UDC 004.8 https://doi.org/10.54251/2616-6429.2025.04.n13

A.S. Talgatov, S.A. Nurgaliyeva*
Master degree student, School of Software Engineering, Astana IT University, Astana, Kazakhstan
PhD in computer science, School of Software Engineering, Astana IT University, Astana, Kazakhstan
***Corresponding author:** symbat.nurgaliyeva@astanait.edu.kz

A REPRODUCIBLE EXPLAINABLE AI PIPELINE FOR TEACHER-FACING STUDENT DIGITAL TWINS

Abstract

Learning management systems record large volumes of student activity, yet their built-in analytics tend to describe the past rather than predict outcomes. We ask a deliberately falsifiable question: do engineered Student Digital Twin features improve grade prediction over a competent baseline drawn from ordinary LMS signals, and are the resulting explanations stable enough to put in front of a teacher? To answer it we built a leakage-aware pipeline that pairs a weekly student-state representation with a gradient-boosting predictor and a perturbation-based explanation layer, then ran a nested feature ablation on public data from two institutions. The result isn't encouraging. Added feature richness does not help in any reliable way: no Twin block beats the baseline by more than one RMSE point on the primary cohort, only one cell shows a clear gain, and explanation rankings shift with the evaluation regime (Kendall τ between 0.32 and 0.79).

Keywords: learning analytics; explainable AI; explanation stability; student performance prediction; digital twin; gradient boosting; reproducibility

Introduction

Learning management systems quietly accumulate a detailed record of how students work: when they log in, what they submit, the marks they receive, and how often they click through course materials. Teachers, however, rarely get a forward-looking picture out of that record. They can't easily see where a particular student stands this week, what the likely end-of-course outcome is, or which signals are driving that estimate. Most learning-analytics studies tune a classifier for accuracy on a held-out split, attach a post-hoc explanation, and report the single best configuration on one cohort. Three practical questions are left open. Does a Digital-Twin representation of student state actually add value over a plain set of LMS features? Are the explanations steady enough to show a teacher? And can the whole loop be built and reproduced on real, public data?

We address these questions with a working prototype and an evaluation that's honest about what it finds. The central object is the student modelled as a dynamic digital twin: a weekly state snapshot at the grain of one row per student per week. We use the term Digital Twin deliberately, but narrowly. Here it means a lean, time-aware state representation, not a full counterfactual simulation engine. The twin keeps a descriptive weekly state and carries a lightweight what-if layer that estimates how the predicted grade would move if a flagged factor reached the cohort median. We treat mechanistic simulation and causal counterfactual reasoning as explicit boundaries rather than implied promises. Around this representation we place a gradient-boosting predictor, a perturbation-based explainable-AI (XAI) layer, and a Next.js teacher interface that reads frozen prediction artifacts.

Training gradient boosting or random forests on the Open University Learning Analytics Dataset (OULAD) and bolting on SHAP explanations after the fact is by now routine [1], [2], [3], [4]. Reported benchmarks reach ROC-AUC 0.993 and F1 0.911 on dropout prediction [1], so adding explainability is no longer a contribution on its own. Continuous-grade regression, which we use alongside classification, is less common, but it isn't enough by itself. Tiukhova et al. [5] found that XAI-derived feature importance for student-success models can be unstable across configurations. We take that stability lens further, out of a single-model setting and into a transfer setting, and measure rank agreement across evaluation splits, across two OULAD courses, and across institutions. Digital

twins in education appear mostly as concept or review papers [6]; working implementations on real student data are essentially missing. Peer-reviewed teacher-facing work publishes model metrics rather than systems, and commercial platforms such as EAB Navigate, Civitas Learning, and Brightspace Insights ship dashboards that are closed and undocumented. Papers that add SHAP seldom discuss how a teacher might misread an explanation, or the gap between describing model behaviour and making a causal claim.

Research gap and key provisions. No published work builds and then honestly evaluates a full-cycle prototype that runs from raw data, through weekly twin snapshots and grade predictions, to per-student explanations shown in a teacher interface, on real public data, while documenting where the added components fail to beat a simpler baseline. This paper targets that gap. The novelty isn't a new algorithm or a higher accuracy number. It's the combination of four things: an end-to-end, open, leakage-aware pipeline from raw OULAD data to teacher-facing weekly explanations; a dependency-driven ablation that tests the Twin representation instead of assuming it; an explanation-stability analysis across splits, courses, and institutions; and an honest multi-cohort finding, including a documented synthetic-data circularity failure mode, of a kind the literature systematically under-reports.

