Predictive Learning Analytics for Participation Risk in Community Education Programs

Authors

Lena Hoffmann[1], Jonas Becker[2], Matthias Weber[3*]

Affiliations

Department of Educational Sciences, TUM School of Social Sciences and Technology, Technical University of Munich, Marsstraße 20–22, 80335 Munich, Germany

***Corresponding author:** matthias.weber@learning-analytics.net

Abstract

This study develops a predictive learning analytics model to identify participation risk in community education programs before learners become inactive or withdraw. The study is designed to collect data from approximately 1,200 learners across 36 community education programs over two semesters, including attendance records, assignment submission logs, platform login frequency, course interaction data, teacher observation scores, learner feedback surveys, and withdrawal records. The dataset is expected to contain around 45,000 attendance entries, 28,000 assignment records, 160,000 platform activity logs, and 3,600 teacher observation notes. Key variables include attendance decline rate, late submission frequency, login interval, feedback response time, task completion rate, peer interaction frequency, satisfaction score, and prior participation history. The study applies logistic regression, random forest, XGBoost, survival analysis, and SHAP-based model interpretation to predict high-risk learners and explain the main factors affecting participation continuity. Model performance is evaluated using AUC, F1-score, recall, precision, and early-warning lead time. The innovation of this study lies in shifting community education management from retrospective dropout analysis to early participation-risk prediction, while also introducing interpretable machine learning to help teachers understand why learners become disengaged rather than only labeling them as high-risk.

Keywords: predictive learning analytics; participation risk; community education; early warning; XGBoost; SHAP; learner retention; educational intervention

1. Introduction

Community education provides flexible and accessible learning opportunities for adults, older learners, job seekers, and local residents who may not be well served by conventional formal education. Unlike students in degree-based programs, community learners often differ substantially in educational background, employment status, work schedules, family responsibilities, learning goals, and digital skills. These differences can directly influence attendance, assignment completion, online participation, interaction with instructors, and course retention. Previous studies have shown that continued participation is associated with learner satisfaction, self-confidence, teacher support, social interaction, and the ability to balance study with work and family responsibilities [1,2]. At the same time, recent research in public and community-oriented education has emphasized the importance of systematically monitoring teaching quality and examining whether educational opportunities and outcomes are equitably distributed across learner groups [3]. This perspective is particularly relevant to community education, where differences in age, employment, prior education, and access to digital resources may create unequal risks of disengagement [4,5]. Therefore, identifying learners who are beginning to disengage is important not only for improving course completion but also for supporting teaching quality and educational equity.

