.26, 13 p < 0.001$, $d = 0.73$). Despite moderate task correlation ($r = 0.61$), MTL provided no advantage over single-task training—a finding robust to excluding withdrawn students (F1 change: -0.6%). In this dataset, task relatedness alone ($r = 0.61$) did not ensure multi-task benefits: pass/fail and withdrawal appear to have partially distinct behavioral antecedents, limiting cross-task transfer.

1 Introduction

Students who will eventually fail or withdraw often show early warning signs in their learning behavior—reduced logins, irregular activity patterns, declining engagement with course materials. If detected early enough, these signals could trigger interventions that redirect academic trajectories. The practical question is whether prediction models can reliably identify such students from behavioral logs, and if so, what modeling approaches work best.

Demographics and prior grades predict outcomes—but only at enrollment. Once the course begins, these features are fixed and cannot track emerging disengagement. A student's daily patterns of interaction with course materials, participation in discussions, and completion of assignments contain information about their current learning state that static measures miss.

Self-Determination Theory predicts that more self-determined students engage consistently over time, whereas externally regulated engagement tends to be irregular and deadline-driven (Ryan and Deci, 2000; Zimmerman, 2002; Pintrich, 2000). We operationalize these ideas using VLE logs: cumulative activity captures overall engagement, while rolling variability and short-term trends capture consistency and momentum (Fredricks et al., 2004). Whether these behavioral proxies actually predict pass/fail and withdrawal under an early-warning setting is tested empirically here.

VLE platforms log every click. The Open University dataset records timestamped interactions across seven modules for over 32,000 students, enabling detailed behavioral profiles. Several studies have leveraged such data to predict course outcomes, typically treating each prediction target in isolation.

Academic performance and dropout risk correlate ($r \approx 0.6$ in our data), raising a modeling question: does a shared encoder capture common precursors, or do the two outcomes depend on different proximal behaviors? Students who struggle academically often disengage before formally withdrawing; conversely, students who remain engaged but perform poorly may need different interventions than those who simply stop participating.

We test this empirically: does training one encoder for both tasks improve prediction over separate models? Correlation alone may not suffice—a student who fails but persists differs behaviorally from one who withdraws entirely, and a shared encoder may not capture both patterns. We focus on two related but distinct outcomes: course pass/fail and formal withdrawal. Using behavioral data from over 26,000 Open University students, we compare BiLSTM-based architectures against gradient boosting baselines across 10 experimental runs. Beyond asking whether deep learning outperforms traditional methods, we examine whether the moderate correlation between academic failure and withdrawal ($r \approx 0.6$) translates into multi-task benefits, and we identify which behavioral features drive prediction accuracy.

2 Related Work

2.1 Predicting Student Outcomes

Before LMS logs became available, dropout prediction relied on demographics—GPA, test scores—with few behavioral signals available (Romero and Ventura, 2010; Baker and Yacef, 2010; Zawacki-Richter et al., 2019; Chen et al., 2020). Tinto (1975) and Bean (1980) showed these static features explained substantial variance in college performance. Such models still inform admissions decisions, but they offer limited leverage for in-course intervention: the predictors are fixed at enrollment.

With LMS logs came behavioral features. Macfadyen and Dawson (2010) showed forum participation and content access predicted grades with moderate accuracy. Okubo et al. (2017) and Whitehill et al. (2017) later applied RNNs and LSTMs (Hochreiter and Schmidhuber, 1997) to clickstreams, finding that temporal modeling outperformed aggregate statistics (Breiman, 2001).

2.2 Multi-Task Learning in Education

Most educational prediction studies train separate models per outcome (Caruana, 1997; Ruder, 2017). Xu et al. (2017) jointly predicted multiple course grades using static features and found modest gains (2–3% F1). To our knowledge, no published study combines MTL with sequential models on behavioral logs; our setting differs in using temporal click sequences rather than cumulative features.

2.3 The OULAD Dataset

The Open University Learning Analytics Dataset (Kuzilek et al., 2017; Umer et al., 2021) provides the behavioral logs we use. Released in 2017, it contains data from 32,593 students across multiple course presentations, including demographic information, assessment scores, and detailed VLE interaction logs. Unlike prior OULAD studies that predict single outcomes from cumulative features, we use 14-day temporal windows and joint pass/fail + withdrawal modeling.

