

SHAP-Centric Interpretability in Educational Machine Learning: A PRISMA Systematic Review of SHAP and LIME

Abdulaziz Hameed Alharbi^{1*}, Midhunchakkaravarthy Janarthanan¹, Divya Midhunchakkaravarthy¹

¹ School of AI, Computing and Multimedia, Lincoln University College, Petaling Jaya, Selangor 47301, Malaysia

ABSTRACT

Machine learning now informs consequential decisions in education. It flags students at risk, forecasts attainment, recommends interventions. The models that predict best are the least transparent, and this opacity now threatens trust, accountability, and the pedagogical legitimacy of these systems. SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are the post-hoc techniques most often proposed as the remedy. How the two are actually used in educational research is far less clear. This review, reported in accordance with the PRISMA 2020 statement, covers 75 peer-reviewed studies published between 2022 and 2026. Five databases were searched (Scopus, SpringerLink, ScienceDirect, ERIC, and IEEE Xplore) and returned 781 records; 669 remained after de-duplication, 135 met the eligibility criteria, and 75 full texts were retrieved, coded, and analysed by model, dataset, task, and explanation method. The picture that comes out of the coding is one-sided. SHAP appears in 74 of the 75 studies. LIME appears in 33, and in 32 of those 33 it accompanies SHAP rather than replacing it. Tree-based models predominate (72 of 75), and the literature concentrates on performance and dropout prediction in higher education. Interpretability is judged almost entirely from the model side; whether educators or students understand the explanations, or change anything because of them, is rarely tested. The review describes this state of practice as SHAP-centric, technically framed interpretability, and uses the gaps it leaves, in human-centred evaluation, longitudinal explanation and fairness, to set out a research agenda for trustworthy educational XAI.

ARTICLE HISTORY

Received April 17, 2026

Accepted June 29, 2026

KEYWORDS

Explainable Artificial
Intelligence,
SHAP,
LIME,
Learning Analytics,
Dropout Prediction,
Systematic Review

1. Introduction

Education has entered an era in which data, rather than intuition alone, increasingly mediates pedagogical and administrative judgement. Machine learning (ML) now underpins adaptive learning platforms, early-warning systems, automated assessment, and intelligent tutoring, and high-capacity models such as gradient-boosted trees and deep neural networks routinely outperform classical baselines on tasks ranging from dropout detection to attainment forecasting (Asselman et al., 2023; Kalita et al., 2025)(Study A31). The very property that confers this accuracy, the capacity to model complex, non-linear interactions, also renders such models opaque: they yield predictions without disclosing the reasoning behind them (Arrieta et al., 2020; Minh et al., 2022). In a domain where an algorithmic output may determine whether a student is flagged as failing, granted admission, or selected for intervention, opacity is not a mere technical inconvenience but a threat to trust, validation, and institutional accountability (Angelov et al., 2021; Guidotti et al., 2018).

Explainable Artificial Intelligence (XAI) has become an important way to improve the transparency of machine learning models. Post-hoc and model-agnostic XAI methods, such as SHAP and LIME, are predominantly used for prediction interpretation. SHapley Additive exPlanations (SHAP) are grounded in cooperative game theory, where Shapley values are used to estimate the contribution of each input feature to a specific prediction. As a result, SHAP can support both local explanations for individual cases and broader global insights into model behaviour (Lundberg & Lee, 2017). Local Interpretable Model-agnostic Explanations (LIME) instead approximate a complex model in the neighbourhood of a single instance with a sparse, interpretable surrogate, yielding case-level rationales applicable to any classifier (Ribeiro et al., 2016). Both promise to convert numerical predictions into feature-level narratives that non-specialists can scrutinise.

Education, however, has requirements that differ from many of the domains where these methods were first developed. Because the main users are often teachers, administrators, and learners rather than data scientists, explanations

need to be easy to understand, relevant to educational practice, and ethically responsible, not only mathematically accurate. Previous studies have used SHAP to rank risk factors in dropout and performance prediction (Dass et al., 2021) while LIME has been used to explain textual feedback and sentiment (Shukla et al., 2026). More recent work has also started to combine both methods in educational contexts (Study A45) (Mustofa et al., 2025; Quimiz-Moreira et al., 2025; Villar & de Andrade, 2024). Yet this literature remains fragmented and case-bound. No synthesis has established, on the basis of the full retrievable evidence, which methods are actually used, for which tasks and models, and whether the resulting explanations are evaluated with the stakeholders they are meant to serve.

1.1 Objectives and Contribution

This review addresses that gap. The current literature review involves analysis of 75 peer-reviewed papers published in the period from 2022 to 2026 in adherence with the PRISMA 2020 criteria. Specifically, the paper focuses on the use of SHAP and LIME for interpretability in education with respect to the educational task, model used, and dataset applied in such studies. The results show that SHAP is used more widely than LIME and that SHAP is mainly used for ordering risk factors and explaining model output while LIME is mostly used in the secondary context. Such an unbalanced pattern deserves special mention since both SHAP and LIME are considered to be equally applicable approaches; however, their use does not match in terms of prevalence. Among the other problems that deserve consideration are the applicability of the explanations for teachers and learners, as well as the stability of such explanations and their relation to fairness.

The review is guided by three research questions:

RQ1. How are SHAP and LIME applied to interpret educational ML models, and with which model families?

RQ2. What are the datasets, education-related activities, and learners involved in such studies, and how are they varied according to different explainability methods?

RQ3. What are the methodological, evaluation-related, and pedagogical limitations of the existing empirical evidence?

The remainder of the paper is organised as follows. Section 2 sets out the theoretical background, including feature attribution and human-centred interpretability. Section 3 presents the PRISMA 2020 methodology. Section 4 reports the synthesis. Section 5 discusses the findings and their implications, Section 6 sets out future research directions, and Section 7 concludes.

2. Theoretical Background

Interpretability research distinguishes models that are intrinsically transparent, such as linear models and shallow trees, from high-capacity models whose behaviour must be reconstructed after training through post-hoc techniques. Educational ML overwhelmingly follows the latter route, because the architectures that achieve state-of-the-art accuracy, ensembles and deep networks, are not transparent by construction. Three intersecting bodies of theory frame the present review: feature attribution, human-centred interpretability, and the sociotechnical conception of trust.

2.1 Feature attribution: Shapley Values and Local Surrogates

SHAP operationalises the Shapley value from cooperative game theory, treating each feature as a player whose marginal contribution to a prediction is averaged over all feature coalitions; the resulting attributions satisfy the desirable properties of local accuracy, missingness, and consistency (Lundberg & Lee, 2017; Salih et al., 2025). Tree-specific estimators render this computation tractable for gradient-boosted models, which partly explains SHAP's affinity with the tabular data that dominate educational prediction (Lundberg et al., 2020). LIME proceeds differently, perturbing an instance, observing the model's responses, and fitting an interpretable surrogate within the local neighbourhood; it is model-agnostic and data-type-agnostic but, by construction, sensitive to the choice of perturbation and sampling (Ribeiro et al., 2016). The two methods thus embody a global-versus-local complementarity rather than a strict opposition, a distinction that the present synthesis shows to be decisive in practice.

2.2 Human-centred and Cognitive Interpretability

A recurrent theme in the wider XAI literature is that algorithmic interpretability, the property that a model can be explained, is not equivalent to cognitive interpretability, the property that an explanation is understood and usable by a particular person (Rong et al., 2023). The value of an explanation is contingent on the recipient's expertise, the decision at hand, and the cognitive load that the representation imposes. In the realm of education, where the end-user is often a layperson rather than technically knowledgeable, the difference is important because what may be mathematically correct may still not be pedagogically effective. The theory supporting these recommendations, informed by human-computer interaction and instructional design, is discussed in Section 5.

2.3 Trust, Acceptance, and Accountability

Whether a stakeholder acts upon an explanation is governed by trust and acceptance, constructs examined through technology-acceptance and trust-calibration frameworks. Evidence indicates that trust is highest when explanations cohere with domain intuition and are well calibrated to model uncertainty, and that domain-specific framings raise educators' acceptance more than generic visualizations (Feldman-Maggor et al., 2025). Accountability provides a normative element in that, within critical educational environments, an explanation becomes a means of making the process contestable, allowing for the justification or rejection of an algorithm-based decision. With increased regulation in AI applications within the field of education, explainability has become less a preference and more a necessity.

3. Materials and Methods

The review is conducted and reported in line with the PRISMA 2020 guidelines (Page et al., 2021; Stracke et al., 2023) for systematic reviews and meta-analyses. The PRISMA statement outlines the identification, screening, eligibility, and selection stages of the literature search to ensure transparency and validity in the review process.