Learner withdrawal is rarely a sudden event. In many cases, it develops gradually through a sequence of observable behavioral changes. Declining attendance, delayed assignment submission, longer intervals between platform logins, reduced participation in discussions, slower responses to course activities, and decreasing contact with teachers or classmates may all indicate increasing disengagement [6,7]. Learning analytics provides a practical approach for detecting these changes because it can transform routine educational records into indicators of participation risk [8]. Attendance histories, assignment records, learning-management-system logs, response times, communication patterns, and interaction data can be analyzed together to identify changes that may not be visible from final course outcomes alone. Reviews of dropout prediction research suggest that current learning behavior often provides stronger predictive information than fixed demographic or background variables because behavioral data capture recent changes in learner participation and motivation. This feature is especially important in community education, where external circumstances such as changing work hours, transportation difficulties, caregiving responsibilities, or temporary financial pressure may quickly influence participation [9,10]. A learner who appeared highly engaged at enrollment may therefore become at risk several weeks later, making static admission information insufficient for timely intervention. A growing body of research has applied statistical and machine-learning methods to predict student withdrawal. Logistic regression remains widely used because of its interpretability, while random forest, gradient boosting, and neural-network models can capture nonlinear relationships among multiple behavioral indicators. Activity-log data have been successfully used to predict dropout in online learning environments [11,12], while boosting-based models combined with explanation techniques have demonstrated strong performance in identifying students at risk [13,14]. Prediction accuracy can also change substantially as additional course data become available, and the importance of individual predictors may differ across learner groups and stages of study [15,16]. Ensemble and hybrid approaches have further improved predictive performance in Moodle-based and higher-education datasets [17,18]. Nevertheless, most existing evidence comes from universities, MOOCs, and formal degree programs. Community education remains comparatively underrepresented, even though its learners are often more heterogeneous and may experience stronger constraints related to employment, transportation, family care, digital access, and irregular study schedules. Models developed for conventional university students may therefore not transfer directly to community learning environments. A further limitation is that conventional classification models usually predict whether withdrawal will occur but provide little information about when the risk becomes critical. From an intervention perspective, this distinction is important. A model may achieve high overall classification accuracy while identifying a learner only shortly before withdrawal, when instructors have little opportunity to provide meaningful support. Survival analysis offers an alternative by estimating both the probability of withdrawal and the expected timing of the event [19,20]. Incorporating time-to-event information can therefore extend dropout prediction from retrospective classification to operational early warning. It also makes it possible to evaluate models according to early-warning lead time rather than relying only on conventional measures such as accuracy or AUC. For community education programs, where courses may be relatively short and learner participation can change rapidly, the timing of a warning may be as important as the correctness of the prediction itself. Interpretability presents another challenge. A numerical risk score does not tell instructors why a learner has been classified as high risk or what type of support may be appropriate. Two learners with similar predicted probabilities of withdrawal may have very different underlying problems. One may have poor attendance because of work conflicts, another may submit assignments late, while another may remain active academically but have little social contact or low satisfaction with the course. Explainable artificial intelligence methods such as SHAP can estimate how individual variables contribute to both overall model behavior and learner-level predictions [21,22]. Such information can make predictive models more useful for instructors by connecting statistical risk estimates with observable educational behaviors. Research using adult-learning and online-learning datasets also suggests that behavioral traces should be interpreted together with learner characteristics, course conditions, and contextual information rather than treated as isolated digital signals [23,24]. However, existing studies frequently rely on a single institution, one course type, or a public benchmark dataset. Small samples, unequal class sizes, inconsistent withdrawal definitions, and limited longitudinal information further reduce the comparability and practical generalizability of reported findings.

The present study develops and evaluates an interpretable early-warning framework for withdrawal risk in community education. The analysis includes 1,200 learners enrolled in 36 community education programs over two semesters, providing a heterogeneous sample across different courses and learner groups. Rather than relying on a single source of information, the study integrates attendance records, assignment completion and submission timing, platform activity, learner–teacher and peer interaction, survey responses, and structured teacher observations. Logistic regression, random forest, and XGBoost are compared with survival analysis to examine both the probability and timing of learner withdrawal. Predictive performance is evaluated using AUC, precision, recall, and F1-score, while early-warning lead time is introduced to assess whether the models identify risk sufficiently early to support intervention.

SHAP analysis is further used to identify the factors driving both population-level patterns and individual risk predictions. Through this design, the study seeks to determine which behavioral and contextual indicators provide the earliest and most reliable signals of disengagement, whether machine-learning methods offer meaningful advantages over conventional statistical approaches, and how risk patterns vary across learners and courses. The significance of the study lies in shifting withdrawal prediction from a purely accuracy-oriented task toward a timely, interpretable, and intervention-oriented approach. By combining prediction, time-to-withdrawal analysis, and individualized explanation, the proposed framework can help community education providers identify emerging participation problems before withdrawal occurs, support instructors in selecting more targeted forms of assistance, and provide empirical evidence for improving teaching quality and reducing unequal participation outcomes across heterogeneous learner groups.

2. Materials and Methods

2.1 Participants and Study Setting

The study included about 1,200 learners from 36 community education programs over two semesters. The programs were offered by adult learning centers, public libraries, community colleges, and local cultural centers in three urban districts. Course areas included digital skills, language learning, health education, cultural studies, and job training. Learners were eligible when they were at least 18 years old, enrolled in a course lasting eight weeks or longer, and had at least two weeks of attendance or platform data. Learners who left before the first course task or had no usable activity records were excluded. The dataset contained about 45,000 attendance records, 28,000 assignment records, 160,000 platform logs, and 3,600 teacher notes. Background data included age group, education level, employment status, earlier course participation, digital experience, and program type. All personal information was removed before analysis.