Prior work on OULAD has achieved varying levels of predictive accuracy depending on the specific task and methodology. Hlostá et al. (2017) reported AUC values around 0.7 for early dropout prediction using gradient boosting. Kuzilek et al. (2017) found that combining demographic and behavioral features

improved over either alone. Our study builds on this foundation while introducing multi-task temporal modeling.

3 Materials and Methods

3.1 Dataset and Preprocessing

We used the OULAD dataset (Kuzilek et al., 2017). We excluded students with fewer than 5 VLE clicks during weeks 2–6 ($n = 5,876$), leaving 26,717 students across seven modules. Each student’s learning trajectory was represented as a sequence of daily interaction summaries over 14-day windows. We chose 14 days because shorter windows (7 days) yielded insufficient temporal context in preliminary experiments, while longer windows (21 + days) reduced the number of usable samples due to sparse activity in early course weeks.

For each day, we extracted multiple behavioral features from the VLE logs. Total click count captured overall engagement intensity. We computed logarithmic transformations to reduce skewness and rolling statistics to capture short-term trends. Activity type distributions reflected the balance between content consumption, assessment work, and forum participation. Temporal derivatives indicated whether engagement was increasing or decreasing. The final feature representation consisted of 12 dimensions per time step, yielding input tensors of shape $(N, 14, 12)$ where N denotes the number of windows. We generated overlapping 14-day windows with stride = 7 days, starting from day 8 of enrollment. Each student contributed on average 4.0 windows (range: 1–6). This yielded 107,942 windows before splitting.

Two binary prediction targets were defined. Academic performance distinguished students who passed their courses from those who failed or withdrew. Dropout risk specifically identified students who formally withdrew before course completion. These targets exhibited moderate class imbalance, with 56% of windows in the pass category and 78% in the non-withdrawal category.

Data were split by student (not by window) to prevent information leakage. We allocated 70% of students to training, 10% to validation, and 20% to testing. The training set contained 18,702 students contributing 75,737 windows; validation had 2,672 students (10,750 windows); test had 5,343 students (21,455 windows).

Importantly, each 14-day window was extracted from the early-to-middle course period (approximately weeks 2–6 of the typical 30-week modules), with prediction targets defined as final course outcomes. This configuration simulates an early warning scenario in which behavioral patterns from the first month of enrollment inform end-of-course risk assessment, providing institutions approximately 20 or more weeks of lead time for targeted intervention.

3.2 Model Architecture

Our multi-task architecture employed a shared encoder followed by task-specific prediction heads. The encoder consisted of a bidirectional LSTM (BiLSTM) with 64 hidden units per direction, processing the 14-step input sequence to produce contextualized representations. We selected 64 units based on validation performance: 32 units underfit the data ($F1 \approx 0.58$), while 128 units showed no improvement over 64 units despite doubling parameters. Bidirectional processing allowed the model to incorporate both past and future context when representing each time step.

The final hidden state from the encoder served as input to two parallel prediction heads. Each head comprised a layer normalization module, a fully connected layer reducing dimensionality to 32 units with ReLU activation, dropout regularization, and a final linear layer producing class logits. This architecture allowed shared learning of behavioral patterns while maintaining flexibility for task-specific decision boundaries.

For comparison, we implemented single-task variants using identical encoder and head architectures but trained separately for each outcome.

The cross-attention variant added a layer between the shared encoder and task heads. For each task, queries came from the task's own head input, while keys and values came from the other task's head input. This allowed each task to attend to the other's representation before final prediction, adding 8,320 parameters ($\sim 5\%$ of total).

Figure 1

Model architectures. (a) Single-task models train separate encoders for each prediction target. (b) Multi-task models share the encoder while maintaining task-specific prediction heads. The shared encoder learns representations that may capture common behavioral patterns across tasks.