3.1 Protocol and Registration

A review protocol specifying the research questions, eligibility criteria, information sources, and screening procedure was defined a priori. The review was not prospectively registered in PROSPERO; this is acknowledged among the limitations in Section 5.7. To support replication, the protocol, the database-specific search exports, and the complete coded dataset are openly available in a public repository on Zenodo (<https://doi.org/10.5281/zenodo.21620430>).

3.2 Eligibility Criteria

Eligibility was specified through a PICOS-style frame adapted for a methodological review (Table 1). From the reviewed literature, the eligible literature consisted of articles published by peer-reviewed journals or conference papers in English that were published between the years of 2022 and 2026, which have used SHAP, LIME, or both, to interpret at least one machine learning (ML) model in an authentic educational environment. The period 2022 was chosen to capture the rapid rise in use of the two algorithms (SHAP and LIME) after their mainstreaming and in light of the significant growth of educational analytics due to the pandemic. Articles were also not included if they were based only on simulated data and not validated with actual learners, if they were some other type of data such as systems outside of education, if they were not published in a peer-review journal or a grey literature source.

Table 1: Eligibility Criteria (PICOS Framework)

Element	Inclusion	Exclusion
Population / Context	Authentic educational settings (higher education, K-12, e-learning, MOOCs, learning analytics, intelligent tutoring)	Non-educational domains, unless insights are explicitly transferable to education
Intervention	Empirical application of SHAP, LIME, or both, with implementation detail	Mention of XAI without empirical implementation
Outcomes	Interpretability, feature attribution, stakeholder trust, fairness, actionability	Predictive performance only, with no explainability analysis
Study Design	Peer-reviewed empirical studies using at least one ML model	Grey literature, preprints, theses; non-English
Time	2022-2026	Outside the window

3.3 Information Sources and Search Strategy

In total, five databases were searched on 1 June 2026, comprising Scopus (n=413), SpringerLink (n=223), ScienceDirect (n=94), ERIC (n=31), and IEEE Xplore (n=20), for a sum of 781 records. Search queries included three concepts combined with Boolean operators, namely explanation technique (“SHAP” OR “LIME” OR “explainable AI” OR “interpretable AI”), model type (“machine learning” OR “deep learning”), and field of research (“education” OR “student” OR “learning analytics”). Full query statements, inclusive of filters and date restrictions, are available on Zenodo (<https://doi.org/10.5281/zenodo.21620430>).

3.4 Study Selection

Selection conformed to the PRISMA 2020 guidelines (Fig. 1). First, de-duplication resulted in the exclusion of 112 duplicate records, resulting in 669 unique records to be screened by title and abstract. In total, 534 articles were found off-

topic or irrelevant to education and/or did not contain an explicit reference to SHAP, LIME, or explainable AI techniques. Thus, a total of 135 records were included in the final stage of selection. Of these, 75 records had been accessed and reviewed based on the complete text, while 60 records were not available during the review window, thus contributing to selection bias, as discussed in Section 5.7. The synthesis is therefore based on the 75 retrieved studies, which are listed in full in Appendix A (Studies A1–A75). Throughout the manuscript, individual included studies are cited in the text using these labels. Table 2 summarises the number of records retained and excluded at each stage.

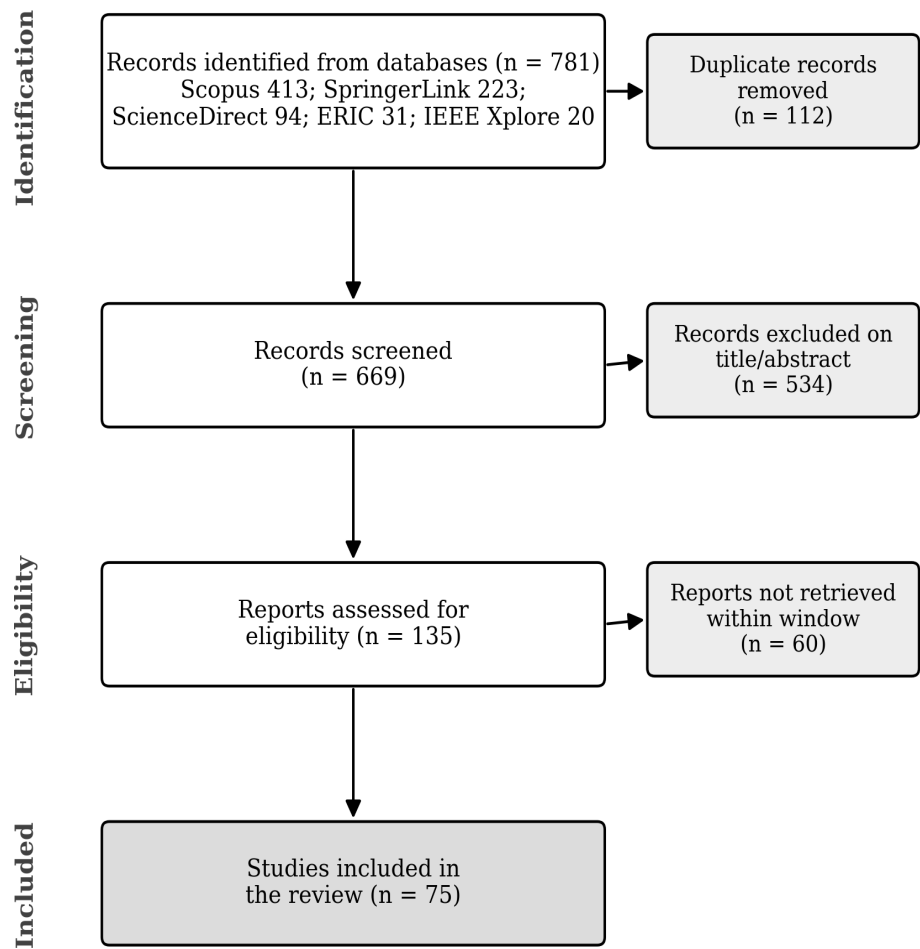


Figure 1: PRISMA 2020 Flow Diagram of Identification, Screening, and Inclusion

Table 2: Records at Each Stage of Selection

Stage	n	Description
Identified Through Databases	781	Scopus 413, SpringerLink 223, ScienceDirect 94, ERIC 31, IEEE Xplore 20
After Duplicate Removal	669	112 duplicates removed
Excluded at Title/Abstract	534	Off-topic, non-education, or no SHAP/LIME/XAI
Eligible at Title/Abstract	135	Met the criteria; full texts sought
Not Retrieved at Full Text	60	Full text unavailable within the review window
Included in Synthesis	75	Retrieved, coded, and analyzed

3.5 Data Extraction and Coding

Data extraction and coding were carried out by the first author, who read the full text of every included study and recorded the relevant information in a structured workbook capturing bibliographic metadata, the ML model family, the XAI method and its interpretability level, the dataset and educational level, the task addressed, the reported performance, the principal findings, and the author-stated limitations. To promote consistency and reduce the chance of oversight across a corpus of this size, a rule-based text-mining script was used as a supporting first-pass aid that flagged candidate mentions of SHAP, LIME, model families, and task keywords in each document. These automatically generated suggestions were not treated as final: the first author manually reviewed and verified every categorical assignment against the full text, correcting

the script's output where necessary, and a method was recorded as used only where the author judged from a close reading that it had been operationalised rather than merely cited or mentioned in passing. The coding scheme, the supporting script and its decision rules, and the complete verified dataset are released so that every assignment can be independently inspected and reproduced. Because extraction was conducted by a single reviewer rather than duplicated by a second independent coder, no formal inter-rater reliability statistic is reported; this is acknowledged in Section 5.7. Where a study addressed more than one task or population, it was assigned to its primary category to avoid double counting. Table 3 illustrates the extracted record for a representative included study.