Materials and Methods

Theoretical analysis: system architecture. The prototype is a monorepo with two services that run independently and a strict read-only boundary between computation and presentation. The ML service (services/ml, Python) holds the OULAD data adapter, the weekly snapshot builder, the feature-engineering pipeline, the experiment runner, the model-training code, and the perturbation-based XAI module. Every experiment writes versioned JSON and CSV artifacts under data/artifacts/experiments. The web service (apps/web, Next.js) is a server-rendered frontend that reads those frozen artifacts at request time and serves the teacher-facing views. It runs no machine learning of its own, so results stay reproducible regardless of the interface, and the interface can be tested without retraining anything. Data moves in one direction. Raw OULAD CSV files enter the adapter, become a weekly snapshot table with one row per student per week, pass through feature engineering into a trained Gradient Boosting model, and emerge as prediction and explanation artifacts in JSON that the web app consumes (Figure 1).

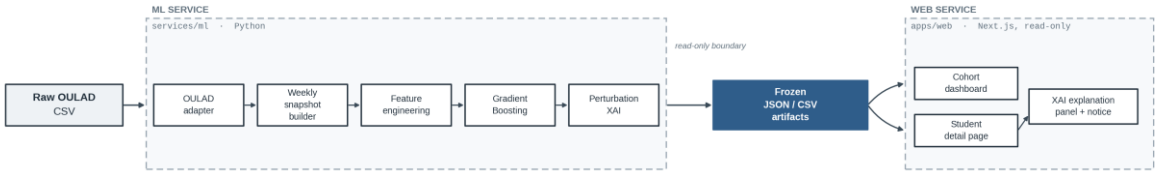

Figure 1 - End-to-end system architecture.

The teacher interface has two parts. A cohort dashboard lists every student in a sortable, filterable table with predicted grade, a pass-risk badge, the current week, and the key LMS signals, and supports one-click filtering to the high-risk group (Figure 2). A student detail page adds summary cards (predicted versus actual grade, risk, mastery, activity, attendance, assignment and quiz averages, and a three-week trend), a weekly trajectory chart, the raw weekly-snapshot table, an explanation panel listing which signals raise or lower the current-week prediction, and a what-if panel that estimates the grade gain if each flagged below-median factor reached the cohort median (Figure 3). The explanation panel carries an explicit limitation notice: the listed factors describe how the model behaves, not why a student is at risk; the score isn't a sole basis for intervention; and the rankings can move across time periods and cohorts. We present the interface as a design demonstrator, not an evaluated intervention, and we make no claim from a user study here.

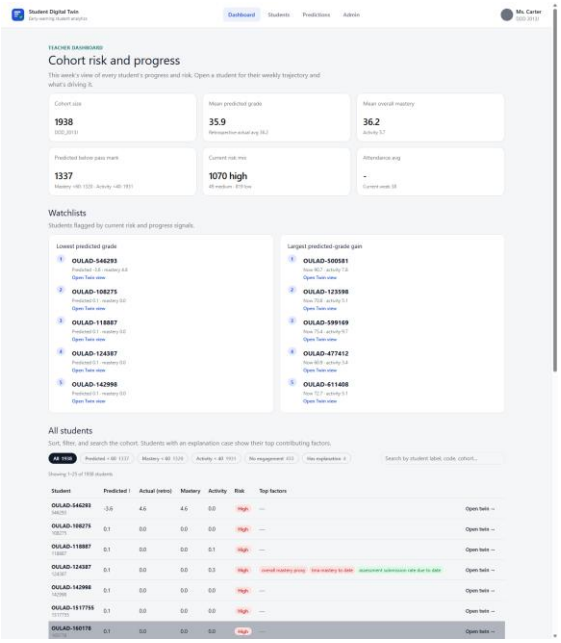

Figure 2 - Teacher cohort dashboard.