3.3 Training Procedure

Models were trained using the AdamW optimizer (Kingma and Ba, 2015) with learning rate 0.001 and weight decay 0.01. We employed cosine annealing learning rate scheduling over 30 epochs with early stopping based on validation loss, using patience of 5 epochs. Batch size was set to 256 samples

To address class imbalance, we applied inverse frequency weighting to the cross-entropy loss:

$$\mathcal{L}_{task} = - \sum_{i=1}^N w_{y_i} \cdot y_i \log(\hat{y}_i)$$

1

where $w_{y_i} = N / (C \cdot n_{y_i})$ denotes the inverse frequency weight for class y_i , with C classes and n_{y_i} samples in class y_i . For multi-task models, the total loss combined both tasks with equal weights:

$$\mathcal{L}_{total} = \mathcal{L}_{academic} + \mathcal{L}_{dropout}$$

Dropout regularization (Srivastava et al., 2014) with rate 0.3 and gradient clipping with maximum norm 1.0 (Pascanu et al., 2013) prevented overfitting and training instabilities. We selected dropout rate 0.3 after observing that lower rates (0.1–0.2) led to validation loss increasing after epoch 15, while higher rates (0.5) slowed convergence without improving final performance.

All experiments were repeated with 10 different random seeds to ensure adequate statistical power. We report means and standard deviations across runs for all metrics.

3.4 Baseline Methods

We compared deep learning approaches against four traditional machine learning baselines. Random forest classifiers (Breiman, 2001) used 50 trees with maximum depth 8 and class-balanced sample weighting. XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) used 50 estimators with maximum depth 4, representing state-of-the-art gradient boosting methods. Logistic regression employed L2 regularization with balanced class weights. All baselines received flattened feature vectors as input, concatenating all time steps into a single 168-dimensional representation. This approach reflects common usage of tree-based methods in learning analytics practice, where temporal features are aggregated rather than processed sequentially, providing a representative comparison of modeling paradigms.

Note that tree-based methods and logistic regression were trained with fixed random states to ensure reproducibility. Because these algorithms are deterministic given fixed hyperparameters and identical train-test splits, their reported standard deviations of zero reflect the absence of stochastic initialization rather than methodological error. Variance in neural network results arises from random weight initialization across seeds.

3.5 Evaluation Metrics

Model performance was assessed using macro-averaged F1 score as the primary metric, which balances precision and recall across classes and is robust to class imbalance.

Statistical analysis employed two complementary approaches. First, paired t-tests evaluated whether observed differences exceeded zero. Second, because non-significant p-values do not establish equivalence, we conducted Two One-Sided Tests (TOST) for equivalence (Lakens, 2017). TOST tests whether the true difference falls within a pre-specified equivalence bound δ . We set $\delta = 0.02$ (corresponding to a 2 percentage point difference in F1), reasoning that smaller differences would be practically negligible. Equivalence was declared when the 95% confidence interval for the difference fell entirely within $[-\delta, +\delta]$. We report 95% CIs throughout for consistency with conventional reporting standards, noting that this represents a more conservative criterion than the 90% CI typically used in TOST.

To quantify the relationship between prediction targets, we computed Pearson correlation and Cramér’s V between academic performance and withdrawal labels. This analysis assessed whether task relatedness—a theoretical prerequisite for multi-task learning benefits—was present in our data.

4 Results

4.1 Task Correlation Analysis

Before examining prediction performance, we assessed the relationship between the two target variables. Academic performance and withdrawal exhibited moderate positive correlation (Pearson $r = 0.61$, $p < 0.001$; Cramér’s $V = 0.61$), confirming that students who withdrew were disproportionately likely to have been struggling academically. This level of task relatedness is typically considered favorable for multi-task learning.

4.2 Overall Prediction Performance

Table 1 presents the academic performance prediction results across 10 independent runs. Multitask learning without cross-attention achieved the highest mean F1 score (0.600), though all three deep learning variants performed within a narrow range (0.597–0.600). All neural approaches outperformed traditional machine learning baselines by 2–4% in F1.

