Explainability in Educational Data Mining and Learning Analytics: An Umbrella Review

Sachini Gunasekara
Faculty of Information Technology
University of Jyväskylä
savisama@jyu.fi

Mirka Saarela
Faculty of Information Technology
University of Jyväskylä
mirka.saarela@jyu.fi

ABSTRACT

This paper presents an umbrella review synthesizing the findings of explainability studies within the Educational Data Mining (EDM) and Learning Analytics (LA) domains. By systematically reviewing existing reviews and adhering to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines, we identified 49 secondary studies, culminating in a final corpus of 10 studies for rigorous systematic review. This approach offers a comprehensive overview of the current state of explainability research in educational models, providing insights into methodologies, techniques, outcomes, and the effectiveness of explainability implementations in educational contexts, including the impact of data types, models, and metrics on explainability. Our analysis unveiled that latent variables, already offering a higher semantic level, are typically easier to interpret, but observed variables can increase actionability. Moreover, while older studies accentuate the benefits of decision tree models for their intrinsic explainability and minimal need for additional explanation techniques, recent research favors more complex models and post-hoc explanation methods. Surprisingly, not a single publication in our corpus discussed metrics for evaluating the effectiveness or quality of explanations. However, a subset of articles in our collection addressed metrics for model performance and fairness in educational settings. Selecting optimal data types, models, and metrics promises to enhance transparency, interpretability, and accessibility for educators and students alike.

Keywords

explainable artificial intelligence, educational data mining, learning analytics, explainability, umbrella review

1. INTRODUCTION

In recent years, the integration of Artificial Intelligence (AI) technologies in educational settings has garnered significant attention for its potential to revolutionize teaching and learning practices [9]. However, the black-box nature of AI algorithms often presents challenges in understanding and interpreting their decision-making processes, particularly in the context of Educational Data Mining (EDM) and Learning Analytics (LA). Explainable Artificial Intelligence (XAI) has emerged as a promising approach to address this issue by providing insights into how AI models arrive at their conclusions [7, 23]. It tries to foster confidence in automated procedures by providing clear insights into the workings of AI algorithms. Through this, it enables, for example, early intervention for troubling students [16], personalized learning experiences [1, 10], and well-informed decision-making in educational contexts by ensuring the interpretability of AI systems [30]. In general, the use of XAI in education is designed to optimize learning outcomes and advance educational fairness by combining the benefits of AI with a comprehensive grasp of its decision-making processes [17]. This paper conducts an umbrella review, synthesizing existing reviews of explainability studies within the EDM and LA fields. By systematically analyzing the findings of multiple reviews, this study aims to offer a comprehensive overview of the current landscape of explainability research in EDM and LA, identify key trends and challenges, and provide insights for future research directions to enhance the interpretability and transparency of AI-driven educational technologies.

2. METHODOLOGY

The approach for employing Systematic Literature Review (SLR) methodology was developed in accordance with the guidelines established by [8]. As research studies on XAI increase across various application domains, the underlying knowledge becomes progressively disorganized [33]. A minimal amount of secondary research, nevertheless, has been carried out with the specific objective of organizing the profuse knowledge about explainability methods and the challenges associated with LA and EDM. As a result, it is essential to conduct an SLR to compile, analyze, and present a comprehensive and unbiased overview of the secondary articles regarding the role and significance of explainability in LA and EDM.

2.1 Research questions

The following research questions (RQs) were developed in light of the need to perform the SLR of the existing approaches for delivering explainable EDM or LA and their assessments in various applications and activities in education:

S. Gunasekara and M. Saarela. Explainability in educational data mining and learning analytics: An umbrella review. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 887–892, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729987

RQ1. What is the influence of different data types and models on the explainability of EDM and LA models?

RQ2. What metrics are suitable for evaluating the explainability of these models?

More precisely, our RQs examine the correlations that have been captured between data types, models, and metrics, challenges, and future research needs for effective implementation of explainability in LA and EDM. Through this, we aim to shed light on the most efficient data types and models that impact the explainability of models and the various explanation structures that were produced in education.