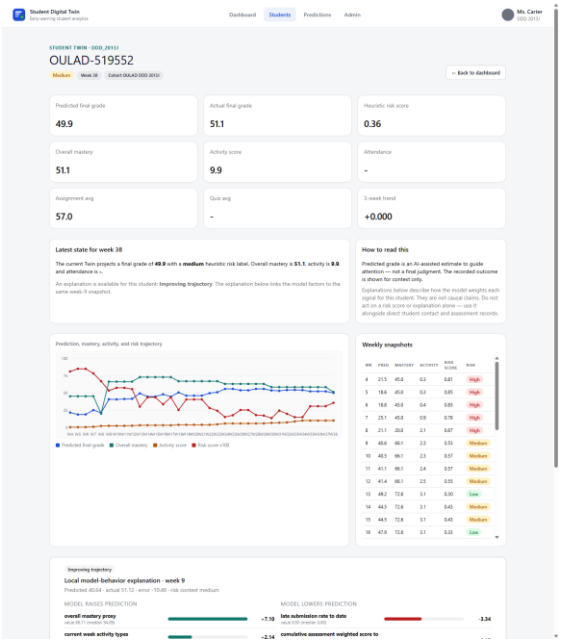

Figure 3 - Per-student detail view.

Experimental part: datasets. We evaluate on real, public, non-circular data from two institutions. OULAD DDD 2013J [7] contributes 1,938 students and 67,830 weekly snapshots over weeks 4 to 38. OULAD BBB 2013J [7] adds 2,237 students and 80,532 weekly snapshots, with a richer dated assessment structure than DDD. The KU Leuven 1819 dataset [8] pools two courses into 1,495 students over weeks 2 to 15. A separate synthetic dataset is used only for controlled methodological checks, and we state its limitations in the Results and Discussion.

Feature sets (nested ablation). We test the Digital-Twin representation by decomposition rather than assuming it. The four feature sets are nested, so each one strictly contains the previous: A_simple is a minimal baseline; B_lms adds the LMS behavioural features that stand in for what the literature typically uses; B_lms + mastery adds a lean mastery-proxy block derived from the dated assessment structure; and C_twin is the full engineered representation, with trends, mastery, indices, and temporal context (Figure 4). Because the sets are nested, any change in performance can be attributed to the block that was added rather than to an unrelated change of features.

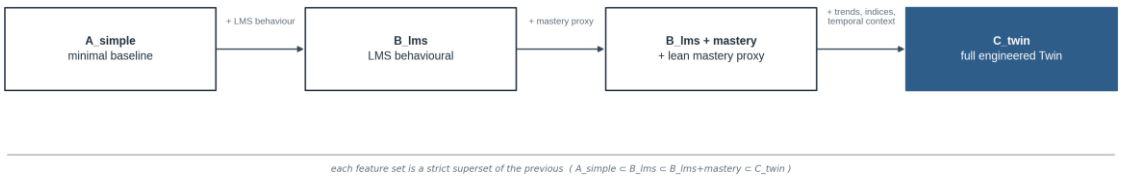

Figure 4 - Nested feature-set ablation.

Experimental part: models and protocol. The model families are deliberately simple and easy to defend: Ridge regression, Logistic Regression, Random Forest [9], and Gradient Boosting [10]. We use two splits. The primary one is a student-grouped split (test_size = 0.25, seed = 42; for DDD, 1,454 training against 484 test students), which keeps each student wholly in train or test. The second is a temporal-forward split that trains on early weeks and predicts later ones. The pipeline is leakage-aware throughout: preprocessing is fit on training data only, forbidden columns are enforced explicitly, and one model (Gradient Boosting) is held fixed across splits so that results aren't flattered by picking the best model per cell. The targets are final_weighted_score, a 0 to 100 regression target, and passed_observed, a binary classification target. Abbreviations follow first use: LMS, learning

management system; XAI, explainable artificial intelligence; OULAD, Open University Learning Analytics Dataset; RMSE, root mean squared error; ROC-AUC, area under the receiver-operating-characteristic curve.

Results and Discussion

This work versus standard baselines. All rows in Table 1 use OULAD DDD 2013J, the B_lms feature set, and the student-grouped split, except the final row labelled this work, which adds the mastery block. RMSE is on the 0 to 100 score scale, while F1 and ROC-AUC refer to passed_observed.