Table 1 Academic performance prediction results. Values are mean $\pm$ standard deviation across 10 random seeds. Best result in bold.		
Model	F1 Score	95% CI
MTL without Cross-Attention	0.600 $\pm$ 0.004	[0.598, 0.602]
MTL with Cross-Attention	0.598 $\pm$ 0.003	[0.596, 0.600]
Single-Task LSTM	0.597 $\pm$ 0.006	[0.593, 0.600]
LightGBM	0.584 $\pm$ 0.000	[0.584, 0.584]
Random Forest	0.577 $\pm$ 0.001	[0.576, 0.578]
Logistic Regression	0.556 $\pm$ 0.000	[0.556, 0.556]
XGBoost	0.555 $\pm$ 0.000	[0.555, 0.555]

4.3 Equivalence Testing

Paired t-tests revealed no significant differences among deep learning variants (Single-Task vs. MTL w/o CA: $t = -1.51$, $p = 0.16$; Single-Task vs. MTL w/ CA: $t = -0.63$, $p = 0.54$). However, absence of significance does not establish equivalence. We therefore conducted TOST equivalence tests with $\delta = 0.02$ (Table 2).

All pairwise comparisons yielded TOST $p < 0.001$, with 95% CIs falling entirely within the equivalence bounds. The three deep learning approaches achieved equivalent performance—not merely a failure to

detect a difference, but positive evidence of equivalence within $\delta = 0.02$ across 10 seeds

Table 2
Equivalence testing results (TOST, $\delta = 0.02$). All comparisons demonstrate statistical equivalence.

Comparison	Mean Difference	95% CI	Equivalent?
Single-Task vs. MTL w/o CA	-0.004	[-0.008, 0.001]	Yes ($p < 0.001$)
Single-Task vs. MTL w/ CA	-0.002	[-0.006, 0.003]	Yes ($p < 0.001$)
MTL w/ CA vs. MTL w/o CA	-0.002	[-0.005, 0.001]	Yes ($p < 0.001$)

Deep learning outperformed gradient boosting: single-task LSTM exceeded LightGBM by 2.2% relative (+ 0.013 absolute F1; $t = 7.26$, $p < 0.001$, Cohen's $d = 0.73$). This margin, while modest, may be practically meaningful for early warning systems.

4.4 Withdrawal Prediction

Table 3 presents withdrawal prediction results. Lower F1 scores reflect severe class imbalance (78% non-withdrawal). MTL without cross-attention achieved the highest F1 (0.489), though with notably higher variance. Cross-attention again conferred no benefit. Withdrawal prediction appears to rely on partially distinct behavioral signals that do not transfer effectively from the academic task.

Table 3
Withdrawal prediction results. Values are mean $\pm$ standard deviation across 10 random seeds.

Model	F1 Score	95% CI
MTL without Cross-Attention	0.489 $\pm$ 0.025	[0.471, 0.508]
MTL with Cross-Attention	0.479 $\pm$ 0.030	[0.457, 0.501]
Single-Task LSTM	0.476 $\pm$ 0.003	[0.473, 0.478]
Logistic Regression	0.477 $\pm$ 0.000	[0.477, 0.477]
Random Forest	0.468 $\pm$ 0.015	[0.457, 0.479]
LightGBM	0.450 $\pm$ 0.000	[0.450, 0.450]
XGBoost	0.442 $\pm$ 0.000	[0.442, 0.442]

4.5 Sensitivity Analysis

A potential concern with our multi-task formulation is label overlap: students who withdrew were classified as failures in the academic task while simultaneously serving as positive cases in withdrawal prediction. To assess whether this overlap artificially inflated performance or confounded multi-task learning dynamics, we conducted a sensitivity analysis excluding withdrawn students from the academic task, predicting only Pass versus Fail.

Performance decreased by only 0.6% when using clean labels (Table 4), a difference well within the confidence intervals of the original analysis. This confirms that label overlap did not meaningfully inflate performance estimates. The minor decrease likely reflects reduced training set size (58,872 vs. 75,737 samples) rather than label leakage. Critically, these results indicate that our primary findings—the equivalence of single-task and multi-task approaches—are robust to alternative operationalizations of academic performance.

Table 4
Sensitivity analysis: academic prediction with clean labels (Pass vs. Fail only, excluding Withdrawn). Results for single-task LSTM across 5 random seeds; reduced replication count reflects computational constraints for this supplementary validation, though observed variance remains comparable to the main experiments.