Table 3: Extracted Record for a Representative Included Study

Field	Extracted Value
Study	Ujkani et al. (2024) [A65]
Venue / DOI	Electronics; 10.3390/electronics13214157
ML Model	XGBoost with custom deep-learning models (LSTM, CNN)
XAI Method	SHAP (global and local)
Dataset / Level	OULAD; tabular; higher education (online)
Task	Course-success and at-risk prediction
Reported Performance	Up to 94% classification accuracy
Principal Finding	SHAP attributions identified the dominant determinants of pass/fail outcomes
Author-stated Limitation	Reliance on a single public dataset

3.6 Quality Appraisal

Methodological quality was appraised by the first author using a modified Critical Appraisal Skills Programme (CASP) instrument adapted for empirical ML and XAI research. The 75 studies being appraised were each read in their entirety and assessed according to criteria defined by the researcher; this was achieved through the use of a 5-point scale, where 1 (very poor) represented the lowest score and 5 (very good) represented the highest. The factors being assessed included: the level of clarity associated with objectives; the extent to which the dataset was described in a clear and transparent manner; the reproducibility of the analysis model pipeline; the justification for, and appropriateness of, XAI used; and the extent to which limitations of the study have been acknowledged. To ensure that consistency was maintained throughout the dataset, an established scoring rubric was utilised to achieve the above. The text-mining resources described in Section 3.5 were used only as a means to assist in identifying these relevant sections in the original study documentation, whereas the actual final evaluations were made based solely on the author's own judgment and evaluation rather than through the use of automated means. An overview of the rubric used and the individual study scores are included with the dataset.

Most of the studies assigned good appraisal scores overall. The mean appraisal scores of the five dimensions ranged from 4.40 to 4.91, with a mean dimension score of 4.65, and the standard deviation of the dimension scores across dimensions was 0.22. All studies reviewed in the process of creating this paper were above the minimum score of 3.0 to be retained; see Table 4 for specific scores for all included studies. The strongest dimensions were dataset description and limitations of the study(s). Reproducibility received the lowest appraisal scores and illustrates a lack of available code, implementation information, and hyperparameter data within the literature reviewed.

The high appraisal scores should be interpreted cautiously. The studies that are included in the review were identified from existing databases and were accessible for this review. Thus, the appraisal serves as an indicator of how well the literature on this topic reports results rather than an indicator of the overall quality of the broader research field. Because the studies included in this review demonstrate heterogeneity and do not have sufficient comparability to use the appraisal to generate a weighted pooled estimate, the appraisal was used to determine the level of confidence in the synthesis of this literature and to identify that reproducibility is one of the major weaknesses in the literature and needs to be addressed in future research.

Table 4: Quality-appraisal Summary by Dimension (n = 75)

Dimension	Mean (SD)	Observation
Clarity of Objectives	4.55 (0.68)	Most studies stated explicit aims or research questions
Dataset Transparency	4.91 (0.29)	Datasets, sizes, and features were generally well documented
Reproducibility	4.40 (0.65)	Weakest dimension; code and hyperparameters unevenly reported
XAI Justification	4.51 (0.53)	SHAP was justified more rigorously than LIME
Limitations Acknowledged	4.89 (0.35)	Limitations and future work were routinely discussed

3.7 Synthesis and Certainty of Evidence

Considering the diversity among the different models, data sets, tasks, and outcomes measured, conducting a quantitative meta-analysis was impractical. Information about these aspects was summarized descriptively using frequency tables, cross tabulation, and themes. The strength of the thematic statements was determined through informal methods based on the number and consistency of supporting studies, which is reflected in the discussion of these themes. No GRADE or CERQUAL assessment was made, following the spirit of the methodology that guided this systematic review and in line with conventions in educational research and computational science.

3.8 Ethical Considerations

In spite of the fact that there was no primary data collection or involvement of human participants in this systematic review, ethical concerns related to machine learning in education have been considered, including fairness, learner-data privacy, transparency in high-stakes decision-making, and accountability for actions taken based on algorithmic outputs. While explicit analysis of fairness and potential bias was found in relatively few papers (6 out of 75 reviewed; 8%), proposals on how to mitigate these risks were even rarer. Many of these issues are documented quantitatively, as laid out in Section 4 of the paper, thereby revealing a disparity between ethical aspirations and demonstrated practices, which is addressed further in Sections 5 and 6.

4. Results

Each of the 75 included studies was coded for its explanation method, model family, task, and learner population, and the frequencies reported in this section are drawn directly from that coded dataset (Appendix A, Studies A1–A75). Two regularities organize the findings that follow. First, the literature is overwhelmingly concerned with structured, tabular prediction by tree-based models, the setting in which SHAP is computationally cheapest and theoretically best grounded. Second, interpretability is appraised almost exclusively from the model side, with the human reception of explanations left largely unexamined. Table 5 presents a representative cross-section of the corpus.

Table 5: Representative Included Studies by Model, Explanation Method, and Task

First Author (Year)	ML Model(S)	XAI Method	Education Level	Task	Study ID
Johora (2025)	XGBoost; LightGBM	SHAP + LIME	Higher ed	Performance prediction	A29
Abukader (2025)	LightGBM	SHAP	Higher ed	Performance prediction	A2
Mustofa (2025)	XGBoost; LightGBM	SHAP + LIME	K-12 / school	Dropout / retention	A45
de la Hoz (2026)	XGBoost; Random Forest	SHAP + LIME	Higher ed	At-risk identification	A12
Qu (2022)	Random Forest; Decision Tree	SHAP	MOOC / online	Engagement / behaviour	A51
Kim (2026)	SVM; Logistic Regression	SHAP	Higher ed	Sentiment / feedback	A38
Kesgin (2025)	XGBoost; Random Forest	SHAP + LIME	K-12 / school	Fairness / bias	A35
Kar (2024)	Random Forest; Decision Tree	SHAP + LIME	MOOC / online	Assessment / grading	A33
Li (2025)	LSTM; CNN	LIME	Higher ed	Recommendation	A40

4.1 SHAP as the de Facto Standard for Tabular Prediction

SHAP is, by a wide margin, the dominant explanation method, appearing in 74 of the 75 studies (Fig. 3). Its prevalence is inseparable from the modelling choices of the field: 72 of the 75 studies employ tree-based learners, XGBoost, random forest, gradient boosting, LightGBM, or CatBoost, whose native compatibility with tree-SHAP makes both global feature rankings and local attributions computationally inexpensive. SHAP accordingly anchors the two most common tasks, performance and grade prediction (21 studies) and dropout or retention (11), where it furnishes dataset-wide importance rankings alongside instance-level rationales. Reported attributions converge on a familiar set of determinants, prior attainment, engagement and attendance signals, and assessment activity. The near-universal adoption of SHAP indicates that it has become not one option among several but the default grammar through which educational ML on tabular data is now interpreted.

4.2 LIME is Rarely used Alone and Predominantly Augments SHAP

LIME presents a markedly different profile. It appears in 33 of the 75 studies but, in a finding that qualifies the conventional portrayal of SHAP and LIME as competing alternatives, is almost never deployed in isolation: a single study uses LIME without SHAP, a course-recommendation model (A40), while the remaining 32 pair it with SHAP. Where it appears, LIME typically supplies instance-level explanations atop a primarily SHAP-based analysis, and several authors

justify the pairing precisely by reference to LIME's sensitivity to perturbation, electing to corroborate rather than to rely upon it. The practical landscape is therefore better described as SHAP augmented, where warranted, by LIME than as a contest between two methods. Fig. 2 reports the distribution of learner populations, which is dominated by higher education (50 studies), followed by K-12 or school settings (16) and MOOC or online contexts (7).

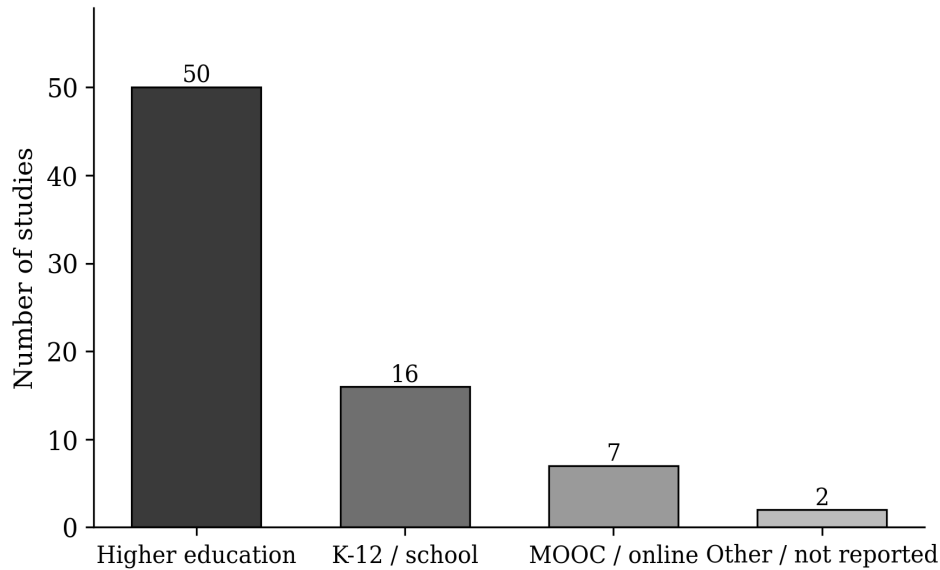


Figure 2: Distribution of the 75 included Studies by Educational Level

4.3 The Complementary Configuration of SHAP and LIME

The 32 studies that combine the two methods reveal a consistent division of labour: SHAP provides the global, dataset-wide ranking of determinants, while LIME explains individual or atypical predictions, so that an educator may move between cohort-level insight and case-level diagnosis. Because LIME appears essentially only in this supporting capacity, the configuration of methods across the corpus resolves into three mutually exclusive groups, SHAP alone (42 studies), SHAP with LIME (32), and LIME alone (1), shown alongside the overall SHAP and LIME counts in Fig. 3. This distribution is the clearest single expression of the review's central finding: interpretability in educational ML is not a balanced negotiation between two techniques but a SHAP-centred practice in which LIME plays an occasional, complementary role.

