Journal of Emerging Multidisciplinary Research
Volume 2, Issue No. 3, Jul – Sep, 2026 ISSN Number: 3139-3292
Frequency: Quarterly

Explainable AI for Early Detection of Academic Disengagement in Blended Learning Environments

Abhilasha Dubey
Assistant Professor
School of Management and Commerce, Vikrant University
Gwalior, Madhya Pradesh, India
smc_abhilasha@vikrantuniversity.ac.in

Abstract*— Learning management systems generate detailed behavioural traces, and predictive models built on those traces can identify students at risk of failure or withdrawal well before conventional assessment does. As these models have moved from research prototypes toward institutional deployment, opacity has become the binding constraint: an instructor asked to act on an unexplained risk flag has no basis for choosing an intervention, and no basis for deciding whether the flag is wrong. Explainable artificial intelligence has been proposed as the remedy, with feature attribution methods now standard in the learning analytics literature. This review examines that proposal critically. It argues that the field has adopted attribution methods to satisfy a legitimate demand for transparency, but has largely not asked whether the explanations produced are of the kind instructors actually need. Three problems are developed. Disengagement is treated as equivalent to its behavioural proxies, so models explain platform inactivity rather than the underlying construct, and blended delivery makes this substitution especially unsafe because much learning occurs off-platform. Attribution explains the model rather than the student, and the features that carry the most attribution are frequently not the features an instructor can act on. And the evaluation literature measures predictive accuracy while the outcome that matters — whether flagged students are helped — is rarely measured at all. The paper proposes reorientation toward actionable and contestable explanation, evaluation against intervention outcomes, and explicit treatment of the equity and self-fulfilling-prophecy risks that early warning systems carry.

***Keywords*— Learning Analytics; Explainable Artificial Intelligence; Student Disengagement; Early Warning Systems; Blended Learning; Educational Data Mining**

I. Introduction

Blended instruction, combining scheduled contact with substantial online activity, is now the default configuration in much of higher education. It produces a durable by-product: a record of when students access materials, how long they spend, what they submit and when, and how they participate in forums and formative assessment. Learning analytics has developed to make use of that record, with early identification of students at risk among its most-pursued applications.

The prototype is Course Signals at Purdue, which combined grades, prior academic history and learning management system interaction to generate a traffic-light indication for instructors and students (Arnold & Pistilli, 2012). Comparable systems followed. Macfadyen and Dawson (2010) demonstrated the proof of concept for mining LMS data as an early warning instrument. Jayaprakash et al. (2014) reported an open-source early alert initiative and, importantly, examined what happened when alerts were acted upon. Subsequent work has refined the prediction problem substantially, including the question of how early in a course useful prediction becomes possible (Adnan et al., 2021) and how early-warning systems can be constructed from routinely available data (Akcapınar et al., 2019; You, 2016).

Accuracy improved. Adoption did not follow at the same rate, and the reasons are instructive. A model that outputs a risk score without justification asks an instructor to act on authority alone. Instructors respond to that request as one would expect: they discount it, particularly after encountering false alarms, and trust once lost is not readily recovered.

Explainable artificial intelligence has been offered as the resolution. Local attribution methods — principally Shapley additive explanations (Lundberg & Lee, 2017) and local surrogate explanation (Ribeiro et al., 2016) — decompose an individual prediction into contributions from input features, and are now routinely applied to dropout and disengagement models. The literature treats this as substantially settling the transparency problem.

This paper argues that it does not. The explanations produced are explanations of model behaviour, delivered in the vocabulary of the feature set, and the gap between that and what an instructor needs in order to act well is wider than the field acknowledges. Sections 3 to 6 develop the argument across the construct, the explanation, the timing of prediction and the evaluation; Sections 7 and 8 address risks and a research agenda.

II. Scope and Method of Review

This is a critical narrative review. It covers work that predicts academic disengagement, at-risk status, dropout or non-completion from behavioural or institutional data, with emphasis on studies that incorporate interpretability. Both fully online and blended contexts are included, with blended treated as the primary case and massive open online course research used comparatively — its dropout base rates and its absence of a contact component make it a different problem, and results transfer only with care.

Methodological literature on attribution is drawn upon directly rather than through its educational applications, since several of the limitations at issue are properties of the methods themselves.

Studies were selected for their contribution to the argument rather than for exhaustive coverage. Where a claim about prevailing practice is made, it reflects the pattern across the

studies examined; the review does not attempt to quantify that pattern, and does not claim to.

III. What is Being Predicted

The first difficulty is conceptual. Student engagement is a multidimensional construct spanning behavioural participation, emotional investment and cognitive effort, and disengagement is its progressive withdrawal. The construct is well theorised in the education literature and has been synthesised specifically for higher education contexts.

Predictive models do not operate on that construct. They operate on its digital traces: login frequency, resource access counts, time-on-page, submission timing, forum posts. The substitution is necessary and it is rarely examined.

Three consequences follow, and blended delivery sharpens all of them.