Table 1 - This work versus standard baselines on OULAD DDD 2013J

Approach	RMSE	F1	ROC-AUC	Per-student XAI	Teacher UI	Open
Logistic Regression (ours)	n/a	0.854	0.951	No	No	Yes
Random Forest (ours)	13.633	0.861	0.947	No	No	Yes
GB LMS-only (ours)	12.658	0.863	0.953	No	No	Yes
GB + Twin + XAI (this work)	**12.724**	**0.861**	**0.953**	**Yes**	**Yes**	**Yes**
GB + SHAP (literature) [1]	n/a	0.911	0.993	Yes	No	Partial
Commercial (EAB Navigate)	unknown	unknown	unknown	Partial	Yes	No

Gradient Boosting on B_lms is the strongest baseline, with RMSE 12.658, F1 0.863, and ROC-AUC 0.953; Ridge regression is the weakest at RMSE 14.318. Logistic Regression is already competitive (F1 0.854, ROC-AUC 0.951), which fits the fact that the dominant predictor, assessment submission rate, is roughly linear in the log-odds of passing. Adding the mastery block moves RMSE by +0.066, slightly worse, and F1 by −0.002. On this cohort and split it brings no improvement. We report that plainly: it's the negative result that motivates the rest of the analysis.

When does the Twin help, and when not?

The full nested ablation, under fixed-model reporting, gives an answer that depends on the course and the split (Table 2).

Table 2 - Twin benefit by cohort and split (RMSE deltas vs. B_lms; negative = improvement)

Cohort	Student-grouped	Temporal-forward
DDD 2013J	null (mastery +0.061)	weak: mastery −0.381
BBB 2013J	null (all within 0.087)	largest gain: mastery −1.026

On DDD, no Twin block beats B_lms by more than one RMSE point on either split. The full C_twin stays within ±0.025 RMSE of the baseline, which is non-inferior rather than better, and the minimal A_simple is clearly worse (+1.207 grouped, +4.105 temporal-forward). The LMS layer, in other words, already carries the signal. On BBB, which has a richer assessment structure, the mastery block improves the temporal-forward split by 1.026 RMSE, dropping it from 6.311 to 5.284, and the full C_twin improves it by 0.765. Yet the student-grouped split stays null, and the trend and index blocks are null on both courses.

So the value of engineered Twin features is heterogeneous. It appears only under forward-time prediction on a course with rich assessment structure, and it disappears under the stricter student-grouped regime (Figure 5). To check that the one positive cell isn't an artefact of a lucky split, we ran a student-clustered bootstrap: 5,000 resamples of the 559 held-out test students, seed 42, re-executing the frozen BBB pipeline. Because gradient boosting is sensitive to its software environment, the reproduced point estimate differs from the stored one (−2.16 against −1.03 RMSE), but the 95%

confidence interval of the mastery RMSE reduction sits entirely below zero ([−2.54, −1.80]; the reduction held in all 5,000 resamples). The direction, then, isn't noise, even though the magnitude is tied to the out-of-time extrapolation regime, so we treat this as the single cell where Twin features clearly help rather than as a stable effect size. The OULAD classification target also stays genuinely predictive throughout (best-model F1 between 0.83 and 0.93, never 1.000), which is what makes the negative finding trustworthy.