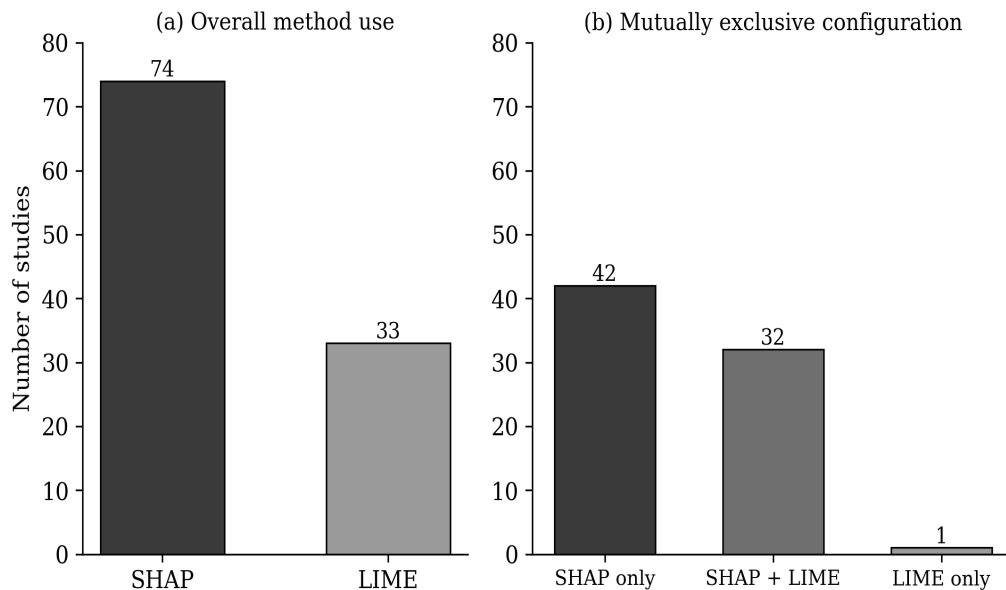


Figure 3: Explanation-method Configuration across the 75 included Studies: Overall Presence of SHAP and LIME (Left) and the Mutually Exclusive SHAP-alone, SHAP-with-LIME, and LIME-alone Partition (Right)

4.4 Distribution Across Educational Tasks

Table 6 cross-tabulates the corpus by task and method configuration. Performance and grade prediction (21 studies) and dropout or retention (11) together account for nearly half of the literature, followed by sentiment or feedback analysis and at-risk identification (9 each), engagement or behaviour modelling (8), and fairness or bias auditing (6); the remainder comprise assessment or grading (4), recommendation (2), and other tasks (5), so that the nine categories sum to the full corpus of 75. SHAP is present across every task, whereas the combined SHAP-and-LIME configuration is proportionally most frequent in at-risk identification, dropout, and fairness work, precisely the settings in which the explanation of an individual learner's predicted trajectory carries the greatest consequence. The task distribution is depicted in Fig. 4.

Table 6: Distribution of Studies by Educational Task and Explanation Configuration

Educational Task	SHAP Only	SHAP + LIME	LIME Only	Total	Study IDs (Appendix A)
Performance / Grade Prediction	15	6	0	21	SHAP: A2, A7, A10, A17, A26, A43, A44, A46, A49, A57, A59, A62, A69, A70, A72 SHAP + LIME: A15, A20, A29, A32, A34, A56
Dropout / Retention	5	6	0	11	SHAP: A9, A21, A37, A39, A66 SHAP + LIME: A5, A8, A45, A47, A50, A61
Sentiment / Feedback	5	4	0	9	SHAP: A6, A13, A28, A38, A60 SHAP + LIME: A18, A52, A55, A67
At-Risk Identification	3	6	0	9	SHAP: A1, A4, A58 SHAP + LIME: A12, A30, A31, A53, A65, A71
Engagement / Behaviour	4	4	0	8	SHAP: A19, A36, A51, A75 SHAP + LIME: A16, A23, A24, A48
Fairness / Bias	2	4	0	6	SHAP: A25, A64 SHAP + LIME: A22, A27, A35, A54
Assessment / Grading	3	1	0	4	SHAP: A14, A73, A74 SHAP + LIME: A33
Recommendation	1	0	1	2	SHAP: A63 LIME: A40
Other	4	1	0	5	SHAP: A11, A41, A42, A68 SHAP + LIME: A3

Note: Categories are mutually exclusive and sum to 75. LIME was used alone in a single study; in every other instance it accompanied SHAP.

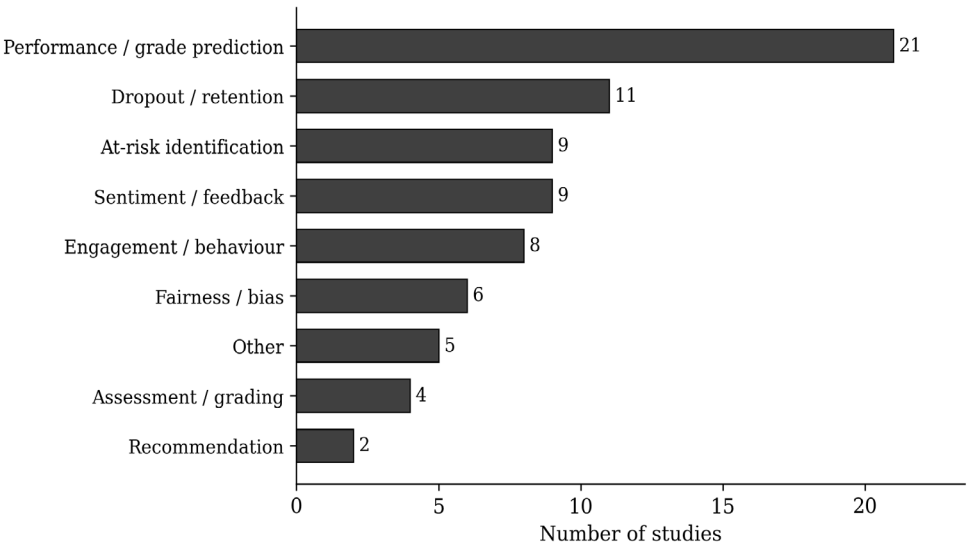


Figure 4: Distribution of the 75 included Studies by Educational Task

4.5 Corpus Composition: Recency, Models, and Datasets

The corpus is recent. Only ten of the 75 studies predate 2024 (A5, A10, A19, A26, A34, A44, A51, A60, A71, A74), while the remaining 65 appeared between 2024 and 2026, in the final three years of the search window; Fig. 5 shows the distribution by publication year. This concentration indicates that the literature analysed here has formed very rapidly, almost entirely since SHAP and LIME entered mainstream use.