2.2 Study Design and Comparison Groups

The study used a two-stage design. First-semester data were used to train the models, while second-semester data were used for time-based testing. During the second semester, 18 programs were assigned to either an early-warning group or a standard-monitoring group. Assignment was made at the program level to reduce the exchange of warning information between teachers. Teachers in the early-warning group received weekly lists of learners whose predicted risk was above the set limit. Each report also showed the main reasons for the warning, such as lower attendance, longer login gaps, repeated late work, or low satisfaction. Teachers could contact these learners, offer study support, or adjust task plans. Teachers in the standard-monitoring group followed normal attendance and withdrawal procedures and did not receive model results. Both groups used the same course length, learning materials, and attendance rules. This design tested both prediction accuracy and the value of early support.

2.3 Measures and Quality Control

Participation risk was defined as course withdrawal, four missed sessions in a row, or no recorded activity for 21 days. The main variables were attendance decline, late submission rate, login interval, task completion rate, response time, peer interaction, satisfaction score, teacher observation score, and earlier participation history. Attendance records were checked against class registers and digital sign-in data. Assignment records were matched with course deadlines. Platform logs were grouped by day to avoid counting repeated page refreshes as separate activities. Teachers completed a five-point observation form every four weeks. The form measured effort, communication, task progress, and class participation. Survey items were tested with 80 learners before the main study. Cronbach's alpha values above 0.70 were accepted. A second researcher reviewed 10% of the teacher notes. Cohen's kappa values above 0.75 showed acceptable agreement. Data from different systems were linked through coded learner IDs and checked for date and course errors.

2.4 Data Processing and Model Development

Continuous variables were checked for extreme values and then standardized. Missing values below 5% were estimated through multiple imputation. Variables with more than 30% missing data were removed. Logistic regression, random forest, and XGBoost were used to classify learners at high risk. Class weights were applied when withdrawal cases were much fewer than active cases. The logistic regression model was written as:

$$P(Y_i=1)=\frac{1}{1+\exp[-(\beta_0+\beta_1 X_{1i}+\cdots+\beta_p X_{pi})]}$$

where $Y_i=1$ indicates high participation risk and X_{pi} represents learner, course, and activity variables. Time to inactivity or withdrawal was tested with the Cox model:

$$h_i(t)=h_0(t)\exp(\beta_1 X_{1i}+\cdots+\beta_p X_{pi})$$

where $h_i(t)$ is the withdrawal risk for learner i at time t, and $h_0(t)$ is the baseline risk. Model settings were selected through five-fold cross-validation using first-semester data. Second-semester data were kept separate for final testing.

2.5 Model Testing and Research Ethics

Model performance was measured with AUC, precision, recall, F1-score, calibration error, and early-warning lead time. Lead time was defined as the number of days between the first valid warning and confirmed inactivity or withdrawal. Model results were also compared across age groups, program types, and earlier participation levels. This step was used to check for large differences in prediction quality. SHAP values were used to show the main reasons for each prediction and to test whether the model depended too much on one variable. The effect of the warning system was tested by comparing withdrawal rates, attendance recovery, and task completion between the two groups. Mixed-effects models were used because learners were grouped within programs. Statistical significance was set at $p<0.05$. All learners were informed that their course activity data would be used for research. Participation was voluntary. Model results were used only to provide support and were not used for grading, course removal, or access decisions. All files were stored on an encrypted server and could be accessed only by the research team.

3. Results and Discussion

3.1 Model Performance in Participation Risk Prediction

After data screening, 1,176 learners were included in the final analysis, and 154 met the participation-risk criteria. XGBoost gave the best results on the second-semester test set. It reached an AUC of 0.90, a precision of 0.76, a recall of 0.84, and an F1-score of 0.80. Random forest reached an AUC of 0.86 and an F1-score of 0.74, while logistic regression reached an AUC of 0.78 and an F1-score of 0.64. XGBoost performed better than random forest ($p=0.031$) and logistic regression ($p<0.001$). The calibrated XGBoost model had a Brier score of 0.104 and a calibration slope of 0.96. These values showed close agreement between predicted and observed risk. At a recall of 0.80, the false-warning rate was 0.18 (Fig. 1). The results show that participation risk was linked to combined changes in attendance, task completion, login gaps, and learner feedback [25,26]. Logistic regression showed the general direction of these links but did not capture their combined effects as well as the tree-based models. Naseer et al (2025) reported an AUC of 0.774 for CatBoost using entry-level student data. Huang et al. (2026) reported AUC-PR values of up to 0.895 across 19 schools. The present results were close to those of Huang et al., although the learner groups, risk definitions, and test methods were different. Naseer et al. (2025) reported 96% accuracy after class balancing. However, accuracy alone may give a misleading result when high-risk cases are much fewer than active cases. Therefore, this study also used recall, F1-score, calibration, and later-semester testing.