Label Definition	F1 Score	Change
Original (Pass vs. Fail+Withdrawn)	0.597 ± 0.006	
Clean (Pass vs. Fail only)	0.593 ± 0.008	-0.6%

4.6 Behavioral Predictors of Academic Success

To address our third research question regarding which behavioral patterns most strongly predict outcomes, we conducted feature importance analysis using random forests (Table 5).

Table 5
Feature importance for academic performance prediction. Importance values represent aggregated contribution across all time steps.

Feature Type	Importance	Psychological Interpretation
Cumulative Average	34.6%	Sustained engagement
Rolling Mean (3-day)	14.9%	Recent activity trend
Rolling Std (3-day)	14.3%	Engagement consistency
Click Change Rate	9.3%	Behavioral momentum
Log Click Count	7.3%	Overall activity level

Cumulative engagement dominated (34.6%), indicating that sustained participation—not sporadic intense activity—characterizes successful students. This aligns with SDT’s prediction that intrinsic motivation manifests through consistent engagement (Ryan and Deci, 2000). Rolling statistics also mattered: engagement stability matters beyond raw activity levels.

Behaviors in Days 11–14 were most predictive, accounting for 45% of feature importance (Fig. 2). Engagement patterns near assessment deadlines appear most diagnostic of eventual outcomes,

consistent with deadline-driven behavior research (Steel, 2007). For early warning systems, this implies a critical period—intervene earlier, while patterns are still malleable.

4.7 Ablation Analysis

To understand the contribution of different model components, we conducted ablation experiments. Removing the bidirectional processing reduced F1 by 2.1%, confirming that future context aids prediction. Reducing hidden dimensionality from 64 to 32 units decreased performance by 1.8%, suggesting the model benefits from representational capacity. Eliminating dropout regularization led to overfitting, with validation performance degrading after fewer epochs.

5 Discussion

5.1 Temporal Modeling Provides Modest Benefits

BiLSTM models outperformed gradient boosting by 2.2% in F1 (0.597 vs. 0.584 for LightGBM). This margin, while statistically reliable ($t = 7.26$, $p < 0.001$), is modest in absolute terms. Temporal modeling may help because sequential patterns—such as declining engagement over consecutive days—carry information that aggregated features obscure. Ablation analysis supports this interpretation: removing bidirectional processing reduced F1 by 2.1%, indicating that both past and future context within the 14-day window contribute to prediction.

Temporal deep learning offers incremental improvements over well-tuned tree-based methods in this setting. Whether the 2.2% gain justifies the additional implementation complexity depends on institutional priorities and infrastructure constraints.

5.2 Multi-Task Learning Achieved Equivalent Performance

MTL did not improve prediction. Equivalence testing confirmed that single-task LSTM, MTL with hard parameter sharing, and MTL with cross-attention all achieved F1 scores within 0.3% of each other (all TOST $p < 0.001$, $\delta = 0.02$). The cross-attention mechanism, intended to capture task dependencies explicitly, provided no advantage over simpler architectures.

MTL may fail despite $r = 0.61$ because pass/fail and withdrawal share only coarse risk factors while depending on different proximal behaviors: some students persist with low performance, whereas others disengage entirely. A student who fails but keeps attending exhibits different engagement dynamics than one who withdraws; these differences limit what the shared encoder can exploit.

Additionally, both tasks have ample training data (> 75,000 examples). Ding et al. (2018) found that MTL helped predict rare phenotypes in healthcare but offered limited gains for common outcomes—consistent with our setting. When single-task models have enough data to learn robust representations on their own, the regularization benefit of shared weights diminishes.

5.3 Boundary Conditions for Multi-Task Learning

Task relatedness ($r = 0.61$) did not translate into MTL benefits. This null result is informative: the nature of the relationship matters as much as its magnitude. Pass/fail and withdrawal share variance but may arise from partially dissociable processes—a student who fails but persists differs motivationally from one who withdraws, and VLE click logs may not capture these distinctions.

For practice, researchers should not assume that conceptually related outcomes will benefit from joint modeling. Empirical validation is necessary. Task combinations that share proximal behavioral causes—such as assignment submission patterns and quiz performance—may yield MTL gains where our distal outcome pairing (final grades and withdrawal) did not.