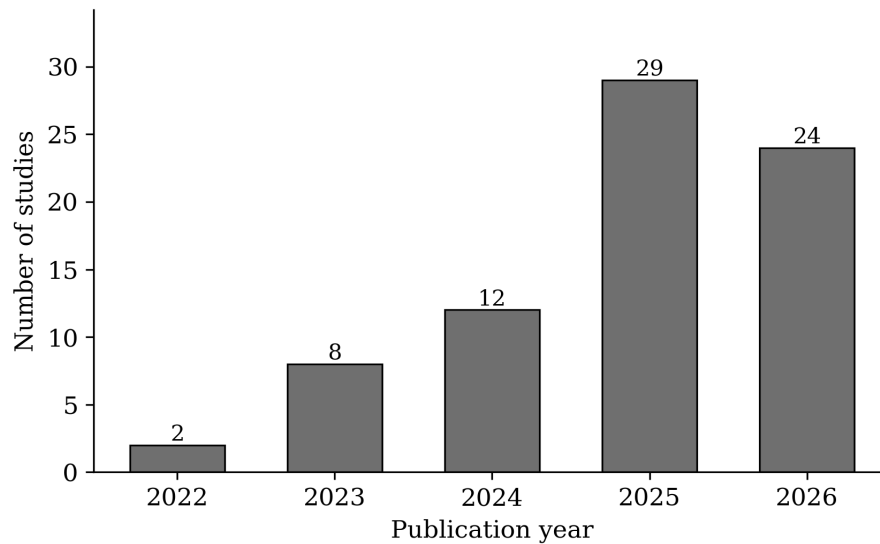


Figure 5: Distribution of the 75 included Studies by Publication Year

Tree-based methods dominate. At least one such model appears in 72 of the 75 studies, with random forest the most common (63 studies), followed by gradient boosting (55), decision tree (54), XGBoost (53), LightGBM (27) and CatBoost (17); the overall pattern of model-family use is summarised in Fig. 6. Deep-learning components turn up in 62 studies, mostly LSTM, CNN, transformer and multilayer-perceptron architectures. In almost every case, though, they are trained alongside the tree ensembles rather than in place of them. This preference is mutually reinforcing with the dominance of SHAP, since the efficiency of tree-SHAP rewards exactly these models. The empirical base, however, rests on a narrow set of public benchmarks: The Open University Learning Analytics Dataset and Kaggle education sets each underpin 11 studies and PISA a further 10. Such reliance on clean, curated corpora, while convenient for reproducibility, attenuates the noise and heterogeneity of authentic institutional data and is revisited among the limitations in Section 5.7.

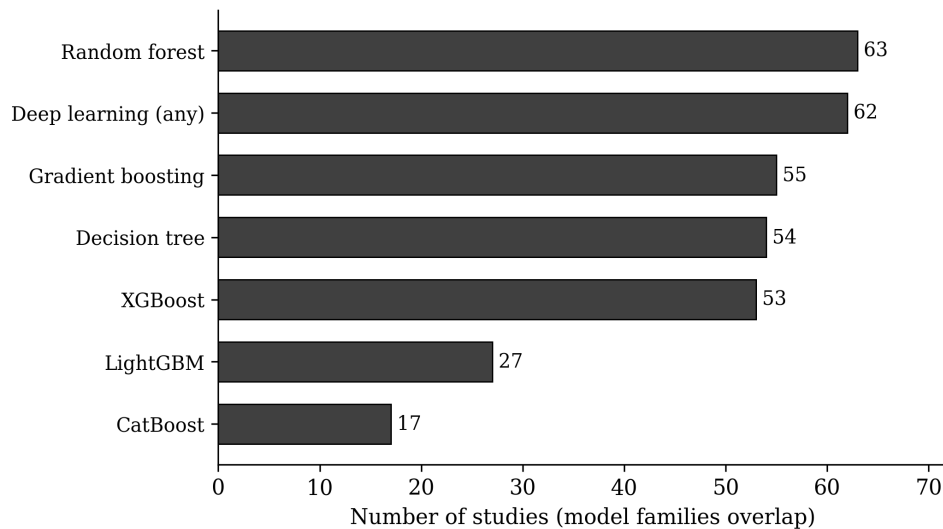


Figure 6: Model Families used Across the 75 included Studies (Counts Overlap, as many Studies use Several Models)

4.6 The Evaluation of Interpretability: A Model-side Bias

A defining characteristic of the corpus is the locus of evaluation. Most studies judge their explanations on technical grounds, through feature attribution, fidelity or stability checks. The person who has to read the explanation and act on it rarely enters the assessment at all. Teachers, students or administrators are involved in only a minority of studies, and even fewer report any measure of trust, usability or impact on decisions. Where a stakeholder framing is present, authors contend that SHAP's global rankings suit cohort-level dashboards while local explanations suit individual decisions, but direct evidence on educator comprehension and behaviour remains sparse. This asymmetry, interpretability demonstrated

technically yet seldom validated with its intended audience, is among the most consequential findings of the review and frames the human-centred argument of Section 5.

4.7 Methodological Gaps Observed Across the Corpus

Five shortcomings recur across the included studies, and each one can be quantified. First, the evidence base is methodologically narrow. Most of the corpus rests on tree ensembles trained on a small set of tabular benchmarks, while multimodal data, video, audio and behavioural signals other than clickstreams remain almost untouched. Second, fairness receives little attention. As noted earlier, only 6 of the 75 studies carry out an explicit bias or fairness analysis, and concrete mitigation steps are rarer than that. Third, evaluation sits mainly on the model side, and how stakeholders receive the explanations is examined far less often (Section 4.6). Fourth, the literature is cross-sectional and static: explanations are presented as snapshots, and almost no study tracks how a learner's attributions evolve over time, restricting their value for progress monitoring. Fifth, the geographic and contextual concentration of the corpus in well-resourced settings leaves the behaviour of these methods in low-resource educational systems essentially unknown. Together these gaps delineate the agenda that Sections 5 and 6 pursue: a move from technically successful attribution toward adaptive, fair, and educationally validated explanation at scale.

5. Discussion

Read together, the 75 studies depict a field that has resolved the question of which explanation method to use, and has done so in favour of SHAP, while leaving the harder questions of for whom, to what effect, and at what cost largely open. The discussion that follows interprets this settlement along four axes, the technical trade-offs between the two methods, the cognitive distance between explanation and understanding, the ethical obligations specific to education, and the institutional conditions for adoption, before consolidating the implications into a design framework.

5.1 Technical Trade-offs and the Logic of SHAP-centrality

The dominance of SHAP is not arbitrary but follows from a coherent technical logic. Grounded in cooperative game theory, SHAP yields consistent, globally coherent attributions that are inexpensive to compute for the tree ensembles that dominate tabular educational prediction (Lundberg et al., 2020; Salih et al., 2025). LIME, by contrast, offers flexible, model-agnostic local explanations but is destabilised by its reliance on perturbation and surrogate approximation, so that closely related instances may receive divergent rationales (Visani et al., 2022). The corpus shows that practitioners have internalised precisely this trade-off: they adopt SHAP as the primary instrument and enlist LIME, where at all, to corroborate individual cases rather than to carry the explanatory burden. The consequence is a de facto standardisation that brings comparability and efficiency but also concentrates the field's epistemic risk in a single method whose own assumptions are seldom interrogated.

5.2 From Attribution to Understanding

A model can be interpretable in the technical sense without its explanations meaning anything to the people who receive them. The corpus shows a field that has worked hard on the technical side and paid far less attention to the human one. Attributions are produced and checked against fidelity or stability criteria as a matter of routine; whether teachers and learners can make sense of them is hardly ever tested. The gap is conceptual as much as empirical. Knowing that a feature contributed a certain amount to a prediction does not, by itself, tell an educator what to do next. Closing it requires that interpretability be reconceived as a problem of communication and design rather than of computation alone, an argument that the human-centred XAI literature has advanced and that the present evidence renders urgent for education (Rong et al., 2023).

5.3 Fairness and the Ethics of Educational Explanation

Explanation in education carries an ethical charge absent from many other domains, because its outputs bear upon the life chances of minors and young adults. It is therefore striking that fairness analysis is so rare across the corpus, and that the diagnosis of bias, where it occurs, is seldom matched by mitigation. Explaining a model is not the same as justifying its use; an attribution may render a decision transparent while leaving its fairness unexamined. The review thus argues that educational XAI must be held to a higher standard than transparency alone, coupling attribution with fairness auditing and with the capacity for stakeholders to contest or override algorithmic logic (Aler Tubella et al., 2024; Deho et al., 2022; Zhu et al., 2025).

5.4 Toward a Framework for Trustworthy Educational XAI

If the discussion so far makes one thing clear, it is that no single technique is sufficient. Even SHAP and LIME used

together cannot meet every demand that educational decision-making places on an explanation, since the stakes are high and the requirements pull in several directions at once. A broader, human-centred framework is needed. The review therefore sets out seven principles for developers, educators and policymakers:

Model-task fit. Match the explainer to the task and architecture, SHAP for structured prediction, LIME for instance-level corroboration, and intrinsic models where stakes are highest.

Cognitive translation. Render attributions in layered forms, technical detail beneath plain-language summaries, so that non-specialists can act upon them.

Pedagogical contextualisation. Tie feature attributions to learning constructs and concrete interventions rather than to raw variables.

Interactive interrogation. Allow educators to explore counterfactual and what-if scenarios rather than receive static outputs.