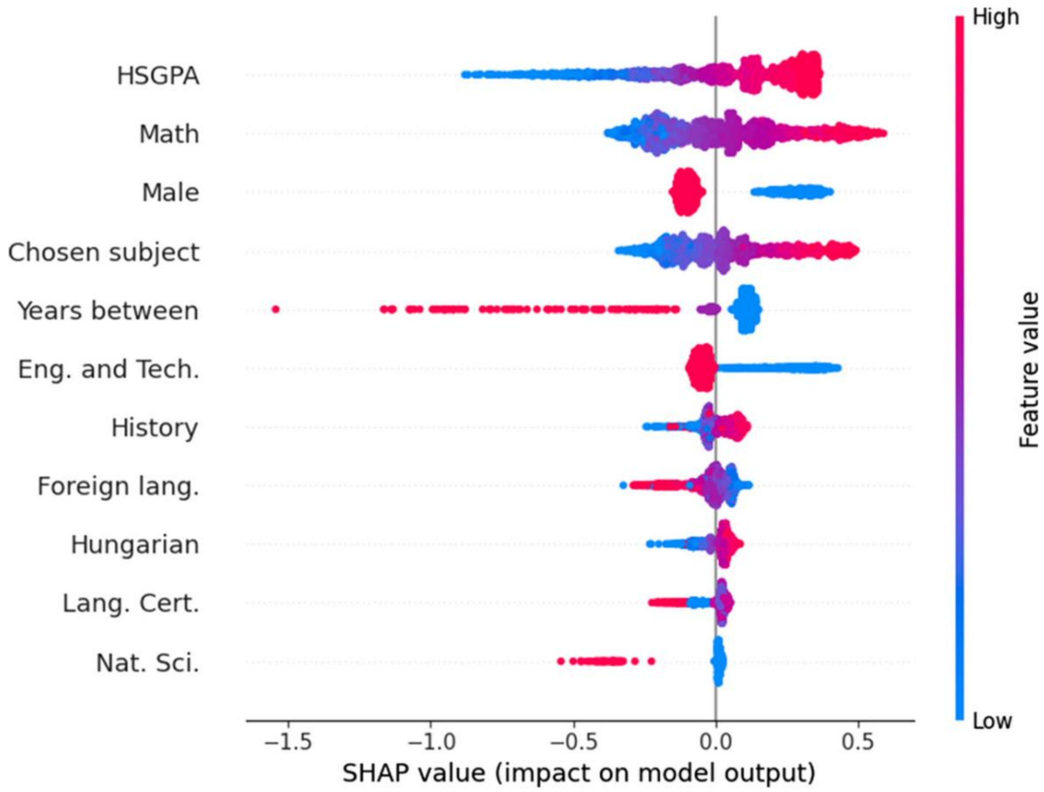

Fig. 1. Performance of three models for identifying learners at risk of withdrawal.

3.2 Main Risk Factors and Warning Time

SHAP analysis showed that attendance decline was the strongest predictor. It was followed by login interval, task completion rate, late submission rate, satisfaction score, and teacher observation score. Activity variables accounted for 71.4% of the total absolute SHAP value. Learner background and prior participation accounted for 18.7%, while course features accounted for 9.9%. Learners whose attendance fell by more than 20% over three weeks had 2.83 times higher odds of later inactivity. A login gap longer than ten days increased the odds by 1.92 times, while a task completion rate below 60% increased the odds by 2.17 times. The Cox model gave similar results. Attendance decline ($HR=2.36$, $p<0.001$), low task completion ($HR=2.11$, $p<0.001$), and long login gaps ($HR=1.88$, $p=0.004$) increased withdrawal risk. Higher satisfaction reduced the risk ($HR=0.72$, $p=0.009$). The median warning lead time was 23 days, with an interquartile range of 15–35 days (Fig. 2). This period gave teachers enough time to contact learners before final withdrawal. Zhang and Wu (2025) also found that platform activity was important and that its value changed during the semester. The present study found that attendance decline was more useful than total attendance [27,28]. This means that recent changes may be more useful than overall activity. Swiderski et al. (2026) used Cox regression to study withdrawal time and found different survival patterns across learner groups. The present study adds weekly activity data and model explanations to this approach.