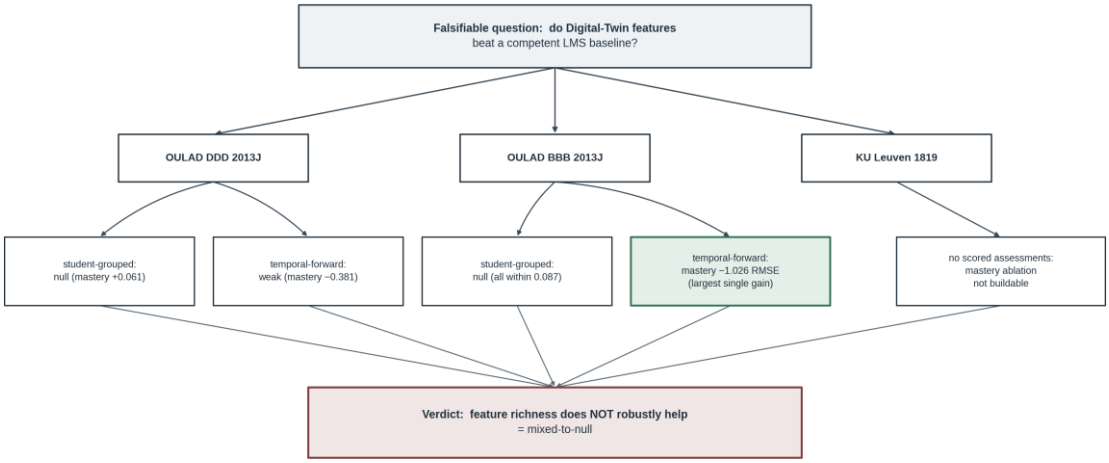

Figure 5 - Evaluation arc and overall verdict.

XAI and explanation stability. Explanations rely on held-out permutation importance, model-native importance, and a one-feature local median-replacement perturbation. We avoid SHAP [11] on purpose. Shapley-value attributions depend on a background or conditional-expectation estimate that's fragile on the small per-student samples surfaced in the teacher UI, and they carry documented conceptual and robustness problems as feature-importance measures [12], [13]. The perturbation method instead works directly on the single prediction row a teacher sees, with no background-set assumption. On DDD the dominant factor is the assessment submission rate due to date (importance share 0.28 to 0.38, and 0.381 in the headline model). Once mastery is included, the co-circular overall mastery proxy joins it (0.23 to 0.43). The genuinely exogenous signals, whether a student is unregistered by a given week (0.13 to 0.20) and VLE clickstream (0.02 to 0.06), are interpretable but carry modest weight (Table 3).

Table 3 - Top global importance shares (DDD 2013J, fixed model)

Feature set	Rank	Feature	Importance share
B_lms	1	activity_score_to_date	0.701
B_lms	2	avg_assignment_score_to_date	0.145
B_lms	3	avg_quiz_score_to_date	0.086
B_lms + mastery	1	activity_score_to_date	0.648
B_lms + mastery	2	overall_mastery	0.178
B_lms + mastery	3	avg_assignment_score_to_date	0.090

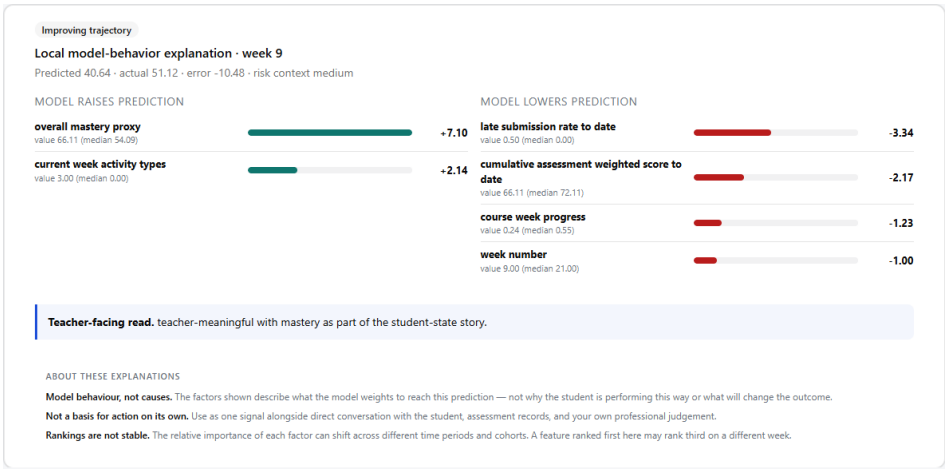

Figure 6 - Per-student explanation panel.

The rankings aren't invariant. Within a course, across splits, Kendall τ is 0.79, 0.68, and 0.55 for B_lms, +mastery, and C_twin (mean 0.67); none reach 0.90. Across courses, comparing DDD with BBB, τ falls to a range of 0.32 to 0.61 (mean 0.52), less stable still. Across three institutions, on an engagement-only task, the mean τ is about 0.56; clicks and active days keep recurring as the top drivers, but the middle and lower ordering is institution-specific (Figure 7).