Fairness integration. Embed bias visualisation, counterfactual fairness, and uncertainty calibration as first-class components, not afterthoughts.

Explanation literacy. Develop stakeholders' capacity to read and critically appraise explanations through professional development.

Evaluation with stakeholders. Validate explanations against educator and learner comprehension and decision impact, not fidelity metrics alone.

5.5 Implications for Practice and Policy

The findings carry a practical warning for institutions. A system is not explainable just because it produces SHAP outputs, and ticking that box at procurement should not settle the question. Interpretability is better understood as a capability the organisation has to build. In practice that means data governance that takes fairness seriously, and staff who can read an explanation and question it; it also means procurement rules that ask for auditable systems, not just for impressive plots. For policymakers the lesson is different. The evidence is concentrated in well-resourced higher-education settings, and methods validated there cannot be assumed to work in schools, or in low-resource systems, without adaptation.

5.6 Limitations of the Reviewed Literature

Several limitations of the source literature qualify these conclusions. The evidence rests disproportionately on a small set of public benchmarks, attenuating ecological validity; stakeholder evaluation is scarce; the literature is cross-sectional, precluding inference about the durability of explanation-informed interventions; fairness is under-examined; and a probable bias toward positive results may inflate the apparent maturity of the field.

5.7 Limitations of this Review

This review is itself subject to limitations. It was not prospectively registered. Of 135 eligible reports, 60 (44%) could not be retrieved within the review window, despite searches of institutional holdings and open-access sources. The synthesis therefore rests on the 55% that were obtainable. Comparing the available metadata (publication year and venue type) showed no marked difference between retrieved and non-retrieved reports, but the possibility of selection bias cannot be excluded, and the frequency estimates describe the retrievable literature rather than the eligible literature as a whole. Data extraction, coding and quality appraisal were all carried out by the first author, who read every included study in full. A rule-based text-mining script served as a first-pass aid; its suggestions were checked, and where necessary corrected, against the full texts. There was no second, independent coder, so no formal inter-rater reliability statistic could be computed. The coding scheme, the appraisal rubric, the script and the complete verified dataset are released instead, so that every assignment can be inspected and the synthesis repeated. Finally, the review is limited to SHAP and LIME. That choice reflects how dominant the two methods are, but it leaves out newer interpretability techniques that are starting to appear in educational research.

Finally, the included studies were published across a range of journals. A proportion, however, remained concentrated in journals such as IEEE Access and Education and Information Technologies. This pattern reflects the current publication landscape in educational machine learning and explainable AI (Türkmen, 2025). Study selection was conducted in accordance with the predefined eligibility criteria and the PRISMA 2020 guidelines (Page et al., 2021). Studies were selected solely on the basis of their relevance and methodological eligibility, irrespective of the journal in which they were published.

6. Future Research Directions

The gaps documented above translate into a concrete agenda. Five directions are especially pressing.

6.1 Human-Centred and Context-aware Explanation

Future work should move the locus of evaluation from the model to the stakeholder, developing and validating adaptive interfaces that calibrate the complexity and language of an explanation to the user's expertise, the decision at hand, and the cognitive load it imposes, and measuring whether such explanations actually improve educator decisions (Arora et al., 2021; Miller, 2019).

6.2 Multimodal and Longitudinal Interpretability

As education incorporates audio, video, and sensor data, attribution methods must extend beyond tabular and textual inputs, and static snapshots must give way to longitudinal accounts that track how a learner's explanations evolve over time, the precondition for using XAI in genuine progress monitoring (Di Mitri et al., 2018).

6.3 Fairness-Aware Explanation

Given the scarcity of fairness work, future research should embed counterfactual reasoning and bias auditing within the explanation pipeline itself, so that the act of explaining a decision also surfaces, and helps to remedy, its inequities (Madaio et al., 2020; Pham et al., 2025).

6.4 Equity Across Educational Contexts

Most of the evidence comes from well-resourced settings, and redressing that imbalance will take deliberate effort. Explainers need to be lightweight and able to run offline. Interfaces have to make sense in the local cultural context. And human-in-the-loop designs should be built together with the educators working in under-resourced systems, not handed down to them.

6.5 Standardisation and Governance

Finally, the field would benefit from shared educational XAI benchmarks and from interpretability protocols aligned with emerging regulation, so that explanation becomes an auditable institutional practice rather than an ad hoc technical addition (Simbeck, 2024).

7. Conclusion

SHAP now sits at the centre of interpretability work in educational ML. It appears in 74 of the 75 studies and has become the default way of reading the tree-based models that dominate tabular prediction. LIME appears in 33 studies, almost never on its own, and in practice plays a supporting role beside SHAP rather than competing with it. What prevails, in short, is a SHAP-centric and technically framed paradigm.

That consolidation is an achievement. It is also a limit. The field has learned how to attribute a prediction to its features; it has not yet learned whether educators understand those attributions, whether they treat learners fairly, or whether they still mean anything months later, as a student's situation changes. So, the central question has shifted. It is no longer which method to pick, but how to make explanations comprehensible to the people they are meant for, equitable in their effects, and valid over time. A field that takes this seriously would gain more from explainable AI than a window into its models. It would have a basis for educational decisions that can be trusted and accounted for.

8. Acknowledgements

The authors would like to thank the School of AI, Computing and Multimedia, Lincoln University College, for its support throughout this work.

9. Data Availability

The search exports, the coded and verified data-extraction workbook, the coding scheme and the supporting text-mining script with its decision rules, the quality-appraisal rubric and per-study scores, and the completed PRISMA 2020 checklist are openly available in a public repository (Zenodo; <https://doi.org/10.5281/zenodo.21620430>) and are also provided as supplementary materials, enabling full re-execution of the synthesis.

10. References

Aler Tubella, A., Mora-Cantalops, M., & Nieves, J. C. (2024). How to teach responsible AI in Higher Education: challenges and opportunities: A. Aler Tubella et al. *Ethics and Information Technology*, 26(1), 3.

- Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5), e1424.
- Arora, A. K., Rodriguez, C., Carver, T., Teper, M. H., Rojas-Rozo, L., & Schuster, T. (2021). Evaluating usability in blended learning programs within health professions education: a scoping review. *Medical science educator*, 31(3), 1213-1246.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115.
- Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, 31(6), 3360-3379.
- Dass, S., Gary, K., & Cunningham, J. (2021). Predicting student dropout in self-paced MOOC course using random forest model. *Information*, 12(11), 476.
- Deho, O. B., Zhan, C., Li, J., Liu, J., Liu, L., & Duy Le, T. (2022). How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? *British Journal of Educational Technology*, 53(4), 822-843.
- Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, 34(4), 338-349.
- Feldman-Maggor, Y., Cukurova, M., Kent, C., & Alexandron, G. (2025). The impact of explainable AI on teachers' trust and acceptance of AI EdTech recommendations: The power of domain-specific explanations. *International Journal of Artificial Intelligence in Education*, 1-34.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 1-42.
- Kalita, E., Alfawar, A. M., El Aouifi, H., Kukkar, A., Hussain, S., Ali, T., & Gaftandzhieva, S. (2025). Predicting student academic performance using Bi-LSTM: a deep learning framework with SHAP-based interpretability and statistical validation. In *Frontiers in Education* (Vol. 10, pp. 1581247). Frontiers.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1), 56-67.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1-38.
- Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 55(5), 3503-3568.
- Mustofa, S., Emon, Y. R., Mamun, S. B., Akhy, S. A., & Ahad, M. T. (2025). A novel AI-driven model for student dropout risk analysis with explainable AI insights. *Computers and Education: Artificial Intelligence*, 8, 100352.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., & Brennan, S. E. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*, 372.
- Pham, N., Ngoc, H. P., & Nguyen-Duc, A. (2025). Fairness for machine learning software in education: A systematic mapping study. *Journal of Systems and Software*, 219, 112244.
- Quimiz-Moreira, M., Delgadillo, R., Parraga-Alava, J., Maculan, N., & Mauricio, D. (2025). Factors, prediction, explainability, and simulating university dropout through machine learning: A systematic review, 2012–2024. *Computation*, 13(8), 198.