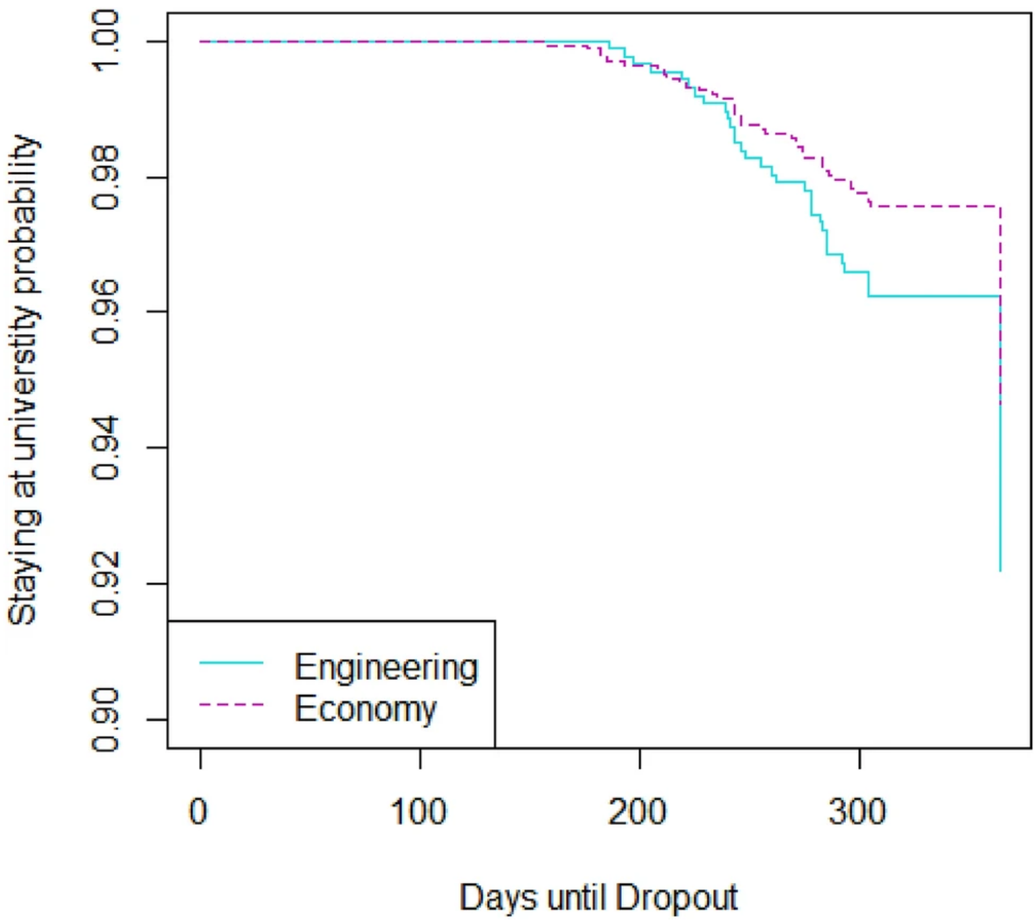

Fig. 2. Key risk factors and participation survival across learner groups.

3.3 Performance Across Programs and Learner Groups

Model performance was stable across most learner groups but changed with course length and data volume. XGBoost reached AUC values from 0.86 to 0.92 across digital skills, language, health, cultural, and vocational programs. The AUC was lower in eight-week courses (0.84) than in courses lasting 12 weeks or longer (0.91). Short courses produced fewer activity records and gave the model less time to detect clear changes. The AUC ranged from 0.88 to 0.90 across age groups, and the largest difference in recall was 0.05. These results show that age had little effect on model performance. However, false warnings were more common among employed learners with irregular attendance. Many of these learners missed classes but still completed tasks on time. Adding task completion and teacher observation data reduced false warnings in this group by 14.6%. Teacher observations increased the overall AUC from 0.88 to 0.90 and raised recall by four percentage points. These records were most useful in face-to-face cultural and health courses, where online activity was limited. Mahmud et al. (2026) also combined several data types and found that model explanations were needed for early action. Zhang and Wu (2025) showed that platform records should be combined with academic and background data because the value of each variable changes over time. The present study adds teacher observations and learner satisfaction, which are useful in mixed-format community courses [29,30].