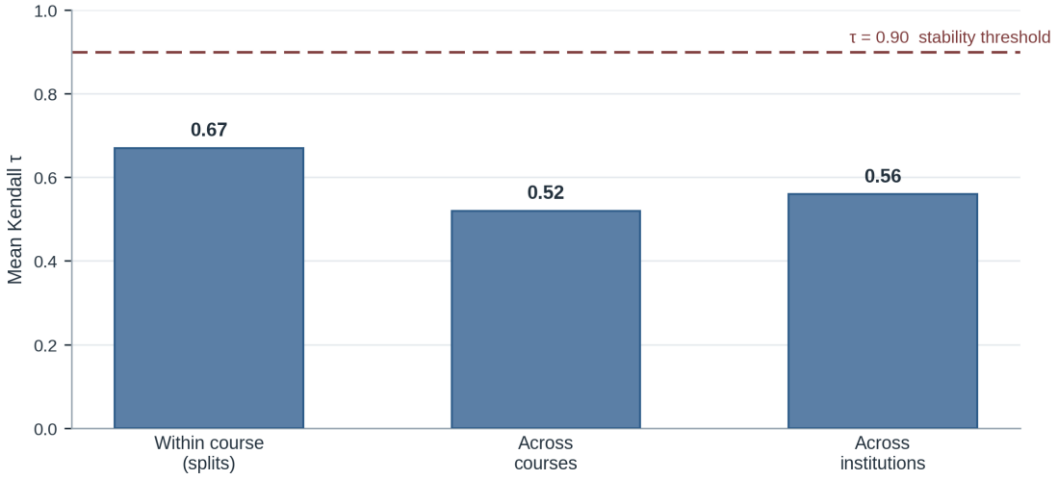

Figure 7 - Explanation-stability summary.

A controlled synthetic faithfulness probe confirms that the method is faithful when the ground truth is known: at the final course week, permutation importance recovers the generator's weight ordering exactly (Kendall τ = 1.0). The same few features dominate across regimes, but their relative priority reorders, which is the practical point for anyone putting these rankings in front of a teacher.

Second institution. KU Leuven exposes clickstream and course structure but has no intermediate scored assessments and no continuous grade, only a binary PASSED outcome. The mastery and Twin ablation therefore can't be built there at all, which is itself a cross-institution heterogeneity result. On its engagement-only task, prediction is modest (F1 0.75 to 0.76, ROC-AUC 0.65 to 0.72), and a richer engagement set does not beat a minimal two-feature one (fixed-model F1 deltas of −0.015 and +0.000). A matched three-institution synthesis confirms that a two-feature engagement baseline of clicks plus active days is hard to beat, with F1 deltas spread between −0.016 and +0.026.

Discussion. Three points follow. First, on the central question, whether Twin features beat a competent LMS baseline, the honest answer across three real cohorts and two institutions is: not robustly. The benefit is real but narrow, confined to BBB, the temporal-forward split, and the mastery block, and it vanishes under the stricter student-grouped regime, on other courses, and on other

blocks. Reporting only the favourable cell, as is common, would have manufactured a Twin wins narrative that the rest of the evidence does not support. Second, explanations are regime-dependent. Importance rankings reorder across splits, courses, and institutions, with a direct product consequence: a teacher UI has to present XAI output as how the model weights each signal for this prediction, not why this student is at risk. Our interface encodes exactly that caveat. Third, the contribution is best read on three axes rather than accuracy alone. On accuracy this work is comparable to standard baselines and below the published best (ROC-AUC 0.953 against 0.993 [1]), which is expected and not a weakness. On explainability it offers per-student, background-free perturbation explanations with documented stability limits. On teacher-facing completeness it's, as far as we know, the most complete openly documented artifact of its kind.