5.4 Psychological Interpretation of Behavioral Predictors

Cumulative engagement accounted for 34.6% of feature importance—sustained participation, not sporadic bursts, characterizes successful students. In our early-warning window (weeks 2–6), this pattern is consistent with SDT’s prediction that intrinsic motivation manifests through consistent engagement (Ryan and Deci, 2000; Fredricks et al., 2004). We note, however, that these behavioral proxies do not directly measure motivation; the interpretation remains correlational.

Rolling statistics (trend and variability) contributed 29% of predictive power. In our weeks 2–6 window, students with stable engagement—low day-to-day variability and non-declining trends—performed better (Schunk and Zimmerman, 2012). Monitoring engagement consistency may be more informative than tracking total volume.

Features from Days 11–14 contributed 45% of total importance, revealing a critical period effect. Engagement patterns near assessment deadlines appear most diagnostic of eventual outcomes, consistent with research on deadline-driven behavior (Steel, 2007). Early intervention may be critical: once engagement patterns stabilize (around week 4–5 in our data), changing trajectories becomes harder.

5.5 Implications for Learning Analytics

In early warning systems (Tsai et al., 2020; Lim et al., 2021), temporal modeling of VLE logs offers incremental gains over aggregate features, but gains are modest (2.2%). Whether this justifies the added complexity depends on infrastructure and priorities. Engagement consistency and cumulative activity—not raw click volume—appear most diagnostic; students with declining trends or high variability may warrant outreach. Finally, MTL added no value here. When training data are abundant, single-task models may suffice.

5.6 Limitations and Future Directions

Data came from a single institution; replication across diverse settings is needed. We used only VLE click logs, omitting discussion content, assessment text, and demographics. Baselines used flattened

features; more sophisticated temporal feature engineering for tree models might narrow the gap. Our MTL formulation used equal-weight loss combining; uncertainty weighting (Kendall et al., 2018) or gradient-based task balancing might yield different results. We also did not implement instance-level interpretability.

Future work could incorporate NLP analysis of discussion posts to capture engagement quality. Longitudinal designs tracking students across multiple courses would reveal how behavioral patterns evolve. Most importantly, coupling prediction models with randomized intervention trials would test whether early warnings translate into improved outcomes when acted upon.

6 Conclusion

BiLSTM models outperformed gradient boosting baselines by 2.2% ($F1 = 0.600$ vs. 0.584 ; $t = 7.26, 307$ $p < 0.001$), indicating that temporal modeling captures behavioral dynamics that aggregate features miss. However, multi-task learning provided no advantage over single-task training: all three deep learning variants achieved equivalent performance (TOST $p < 0.001$, $\delta = 0.02$), with 95% CIs for pairwise differences falling within $[-0.008, 0.003]$. This null result held despite moderate task correlation ($r = 0.61$).

Task relatedness alone ($r = 0.61$) did not produce MTL gains. Pass/fail and withdrawal, while correlated, appear to have partially distinct behavioral antecedents. Cumulative engagement (34.6% importance) and behavioral consistency (29%) emerged as the strongest predictors, consistent with self-regulated learning theory.

For practitioners, temporal deep learning offers incremental gains, but the added complexity of MTL may not be justified when training data are abundant. Single-task models achieved equivalent accuracy with simpler implementation.

Declarations

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This research was supported by the National Education Science Planning Project “Research on Interpersonal Sensitivity and Compensation Education Mechanism of Disadvantaged College Students—Taking Private Universities in Shandong Province as an Example” (DEA140268) and the Shandong Province Undergraduate Teaching Reform Project “Construction and Practice of a Comprehensive Ideological and Political Education System Based on Student Development” (Z2022162).

Author Contribution

CZ conceived the study, implemented the models, conducted the experiments, and wrote the manuscript. SJ supervised the research and revised the manuscript. All authors contributed to the article and approved the submitted version.

Data Availability

The Open University Learning Analytics Dataset used in this study is publicly available at <https://analyse.kmi.open.ac.uk/open-dataset>. Code for reproducing our experiments is available at <https://github.com/alexlocez/code>.