Off-platform learning is invisible. A student who attends every class, reads printed materials and studies with peers may register minimal platform activity. In a fully online course this pattern is genuinely anomalous; in a blended course it may be ordinary. Models trained on platform traces will flag such students, and instructors will learn from experience that the flags are unreliable.

Activity is not effort. Time-on-page records a browser tab, not attention. Access counts record clicks, not comprehension. The relationship between these measures and cognitive engagement is loose and varies by student.

Proxies are gameable once visible. When students learn that activity is monitored, activity responds. Where dashboards are shown to students, the intervention alters the measurement, and models calibrated before disclosure degrade after it.

The consequence is that a model's explanation, however faithful to the model, describes the trace and not the student. An attribution stating that low weekend activity drove a risk prediction is a true statement about the model and an unreliable statement about the learner.

IV. What the Explanations Explain

Attribution methods answer a specific question: how did each input feature contribute to this prediction, relative to a baseline. This is genuinely useful, and the empirical applications bear that out. Studies of disengagement from voluntary assessment activity, for instance, find that attribution values behave coherently — higher platform interaction shifts predictions toward submission, and the direction of effect is stable across students. Comparable analyses in dropout prediction show behavioural completion features carrying more attribution than instantaneous grade standing, which is a substantively interesting finding about what the model has learned.

Four limitations qualify the value of these outputs for instructional decision-making.

Model, not mechanism. Attribution is a decomposition of a fitted function. If the model has learned an association that reflects cohort composition rather than learning behaviour, the explanation will report that association confidently. The explanation cannot detect its own model's errors, and its fluency makes those errors harder to notice rather than easier.

Correlated features distribute attribution unstably. Behavioural traces are strongly intercorrelated: login count, session count and resource access move together. Attribution among correlated inputs depends on estimator choices, so importance rankings across such features should not be read as measurements of influence.

Attribution is not actionability. The features that carry the most attribution are frequently unalterable — prior grade point average, entry qualifications, demographic variables. An explanation dominated by these tells an instructor why the model is confident, not what to do. Explanations restricted to malleable features would be less faithful and more useful, and the tension between the two is real and largely unaddressed.

Correctness is unverified. In most educational applications the explanation is presented and found plausible. Plausibility is a weak test, since attribution methods will produce plausible-looking output from a badly specified model as readily as from a good one. Rudin (2019) argues that for high-stakes decisions the appropriate response is to use models that are interpretable by construction rather than to explain opaque ones after the fact, and the argument applies with force here: many of the features are few, tabular and well understood, which is precisely the regime in which a constrained interpretable model gives up little accuracy.

V. The Timing Trade-Off

A dimension largely absent from the explainability discussion, but central to whether these systems help anyone, is when the prediction is made.

Accuracy and earliness are in direct tension. Behavioural evidence accumulates over a course, so a model given ten weeks of data will outperform one given two. Intervention value runs the other way: a student flagged in week two can still change course, while a student flagged in week ten has already missed most of the material and may be past the point where support is effective. Adnan et al. (2021) address this explicitly by evaluating prediction at successive proportions of course length, and their framing — that the operationally relevant question is not maximum achievable accuracy but accuracy at the point where intervention remains worthwhile — deserves wider adoption than it has received.

Blended delivery complicates the trade-off further. Early-course platform activity in a blended module is a particularly noisy signal, because students are still establishing routines and because the proportion of learning occurring in contact sessions is highest at the start, when foundational material is delivered face to face. The period in which prediction would be most useful is therefore the period in which the available traces are least informative.

Two responses appear in the literature. One models the temporal sequence rather than aggregating it, on the argument that trajectory carries information that summary counts discard — a declining pattern of engagement means something different from a consistently low one, even where the totals match. Approaches of this kind, including sequence models and continuous-time formulations, report that variability in

engagement is itself predictive, with both very low and very high variability associated with elevated risk. The other response accepts lower early accuracy and adjusts the intervention accordingly, using low-cost, low-stigma contact early and reserving intensive support for later, more confident flags.

The second is probably the more practical, and it reframes the design problem usefully: rather than seeking a single optimal threshold, the system matches intervention intensity to prediction confidence. This is straightforward to implement and is rarely how deployed systems are configured.

VI. What is Evaluated

The literature evaluates classifiers. Accuracy, area under the curve, precision and recall are reported comprehensively, and the technical standard has risen.

The outcome of interest is different. An early warning system is worthwhile only if flagged students receive support they would not otherwise have received and are measurably better off as a result. That chain has three links — prediction, intervention, outcome — and the literature concentrates almost entirely on the first.

Jayaprakash et al. (2014) is notable for addressing the full chain, and dashboard studies have examined whether presentation to students changes engagement. But work of this kind remains a small minority, and there are structural reasons why it should be a larger one. Prediction quality and intervention effectiveness are not monotonically related: a highly accurate model attached to an intervention instructors do not trust produces nothing, while a moderately accurate model attached to a well-designed and well-timed intervention may produce a great deal.

Two further evaluation gaps deserve mention. Threshold selection is normally reported as a technical choice, when it is a policy choice about the acceptable ratio of missed students to false alarms, with different costs falling on different parties. And temporal validation is uncommon: cohorts differ year to year, delivery formats change, and a model validated by random partition within one cohort has not been shown to generalise to the next.