Limitations and threats to validity. During development a synthetic target manufactured an apparent mastery win. The synthetic final grade is a deterministic, noise-free closed-form weighted mean of the same behaviours the features re-aggregate (week-10 reconstruction error 0.008, correlation 1.000000), so the resulting R^2 near 0.99 and the saturated passed F1 of 1.000 are algebraic artifacts, not learnable signal. We report this openly as a cautionary methodological result. On OULAD, final_weighted_score is partly fed by assessment scores, so the regression results lean somewhat on within-system accounting; the classification target and the VLE clickstream are the cleaner, exogenous evidence. Overall mastery is close to cumulative LMS performance ($|r| = 0.993$ with average assignment score) and must not be presented as an independent construct. The teacher interface is a design demonstrator with no user evaluation. The evidence spans two OULAD courses at one institution plus a second institution that cannot test the mastery ablation, and the explanations are neither causal nor SHAP. The what-if layer is a descriptive, perturbation-based estimate, not a mechanistic simulation or a causal claim.

Conclusions

We presented an explainable, teacher-facing Student Digital Twin and evaluated it honestly across two institutions. On real, non-circular OULAD data the engineered Twin features don't deliver a consistent advantage over a competent LMS baseline. The result is mixed-to-null on DDD 2013J and heterogeneous on BBB 2013J, where mastery improves the temporal-forward split by 1.026 RMSE and nowhere else, and a second institution can't even build the ablation. The accompanying explanations are regime-sensitive within a course (Kendall τ 0.55 to 0.79) and partly course-specific across courses (0.32 to 0.61). The contribution is therefore methodological and system-level: a reproducible, leakage-aware pipeline with a teacher interface, an explanation-stability analysis, a documented synthetic-circularity failure mode, and an honest multi-cohort finding, rather than an accuracy record. Two practical recommendations follow. Teacher-facing dashboards should present XAI output as model behaviour under a stated evaluation scenario, and they shouldn't assume that richer engineered features improve prediction without per-cohort verification. Future work includes external validation across more institutions and validated intervention and counterfactual reasoning.

Acknowledgements

This research received no external funding. The authors thank the providers of the public datasets used. All datasets are public and openly licensed: OULAD is distributed under CC-BY 4.0 [7], and the KU Leuven activity and performance dataset is distributed under CC-BY 4.0 via Zenodo [8]. The authors aren't affiliated with, and report no competing interest in, the dataset providers. All experiment code, configuration, and frozen result artifacts (versioned JSON and CSV per experiment) are available in the project repository at [repository URL] to support independent reproduction.

References

1. J. López de la Rosa et al., “A Modular and Explainable Machine Learning Pipeline for Student Dropout Prediction in Higher Education,” Algorithms, vol. 18, no. 10, art. 662, 2025, doi: 10.3390/a18100662.
2. S. Boujmiraz, H. Darhmaoui, and A. Drissi el Maliani, “Predicting student performance: a comprehensive

review of machine learning, deep learning, and explainable AI approaches," Computers and Education: Artificial Intelligence, 2026, doi: 10.1016/j.caeai.2026.100396.
3. F. T. Johora, M. N. Hasan, A. Rajbongshi, M. Ashrafuzzaman, and F. Akter, "An explainable AI-based approach for predicting undergraduate students' academic performance," Computers and Education: Artificial Intelligence, 2025, doi: 10.1016/j.caeai.2025.100376.
4. W. Villegas-Ch et al., "Machine learning models for academic performance prediction: interpretability and application in educational decision-making," Frontiers in Education, vol. 10, art. 1632315, 2025, doi: 10.3389/feduc.2025.1632315.
5. E. Tiukhova, P. Vemuri, N. López Flores, A. S. Islind, M. Óskarsdóttir, S. Poelmans, B. Baesens, and M. Snoeck, "Explainable Learning Analytics: Assessing the stability of student success prediction models by means of explainable AI," Decision Support Systems, vol. 182, art. 114229, 2024, doi: 10.1016/j.dss.2024.114229.
6. M. Furini, O. Gaggi, S. Mirri, M. Montangero, E. Pelle, F. Poggi, and C. Prandi, "Digital twins and artificial intelligence as pillars of personalized learning models," Communications of the ACM, vol. 65, no. 4, pp. 98–104, 2022, doi: 10.1145/3478281.
7. J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," Scientific Data, vol. 4, art. 170171, 2017, doi: 10.1038/sdata.2017.171.
8. E. Tiukhova, D. Van Landuyt, B. Baesens, and M. Snoeck, "Open data, private learners: a de-identified student activity and performance dataset for learning analytics," Scientific Data, 2026, Zenodo, doi: 10.5281/zenodo.17087849.
9. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
10. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
11. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 4765–4774.
12. I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," in Proc. 37th Int. Conf. Machine Learning (ICML), 2020, pp. 5491–5500.
13. D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), 2020, pp. 180–186.