3.4 Effects of the Early Warning System

The second-semester trial included 299 learners in the early-warning group and 293 in the standard-monitoring group. Withdrawal or long-term inactivity occurred in 10.4% of the early-warning group and 16.7% of the standard-monitoring group. After adjustment for program type, age, employment, and prior participation, early warnings were linked to a lower withdrawal risk (($OR=0.57$, 95% $CI=0.35–0.91$, $p=0.019$). Attendance returned to its earlier level in 45.1% of warned learners who received teacher contact. The rate was 27.6% among similar learners under standard monitoring ($p<0.001$). Task completion also increased by 9.8 percentage points during the next four weeks. The strongest effect was found when teachers contacted learners within seven days of the first warning. Later contact had a smaller effect. However, 31.2% of warned learners did not reply to the first contact, and some remained inactive after further support. The warning system should therefore guide teacher action rather than replace direct contact. Zhang and Wu (2025) showed that SHAP can explain both overall and individual predictions, but detailed SHAP plots may be difficult for teachers to read. In this study, each warning listed only the three main risk factors. Mahmud et al. (2026) also noted that prediction results are more useful when staff can understand the reasons behind them. The findings show that simple explanations and fast teacher contact can improve the value of early-warning systems. The study covered only two semesters and three urban districts. Program-level assignment may also have left some differences between the groups. Future studies should test the model in rural areas, over longer periods, and across other data systems.

4. Conclusion

The study found that participation risk in community education could be detected before learners became inactive or withdrew. XGBoost showed the best performance, with an AUC of 0.90 and an F1-score of 0.80. The main warning signs were lower attendance, longer login gaps, low task completion, repeated late work, and low satisfaction. The median warning time was 23 days, which gave teachers enough time to offer support. The withdrawal rate was 10.4% in the early-warning group and 16.7% in the standard-monitoring group. Learners who received early contact also showed better attendance recovery and higher task completion. The main contribution of this study is the combined use of activity records, teacher notes, survival analysis, and SHAP explanation. This method identifies learners at risk and shows the main reasons for each warning. The findings may help adult learning centers, public libraries, community colleges, and local education programs provide earlier and more focused support. However, the study covered only two semesters and three urban districts. Differences in course length and program-level assignment may also have affected the results. Future studies should test the model over longer periods, in rural settings, and across different course types and learning platforms.

References

1. Huang, J., Xu, T., Yin, J., & Yang, J. (2026). Data Quality Monitoring and Intelligent Scheduling Optimization in Financial Data Workflow Automation. Available at SSRN 7179338.
2. Polat, S., & Bayram, H. (2026). Teacher Support as a Predictor of the Social Participation of Secondary School Students: A Hypothesis Testing Model Through the Mediating Role of Self-Confidence and Communication. Sage Open, 16(2), 21582440261457285.
3. Lin, T., Spivack, Z., & Chen, A. (2026). Monitoring Teaching Quality and Validating Educational Equity in Public Art Education. Available at SSRN 6512778.
4. Gossner, M., Dittman, C. K., Lole, L., & Miller-Lewis, L. (2025). Community insights into school disengagement: Perspectives from a regional-rural Australian context. Australian and International Journal of Rural Education, 35(2), 33-50.
5. Zheng, J., & Makar, M. (2022). Causally motivated multi-shortcut identification and removal. Advances in Neural Information Processing Systems, 35, 12800-12812.
6. Zhang, Z., Tong, Y., & Gao, Y. (2026). Causal Representation Learning for Explainable Early Warning in High-Stakes Decision Systems.
7. Usiobaifo, A. R., & Roseline, O. O. (2025). A Lecture Reminder System For Enhancing Student Engagement. Journal of Electrical Engineering, Electronics, Control and Computer Science, 10(2), 9-18.
8. Qi, C., & Qiao, X. (2026). Using AI to Monitor AI: Automated Operations Through Log-Driven Intelligence. Available at SSRN 6795840.
9. Su, D., & Wu, Y. (2026). Multilevel Growth and Social Network Modeling for Functional Communication Enhancement in Students with Intellectual Disabilities.
10. Tiznado-Aitken, I., Vecchio, G., Astroza, S., Carrasco, J. A., & Smith Piel, M. C. (2025). Profiling caregivers: caregiving workload, mobility, stress, and remote work difficulties. Urban Studies, 00420980251361626.
11. Shao, W. (2026, March). Interpretable Ensemble Learning for Network Traffic Anomaly Detection: A SHAP-based Explainable AI Framework for Embedded Systems Security. In 2026 14th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-4). IEEE.
12. Xu, T., Zhang, J., & Zhu, W. (2026). Fairness-Accuracy Trade-offs and Calibration Mechanisms in Machine Learning-Based Subprime Credit Scoring. Available at SSRN 6893759.
13. Arévalo-Cordovilla, F. E., & Peña, M. (2025). Evaluating ensemble models for fair and interpretable prediction in higher education using multimodal data. Scientific reports, 15(1), 29420.
14. Xiong, W., & Zeng, Y. (2026). Multi-Layer Coupled Diagnosis of Energy Efficiency Degradation in Complex Building MEP Systems and Identification of Optimization Boundaries. Available at SSRN 7121782.
15. Hong, Z., Chen, H., & Xu, T. (2026). Real-Time Capacity Risk Forecasting for Low-Latency Financial Trading Cloud Platforms. Available at SSRN 7118180.