- 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).
- Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2023). Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence*, 46(4), 2104-2122.
- Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1), 2400304.
- Shukla, P., Singh, J. N., & Mall, S. (2026). Interpretable aspect based sentiment classification of online educational reviews using SVM model and explainable LIME-AI model. *International Journal of Information Technology*, 18(1), 655-667.
- Simbeck, K. (2024). They shall be fair, transparent, and robust: auditing learning analytics systems. *AI and Ethics*, 4(2), 555-571.
- Stracke, C. M., Chounta, I.-A., Holmes, W., Tlili, A., & Bozkurt, A. (2023). A standardised PRISMA-based protocol for systematic reviews of the scientific literature on Artificial Intelligence and education (AI&ED). *Journal of Applied Learning and Teaching*, 6(2), 64-70.
- Türkmen, G. (2025). The review of studies on explainable artificial intelligence in educational research. *Journal of Educational Computing Research*, 63(2), 277-310.
- Ujkani, B., Minkovska, D., & Hinov, N. (2024). Course Success Prediction and Early Identification of At-Risk Students Using Explainable Artificial Intelligence. *Electronics*, 13(21), 4157.
- Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, 4(1), 2.
- Visani, G., Bagli, E., Chesani, F., Poluzzi, A., & Capuzzo, D. (2022). Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 73(1), 91-101.
- Zhu, H., Sun, Y., & Yang, J. (2025). Towards responsible artificial intelligence in education: a systematic review on identifying and mitigating ethical risks. *Humanities and Social Sciences Communications*, 12(1), 1111.

Appendix A. Included Studies (n = 75)

The 75 studies synthesised in this review, listed alphabetically by first author. Complete coded data are provided on Zenodo (<https://doi.org/10.5281/zenodo.21620430>).

- A1.** Abdalkareem, M. (2024). Explainable Models for Predicting Academic Pathways for High School Students in Saudi Arabia. *IEEE Access*. doi:10.1109/access.2024.3369586
- A2.** Abukader, A. (2025). Intelligent System for Student Performance Prediction: An Educational Data Mining Approach Using Metaheuristic-Optimized LightGBM with SHAP-Based Learning Analytics. *Applied Sciences (Switzerland)*. doi:10.3390/app152010875
- A3.** Ahmmed, J. (2026). HPrEd: A Tuned Ensemble With Model-Agnostic XAI to Explain Social Media's Association With Academic Productivity. *IEEE Access*. doi:10.1109/access.2026.3663639
- A4.** Al Hashmi, R.A.M. (2025). Early detection of at-risk health sciences students: a machine learning-based predictive study using midterm grades. *BMC Medical Education*. doi:10.1186/s12909-025-08192-6
- A5.** Alcauter, I. (2023). Explaining Factors of Student Attrition at Higher Education. *Computacion y Sistemas*. doi:10.13053/cys-27-4-4776
- A6.** Asamoah, K.O. (2026). Analyzing user satisfaction in e-learning platforms: A text mining and explainable machine learning approach using unstructured data. *Applied Soft Computing*. doi:10.1016/j.asoc.2026.115450
- A7.** Aslam, M.A. (2024). A Human-Centered Approach to Academic Performance Prediction Using Personality Factors in Educational AI. *Information (Switzerland)*. doi:10.3390/info15120777
- A8.** Bettahi, A. (2025). A Modular and Explainable Machine Learning Pipeline for Student Dropout Prediction in Higher Education. *Algorithms*. doi:10.3390/a18100662
- A9.** Bettahi, A. (2025). Early Student Risk Detection Using CR-NODE: A Completion-Focused Temporal Approach with Explainable AI. *Algorithms*. doi:10.3390/a18120781
- A10.** Charytanowicz, M. (2023). Online Education vs Traditional Education: Analysis of Student Performance in Computer Science using Shapley Additive Explanations. *Informatics in Education*. doi:10.15388/infedu.2023.23
- A11.** Chen, H. (2026). Interpretable machine learning reveals key determinants of creative thinking among East Asian students: Evidence from PISA 2022. *Thinking Skills and Creativity*. doi:10.1016/j.tsc.2025.102069
- A12.** de la Hoz, E. (2026). Actionable learning analytics: predicting university performance levels with interpretable machine learning. *Journal on Efficiency and Responsibility in Education and Science*. doi:10.7160/eriesj.2026.190102
- A13.** Delprato, M. (2026). Effects of COVID-19 education policies on students' learning proficiency in Latin America—Insights from PISA 2022. *International Journal of Educational Development*. doi:10.1016/j.ijedudev.2025.103483
- A14.** Delprato, M. (2026). Identifying the post-pandemic determinants of low performing students in Latin America through interpretable Machine Learning methods. *Engineering Applications of Artificial Intelligence*. doi:10.1016/j.engappai.2026.115154
- A15.** Dima R. Badawi (2026). Enhancing transparency in educational question classification: an explainable AI approach using SHAP and LIME with CNN models. *Neural Computing and Applications*. doi:10.1007/s00521-026-11931-5
- A16.** Erdem, B. (2025). Cheating Detection in Online Exams Using Deep Learning and Machine Learning. *Applied Sciences (Switzerland)*. doi:10.3390/app15010400
- A17.** Ersoz, B. (2026). Evaluating the effects of adjunct VR-haptic simulator practice in preclinical restorative training and predicting student performance using machine learning: a randomized controlled trial. *BMC Medical Education*. doi:10.1186/s12909-026-09072-3
- A18.** Farhah, N. (2025). AI-Driven Innovation Using Multimodal and Personalized Adaptive Education for Students With Special Needs. *IEEE Access*. doi:10.1109/access.2025.3560863
- A19.** Gao, H. (2023). Exploring Gender Differences in Computational Thinking Learning in a VR Classroom: Developing Machine Learning Models Using Eye-Tracking Data and Explaining the Models. *International Journal of Artificial*

Intelligence in Education. doi:10.1007/s40593-022-00316-z

- A20. Ghimire, S. (2024). Explainable artificial intelligence-machine learning models to estimate overall scores in tertiary preparatory general science course. *Computers and Education: Artificial Intelligence*. doi:10.1016/j.caeai.2024.100331
- A21. Gu, H. (2026). Efficient and Interpretable Machine Learning for Student Academic Outcome Prediction. *Mathematics*. doi:10.3390/math14040626
- A22. Gunasekara, S. (2025). Explainable AI in Education: Techniques and Qualitative Assessment. *Applied Sciences (Switzerland)*. doi:10.3390/app15031239
- A23. Hoq, M. (2024). Explaining Explainability: Early Performance Prediction with Student Programming Pattern Profiling. *Journal of Educational Data Mining*. doi:10.5281/zenodo.14246435
- A24. Hossen, M.K. (2025). From data to insights: Using gradient boosting classifier to optimize student engagement in online classes with explainable AI. *Education and Information Technologies*. doi:10.1007/s10639-025-13500-0
- A25. Hou, N. (2026). Enhancing Fairness and Explainability in Student Performance Prediction Using Bias Mitigation Techniques. *Information Technology and Control*. doi:10.5755/j01.itc.55.1.39399
- A26. Hu, J. (2023). Decoding Contextual Factors Differentiating Adolescents' High, Average, and Low Digital Reading Performance Through Machine-Learning Methods. *IEEE Transactions on Learning Technologies*. doi:10.1109/tlt.2023.3281056
- A27. Husayn, M. (2025). GWO-Optimized Ensemble Learning for Interpretable and Accurate Prediction of Student Academic Performance in Smart Learning Environments. *Applied Sciences (Switzerland)*. doi:10.3390/app152212163
- A28. Javadi, B. (2026). Evaluating Classification Machine Learning Models for Identifying External Factors Influencing Student Choices in Virtual Learning Environments. *International Journal of Information and Education Technology*. doi:10.18178/ijiet.2026.16.1.2501
- A29. Johora, F.T. (2025). An explainable AI-based approach for predicting undergraduate students academic performance. *Array*. doi:10.1016/j.array.2025.100384
- A30. Kala, A. (2024). Early Prediction of Student Performance in Face-to-Face Education Environments: A Hybrid Deep Learning Approach With XAI Techniques. *IEEE Access*. doi:10.1109/access.2024.3516816
- A31. Kalita, E. (2025). Predicting student academic performance using Bi-LSTM: a deep learning framework with SHAP-based interpretability and statistical validation. *Frontiers in Education*. doi:10.3389/feduc.2025.1581247
- A32. Kalita, E. (2025). LSTM-SHAP based academic performance prediction for disabled learners in virtual learning environments: a statistical analysis approach. *Social Network Analysis and Mining*. doi:10.1007/s13278-025-01484-1
- A33. Kar, S.P. (2024). Assessment of learning parameters for students' adaptability in online education using machine learning and explainable AI. *Education and Information Technologies*. doi:10.1007/s10639-023-12111-x
- A34. Kartik, N. (2023). XAI-Based Student Performance Prediction: Peeling Back the Layers of LSTM and Random Forest's Black Boxes. *SN Computer Science*. doi:10.1007/s42979-023-02070-y
- A35. Kesgin, K. (2025). Beyond Performance: Explaining and Ensuring Fairness in Student Academic Performance Prediction with Machine Learning. *Applied Sciences (Switzerland)*. doi:10.3390/app15158409
- A36. Khine, M.S. (2025). From Meals to Marks: Modeling the Impact of Family Involvement on Reading Performance with Counterfactual Explainable AI. *Education Sciences*. doi:10.3390/educsci15070928
- A37. Kim, H. (2026). Understanding dropout risks in online courses: SHAP-based analysis of lecture interval patterns in LMS. *Education and Information Technologies*. doi:10.1007/s10639-026-13994-2
- A38. Kim, Hajung (2026). Automated Scoring of Young Learners' Student-Generated Questioning in Elementary Reading: A Machine-Learning Approach. *Computers & Education*. doi:10.1016/j.compedu.2026.105674
- A39. Kwon, J.B. (2026). A Portable, Generalizable Machine Learning Framework for Long-Term Student Dropout Prediction. *IEEE Access*. doi:10.1109/access.2026.3666870
- A40. Li, P. (2025). Application of Deep Learning-Based Personalized Learning Path Prediction and Resource Recommendation for Inheriting Scientist Spirit in Graduate Education. *Computer Science and Information Systems*.