А. Талғатов, С.А. Нургалиева*
Магистр студенті, Бағдарламалық инженерия мектебі, Astana IT University, Астана, Қазақстан
PhD, Бағдарламалық инженерия мектебінің ассистент профессоры, Astana IT University, Астана, Қазақстан
***Корреспондент авторы:** symbat.nurgaliyeva@astanait.edu.kz

МҰҒАЛІМГЕ БАҒЫТТАЛҒАН СТУДЕНТТІҢ ЦИФРЛЫҚ ЕГІЗДЕРІНЕ АРНАЛҒАН ҚАЙТАЛАНАТЫН ТҮСІНДІРІЛЕТІН ЖИ КОНВЕЙЕРІ

Түйін

Оқытуды басқару жүйелері студент әрекетінің үлкен көлемін тіркейді, бірақ олардың кіріктірілген аналитикасы нәтижені болжаудан гөрі өткенді сипаттауға бейім. Біз әдейі терiске шығарылатын сұрақ қоямыз: студенттің цифрлық егізінің инженерлік белгілері әдеттегі LMS сигналдарына негізделген құзыретті базалық модельмен салыстырғанда баға болжамын жақсарта ма және алынған түсіндірмелер мұғалімге көрсетуге жеткілікті тұрақты ма? Жауап алу үшін студент күйінің апталық көрінісін градиенттік бустинг моделімен және пертурбацияға негізделген түсіндірме қабатымен біріктіретін, ағып кетуден қорғалған конвейер құрылды; содан кейін екі мекеменің ашық деректерінде кірістірілген белгілер абляциясы жүргізілді. Нәтиже ойландырады: белгілер молдығын қосу тұрақты түрде көмектеспейді — бірде-бір егіз блогы негізгі когортада базалық модельден бір RMSE пунктінен артық озбайды, тек бір ұяшық айқын өсім береді, ал түсіндірме рангтары бағалау

режиміне қарай өзгереді (Kendall τ 0,32-ден 0,79-ға дейін).

Кілттік сөздер: оқу аналитикасы; түсіндірілетін ЖИ; түсіндірмелердің тұрақтылығы; үлгерімді болжау; цифрлық егіз; градиенттік бустинг; қайталанушылық

А. Талгатов, С.А. Нургалиева*
Студент-магистрант, Школа программной инженерии, Astana IT University, Астана, Казахстан
PhD, ассистент профессор школы программной инженерии, Astana IT University, Астана, Казахстан
***Автор для корреспонденции:** symbat.nurgaliyeva@astanait.edu.kz

ВОСПРОИЗВОДИМЫЙ КОНВЕЙЕР ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ЦИФРОВЫХ ДВОЙНИКОВ СТУДЕНТОВ, ОРИЕНТИРОВАННЫХ НА ПРЕПОДАВАТЕЛЯ

Аннотация

Системы управления обучением фиксируют большие объёмы студенческой активности, однако их встроенная аналитика чаще описывает прошлое, чем прогнозирует результат. Мы ставим намеренно опровергаемый вопрос: улучшают ли инженерные признаки цифрового двойника студента прогноз оценки по сравнению с компетентной базовой моделью на обычных LMS-сигналах и достаточно ли устойчивы получаемые объяснения, чтобы показывать их преподавателю. Для ответа построен защищённый от утечек конвейер, объединяющий недельное представление состояния студента с моделью градиентного бустинга и слоем объяснений на основе пертурбаций; затем проведена вложенная абляция признаков на открытых данных двух учреждений. Результат отрезвляющий: добавление богатства признаков не помогает устойчиво — ни один блок двойника не превосходит базовую модель более чем на один пункт RMSE на основной когорте, лишь одна ячейка даёт явный прирост, а ранги объяснений меняются вместе с режимом оценивания (Kendall τ от 0,32 до 0,79).

Ключевые слова: учебная аналитика; объяснимый ИИ; устойчивость объяснений; прогнозирование успеваемости; цифровой двойник; градиентный бустинг; воспроизводимость