VII. Risks

Three risks recur and are inadequately handled.

Self-fulfilling prediction. A student labelled at risk may be treated differently by instructors and may internalise the label. Where that response reduces attainment, the model's accuracy is partly an artefact of its own deployment. This is not hypothetical and is difficult to detect without randomised allocation of the flag.

Differential error. Models trained on historical outcomes reproduce historical patterns. Where prior attainment correlates with socioeconomic background, first-generation status or disability, error rates will differ across groups even when overall accuracy is high. Reporting disaggregated performance is straightforward and uncommon.

Function creep. A system built to direct support can be repurposed for compliance monitoring or resource allocation. The technical artefact does not distinguish these uses; only governance does, and governance is generally outside the scope of the papers that build the systems.

Consent without alternative. Students agree to learning management system terms as a condition of enrolment, which makes consent to behavioural analysis formally present and substantively empty. Nothing in the technical literature addresses this, and it is not clear that anything in the technical literature could; but studies that describe institutional deployment without describing what students were told, and whether they could decline, omit a material fact about the system being evaluated.

Explainability is often invoked as mitigating these risks. It does not, on its own. An explanation makes a decision inspectable but does not make it contestable, and contestability — a defined route by which a student or instructor can challenge a flag and have it revised — is what the risks actually require. The distinction is practical. An institution that publishes its model's feature attributions has been transparent; an institution that lets a student say "I have been reading the printed notes, not the PDFs" and have that recorded against the prediction has built a corrective mechanism. Only the second constrains error, and only the second gives the student a reason to regard the system as operating on their behalf.

There is a further reason to prefer contestability over explanation alone. The traces on which these models operate are incomplete by construction, and the missing information is disproportionately held by the student. A student knows why their activity dropped in week six; the model has only that it dropped. A system designed to elicit that information converts an inference problem into a communication problem, which is both more tractable and closer to what an instructor would do unaided. Framed this way, the risk flag is best understood not as a judgement to be explained but as a prompt to ask a question, and much of the design difficulty the field has attributed to explainability dissolves once the output is understood in those terms.

VIII. Research Agenda

Construct validation. Behavioural proxies should be validated against independent measures of engagement rather than assumed adequate, with particular attention to blended settings where off-platform activity is substantial.

Actionable explanation. Explanation interfaces should be designed around what an instructor can change, and evaluated on whether instructors make better decisions with them — a question answerable by experiment and almost never asked.

Interpretable-by-design comparison. Studies applying post hoc attribution should report the accuracy cost of an inherently interpretable alternative. Where that cost is small, the case for the opaque model is weak (Rudin, 2019).

Intervention-anchored evaluation. Reporting should extend to what was done with the prediction and what happened to the student, with randomised or quasi-experimental allocation of alerts where feasible.

Disaggregated and temporal reporting. Performance by student subgroup, and across cohorts rather than within them, should be standard.

IX. Conclusion

Predicting academic disengagement from learning management system data is technically mature. The models work, in the narrow sense that they identify students who go on to fail or withdraw earlier than existing institutional processes do.

The move toward explainability has been a genuine improvement over unexplained risk scores, and the empirical attributions reported are largely coherent with what is known about learning behaviour. But the field has treated explanation as an endpoint when it is an intermediate step. The explanations produced describe how a model combined behavioural proxies; instructors need to know which student needs what, and whether the system's judgement can be trusted and challenged in this particular case.

Closing that gap requires work of a different kind from what the field currently rewards: validating proxies against the construct, designing explanations around actionable factors, comparing against interpretable baselines, and evaluating against student outcomes rather than classification metrics. Blended environments make this more pressing rather than less, because it is precisely there that platform traces capture the smallest share of what students actually do.

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, 9, 7519–7539. https://doi.org/10.1109/ACCESS.2021.3049446

Akcapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, 16(1), 1–20. https://doi.org/10.1186/s41239-019-0172-z

Arnold, K. E., & Pistilli, M. D. (2012). Course Signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 267–270). ACM. https://doi.org/10.1145/2330601.2330666

Feng, W., Tang, J., & Liu, T. X. (2019). Understanding dropouts in MOOCs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 517–524.

Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. In *Proceedings of the Second European MOOC Stakeholder Summit* (pp. 58–65).

He, J., Bailey, J., & Rubinstein, B. I. P. (2015). Identifying at-risk students in massive open online courses. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 1749–1755).

Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 12(3), 1353. https://doi.org/10.3390/app12031353

Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1), 6–47. https://doi.org/10.18608/jla.2014.11.3

Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (2017). Awareness is not enough: Pitfalls of learning analytics dashboards in the educational practice. In *Data Driven Approaches in Digital Education* (pp. 82–96). Springer.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30 (pp. 4765–4774).

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, 54(2), 588–599.

Mangaroska, K., & Giannakos, M. (2017). Learning analytics for learning design: Towards evidence-driven decisions to enhance learning. In *European Conference on Technology Enhanced Learning* (pp. 428–433). Springer.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education*, 29, 23–30.