doi:10.2298/csis241125043l

- A41.** Liu, F. (2025). Exploring the Impact of Color and Daytime/Nighttime on Visual Memory in Virtual Reality Teaching Environments Using Explainable Machine Learning. *Intelligent Computing*. doi:10.34133/icomputing.0155
- A42.** Low, J. (2026). Explainable Multi-View Modeling of AI-Driven Personalized Learning Adoption in TVET Systems. *Journal of Logistics, Informatics and Service Science*. doi:10.33168/jliss.2026.0408
- A43.** Luo, Z. (2024). A Method for Prediction and Analysis of Student Performance That Combines Multi-Dimensional Features of Time and Space. *Mathematics*. doi:10.3390/math12223597
- A44.** Mingyu, Z. (2022). An interpretable prediction method for university student academic crisis warning. *Complex and Intelligent Systems*. doi:10.1007/s40747-021-00383-0
- A45.** Mustofa, S. (2025). A novel AI-driven model for student dropout risk analysis with explainable AI insights. *Computers and Education: Artificial Intelligence*. doi:10.1016/j.caeai.2024.100352
- A46.** Noh, H. (2025). Explainable Machine Learning for Predicting Computational Thinking Proficiency in K-12 Students Using National-Level Data. *IEEE Access*. doi:10.1109/access.2025.3634996
- A47.** Nti, I.K. (2026). Explainable machine learning for student dropout prediction and tailored interventions in online personalized education. *Discover Artificial Intelligence*. doi:10.1007/s44163-026-01016-6
- A48.** Ofori, E. (2025). Explainable artificial intelligence in LSTM transformer models for student performance analysis. *Discover Computing*. doi:10.1007/s10791-025-09814-9
- A49.** Olipas, C.N.P. (2026). Predictive modeling and explainability of student employability in the Philippines using random forest and Shapley additive explanations. *Interdisciplinary Journal of Information, Knowledge, and Management*. doi:10.28945/5690
- A50.** Osunbunmi, I. (2025). Artificial intelligence in engineering education research: Using machine learning models to predict undergraduate engineering students' persistence to graduation. *Journal of Engineering Education*. doi:10.1002/jee.70034
- A51.** Qu, Y. (2022). Can We Predict Student Performance Based on Tabular and Textual Data? *IEEE Access*. doi:10.1109/access.2022.3198682
- A52.** Raihan, M. (2024). Bridging Technology and Psychology: AI-Driven Analysis of Student's Class Lecture Activity for Improved Learning Outcomes. *IEEE Access*. doi:10.1109/access.2024.3474848
- A53.** Raji, N.R. (2024). Explainable Machine Learning Prediction for the Academic Performance of Deaf Scholars. *IEEE Access*. doi:10.1109/access.2024.3363634
- A54.** S, H. (2025). Interpretable and ethical learning assessment transformer (IELAT): an explainable transformer model for personalized student assessments. *Cogent Education*. doi:10.1080/2331186x.2025.2485508
- A55.** Sa Yang (2025). Identifying factors influencing online learning outcomes for middle-school students — a re-examination based on XGBoost and SHAP. *Education and Information Technologies*. doi:10.1007/s10639-025-13405-y
- A56.** Sahlaoui, H. (2024). An empirical assessment of smote variants techniques and interpretation methods in improving the accuracy and the interpretability of student performance models. *Education and Information Technologies*. doi:10.1007/s10639-023-12007-w
- A57.** Sanfo, J.-B.M.B. (2025). Application of explainable artificial intelligence approach to predict student learning outcomes. *Journal of Computational Social Science*. doi:10.1007/s42001-024-00344-w
- A58.** Sharma, M. (2026). Blended Ensemble Intelligence for Academic Performance Prediction in Educational Datasets. *SN Computer Science*. doi:10.1007/s42979-026-04919-4
- A59.** Shoaib, M. (2026). Forecasting algorithm for the vocational mobility of graduates with multiple attributes. *Expert Systems with Applications*. doi:10.1016/j.eswa.2025.129463
- A60.** Sieg, M. (2023). Discovering unknown response patterns in progress test data to improve the estimation of student performance. *BMC Medical Education*. doi:10.1186/s12909-023-04172-w

- A61. Silva, F.d.C. (2025). An Investigation Into Dropout Indicators in Secondary Technical Education Using Explainable Artificial Intelligence. *Revista Iberoamericana de Tecnologías del Aprendizaje*. doi:10.1109/rita.2025.3566095
- A62. Sun, Y. (2026). Exploring the causal effects of ICT on creative thinking: A perspective of academic performance differences. *Thinking Skills and Creativity*. doi:10.1016/j.tsc.2025.102004
- A63. Taqatqeh, S. (2026). Early warning and recommendation system for undergraduate students at higher education institutions based on machine learning methods. *PeerJ Computer Science*. doi:10.7717/peerj-cs.3792
- A64. Trejo-Macotela, F.R. (2026). Artificial Intelligence, Academic Resilience, and Gender Equity in Education Systems: Ethical Challenges, Predictive Bias, and Governance Implications. *Education Sciences*. doi:10.3390/educsci16040605
- A65. Ujkani, B. (2024). Course Success Prediction and Early Identification of At-Risk Students Using Explainable Artificial Intelligence. *Electronics (Switzerland)*. doi:10.3390/electronics13214157
- A66. Umendu, E.C. (2026). Enhancing Student Retention in Higher Education Institutions (HEIs): Machine Learning Approach. *Electronics (Switzerland)*. doi:10.3390/electronics15040734
- A67. Wang, H. (2026). Beyond prediction: integrating machine learning and structural modeling to explain student performance. *Evolutionary Intelligence*. doi:10.1007/s12065-026-01192-5
- A68. Xia, Z. (2024). An interpretable English reading proficiency detection model in an online learning environment: A study based on eye movement. *Learning and Individual Differences*. doi:10.1016/j.lindif.2023.102407
- A69. Xie, Y. (2025). An analysis of predictive modeling and factors influencing CET-4 pass rate among Chinese university students: a machine learning-based approach. *Education and Information Technologies*. doi:10.1007/s10639-024-12964-w
- A70. Zarei, M. (2025). A hybrid ANFIS-gradient boosting frameworks for predicting advanced mathematics student performance. *Iranian Journal of Fuzzy Systems*. doi:10.22111/ijfs.2025.52017.9178
- A71. Zhang, C. (2023). E-Learning at-Risk Group Prediction Considering the Semester and Realistic Factors. *Education Sciences*. doi:10.3390/educsci13111130
- A72. Zhang, L. (2025). LSTM-Based Student Performance Prediction Model Incorporating Peer Effects and Explained Through SHAP. *Journal of Engineering (United Kingdom)*. doi:10.1155/je/5731919
- A73. Zine, M. (2025). Evaluating e-learning readiness using explainable machine learning and key organizational change factors in higher education. *Education and Information Technologies*. doi:10.1007/s10639-025-13335-9
- A74. Zine, M. (2023). E-Learning Readiness Assessment Using Machine Learning Methods. *Sustainability (Switzerland)*. doi:10.3390/su15118924
- A75. Zine, M. (2025). Understanding Student Behavioral Dynamics in Moodle Through Chaos Theory and Interpretable Machine Learning. *SN Computer Science*. doi:10.1007/s42979-025-04595-w