References

1. Baker, R. S. and Yacef, K. (2010). Data mining for education. In International Encyclopedia of Education, pages 112–118. Elsevier, Oxford, 3rd edition.
2. Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, 12(2):155–187.
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
4. Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75.
5. Chen, L., Huang, L., Zheng, Y., and Tan, H. = (2020). Utilizing learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development*, 68(4):1961–1990.
6. Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.
7. Ding, D. Y., Simpson, S., Pfister, H., and Bhatt, A. (2018). The effectiveness of multitask learning for phenotyping with electronic health records data. In *Pacific Symposium on Biocomputing 2019*, pages 18–29. World Scientific.
8. Fredricks, J. A., Blumenfeld, P. C., and Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1):59–109.
9. Hlosta, M., Zdrahal, Z., and Zendulka, J. (2017). Ouroboros: Early identification of at-risk students without models based on legacy data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 6–15.
10. Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.
11. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*,

volume 30.

12. Kendall, A., Gal, Y., and Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.
13. Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA.
14. Kuzilek, J., Hlosta, M., and Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific Data*, 4(1):1–8.
15. Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362.
16. Lim, L.-A., Gasevic, D., Matcha, W., Ahmad Uzir, N., and Dawson, S. (2021). Students’ sense-making of personalised learning analytics. *British Journal of Educational Technology*, 37852(5):1912–1929.
17. Macfadyen, L. P. and Dawson, S. (2010). Mining lms data to develop an “early warning system” for educators: A proof of concept. *Computers & Education*, 54(2):588–599.
18. Okubo, F., Yamashita, T., Shimada, A., and Ogata, H. (2017). A neural network approach for students’ performance prediction. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 598–599.
19. Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, pages 3861310–1318.
20. Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In *Handbook of Self-Regulation*, pages 451–502. Academic Press, San Diego, CA.
21. Romero, C. and Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 40(6):601–618.
22. Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv [Preprint]*. Available at: <https://arxiv.org/abs/1706.05098> (Accessed June 1, 2024).
23. Ryan, R. M. and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78.
24. Schunk, D. H. and Zimmerman, B. J. (2012). *Motivation and Self-Regulated Learning: Theory, Research, and Applications*. Routledge, New York.
25. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.
26. Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, 133(1):65–94.
27. Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1):89–125.

28. Tsai, Y.-S., Poquet, O., Gašević, D., Dawson, S., and Pardo, A. (2020). Learning analytics in higher education—challenges and policies: A review of eight learning analytics policies. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, pages 233–242.
29. Umer, R., Susnjak, T., Mathrani, A., and Suriadi, S. (2021). A systematic mapping of the mooc dropout research. Open Learning: The Journal of Open, Distance and e-Learning, 36(3):268–289.
30. Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., and Tingley, D. (2017). Mooc dropout prediction: How to measure accuracy? In Proceedings of the Fourth ACM Conference on Learning @ Scale, pages 161–164.
31. Xu, J., Moon, K. H., and Van Der Schaar, M. (2017). Progressive prediction of student performance in college programs. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, pages 1395–1401.
32. Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? International Journal of Educational Technology in Higher Education, 16(1):1–27.
33. Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory into Practice, 41(2):64–70.