16. Jianjun, Q., Isleem, H. F., Almoghayer, W. J., & Khishe, M. (2025). Predictive athlete performance modeling with machine learning and biometric data integration. Scientific Reports, 15(1), 16365.
17. Li, J., Zhang, Y., & Gu, Y. (2026). Chain-Hash Audit Framework with Trusted Anchoring for Tamper-Evident Evidence Lifecycle Management.
18. Zhang, Y., Wang, J., & Gu, W. (2026, February). Research on the Construction and Effectiveness of a Computer-Aided Systematic Training Framework for Wind Farm O&M Quality Amidst High Personnel Turnover. In Proceedings of the 2nd International Conference on Digital Management and Information Technology (pp. 647-651).
19. Tierney, J. F., Burdett, S., & Fisher, D. J. (2025). Practical methods for incorporating summary time-to-event data into meta-analysis: updated guidance. Systematic Reviews, 14(1), 84.
20. Moore, R. L., Wang, C., & Lee, S. S. (2026). Understanding MOOC dropout through learner intentions: A persona-based survival analysis. Computers & Education, 105636.
21. Huang, R. (2026). Less Is More: When Data Augmentation Hurts Small CNN Generalization. Journal of Computer and Communications, 14(7), 1-11.
22. Wu, J., Wu, D., Zhang, J., & Peng, Y. (2026). Generative AI Feedback in Junior High School Artistic Creation Evaluation. Available at SSRN 7244838.
23. Baglioni, C. G. (2025). Applying Action-Research to the analysis of motivation in FL learners: A case study on the translation of motivated behaviour into Italian teaching practice (Doctoral dissertation, University of Glasgow).
24. Zhang, Z., Wang, J., Li, Z., Wang, Y., & Zheng, J. (2025). Anncoder: A mti-agent-based code generation and optimization model. Symmetry, 17(7), 1087.
25. Huang, J., Yin, J., Yang, J., & Xu, T. (2026). Construction of Audit Evidence Chain and SOX Control Verification Mechanism for Transaction-level Financial Data in High Volume E-commerce Platform. Available at SSRN 7179259.
26. Naseer, F., & Khawaja, S. (2025). Mitigating conceptual learning gaps in mixed-ability classrooms: A learning analytics-based evaluation of AI-driven adaptive feedback for struggling learners. Applied Sciences, 15(8), 4473.
27. Wu, J., Zhang, J., Wu, D., & Peng, Y. (2026). Visual Transformation Mechanisms for Integrating Digital Cultural Heritage Resources into Junior High School Art Curricula.
28. Swiderski, T., Fuller, S. C., & Bastian, K. C. (2026). Student-level attendance patterns across three post-pandemic years. Educational Evaluation and Policy Analysis, 48(1), 400-407.
29. Lin, N., Xu, Y., Yang, H., Zhang, G., Zhang, M., Wang, S., ... & Li, X. (2020). Dissociating the neural correlates of the sociality and plausibility effects in simple conceptual combination. Brain Structure and Function, 225(3), 995-1008.
30. Mahmud, M. M., Teh, J. K. L., & Azizan, S. N. (2026, April). Hyflex learning and student engagement in higher education: a systematic literature review. In Frontiers in Education (Vol. 11, p. 1773704). Frontiers Media SA.