Figures

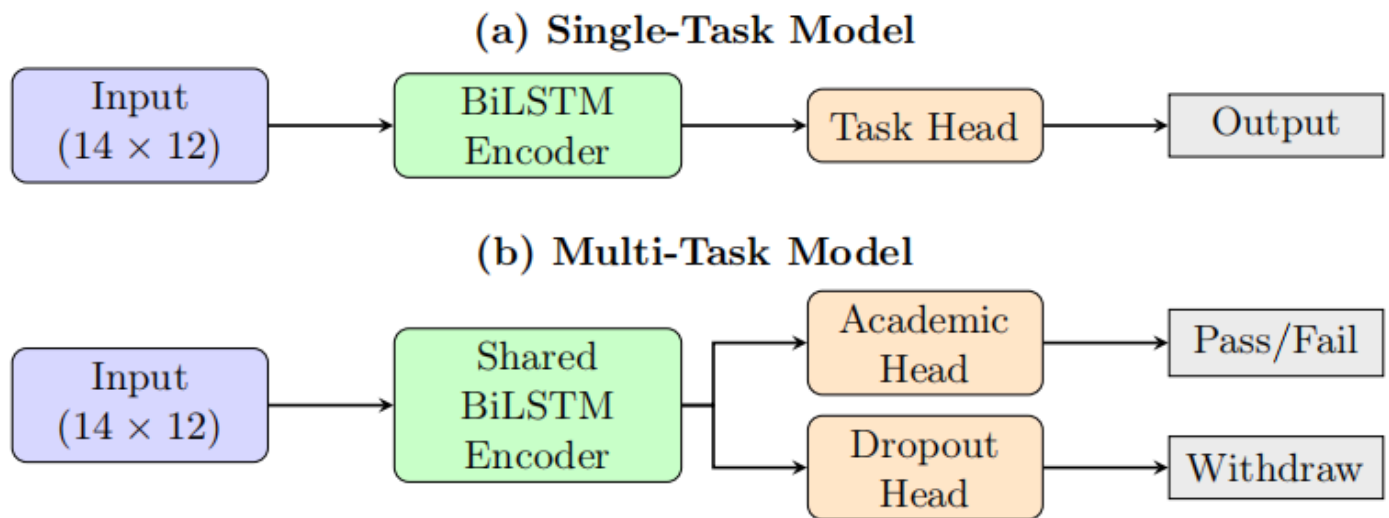


Figure 1

Model architectures. (a) Single-task models train separate encoders for each prediction target. (b) Multi-task models share the encoder while maintaining task-specific prediction heads. The shared encoder learns representations that may capture common behavioral patterns across tasks.

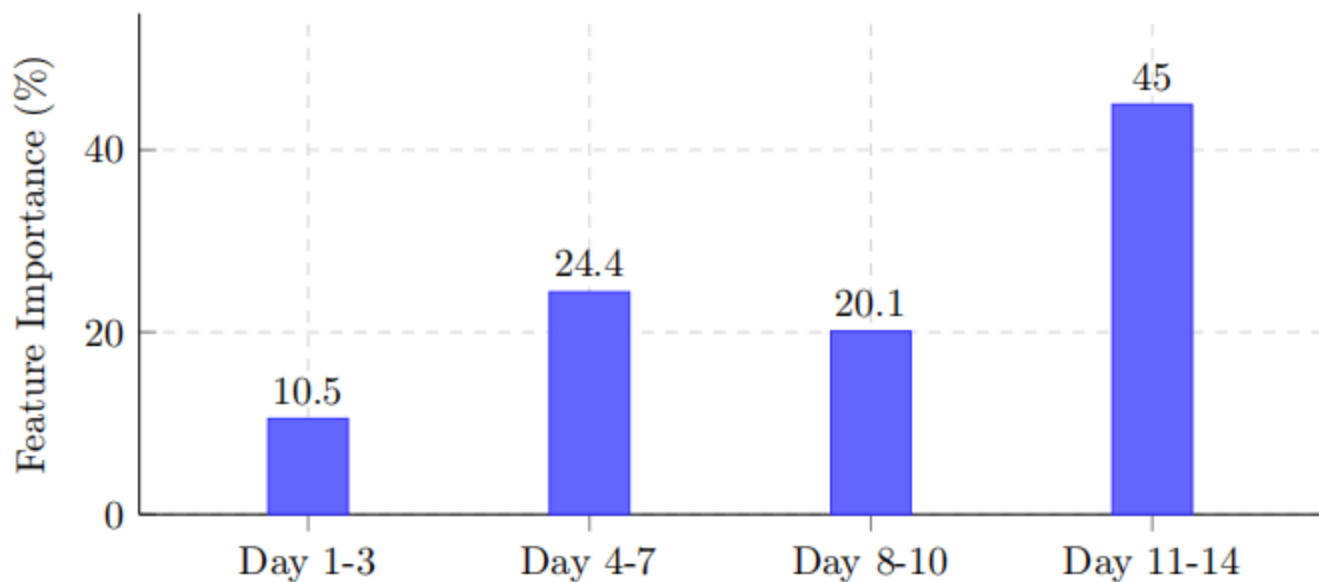


Figure 2

Distribution of feature importance across the 14-day observation window. Later days (Day 11-14) contribute disproportionately to prediction accuracy (45.0%), suggesting that engagement patterns near assessment deadlines are most diagnostic of student outcomes. This “critical period” effect has implications for the timing of interventions.