2.2 Identifying relevant research articles

We conducted a systematic keyword search utilizing the following six databases through the identification phase: ACM digital library, IEEE Xplore digital library, Springer, Science Direct, Web of Science, and SCOPUS. The search query was contextualized in three dimensions: Explainability, AI, and Education. As this work focuses on XAI in EDM and LA, we gave keywords that define each dimension. We looked for terms that represented the dimensions of EDM and LA in the Title, Abstract, and Author keywords fields, and for explainability terms in the full text of the articles. Since XAI is a relatively new field of study, the search result was filtered to include articles published in 2000 or later. All of the keywords for articles regarding explainability and EDM/LA were gathered using the boolean operators ”AND” and ”OR” in the search. The following is a summary of the search strings that were generated and adapted to meet the advanced search criteria of each database (where different spellings and plural are indicated by the wildcard *): ((interpretab*, transparen*,explainab*, explanation*, intelligib*,XAI) AND (”educational data mining” OR EDM OR ”learning analytics” OR LA)). Within each database's search tool constraints, the same search string was used for all of them. Journal articles and conference papers are among the items being looked up.

The selection of these databases is based on their reliability and international, multidisciplinary nature. They serve as extensive knowledge databases, offering comprehensive citation indexing coverage and providing access to the highest-quality data from scientific publications. To address the proposed research questions of the umbrella review known as a “review of reviews” [14] comprehensively, an exhaustive search for research publications was conducted. According to the PRISMA guidelines [24, 27], the SLR followed the analytical process. The primary stages of our entire systematic review are outlined in Figure 1.

2.2.1 *Inclusion and Exclusion criteria*

To find possible research publications, inclusion and exclusion criteria were established; the results are shown in Table 1. Peer-reviewed English-language publications on explainable EDM or LA published in peer-reviewed international conference proceedings and journals met the requirements to be included in the SLR. Articles unrelated to the study topics met the criteria to be excluded from the SLR. These inclusion and exclusion criteria were taken into consideration throughout the article selection process.

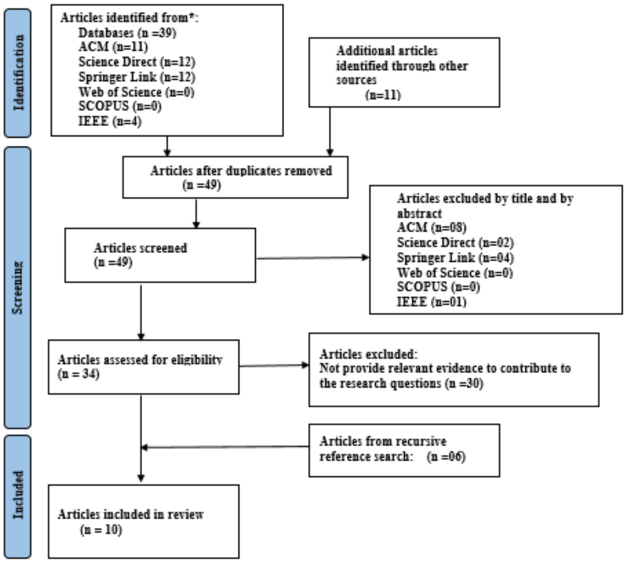
Identification
Screening
Included
Articles identified from*:
Databases (n =39)
ACM (n=11)
Science Direct (n=12)
Springer Link (n=12)
Web of Science (n=0)
SCOPUS (n=0)
IEEE (n=4)
Additional articles identified through other sources
(n=11)
Articles after duplicates removed
(n =49)
Articles excluded by title and by abstract
ACM (n=08)
Science Direct (n=02)
Springer Link (n=04)
Web of Science (n=0)
SCOPUS (n=0)
IEEE (n=01)
Articles screened
(n =49)
Articles assessed for eligibility
(n = 34)
Articles excluded:
Not provide relevant evidence to contribute to the research questions (n =30)
Articles from recursive reference search: (n =06)
Articles included in review
(n = 10)

Figure 1: Umbrella review PRISMA chart.

2.2.2 *Data extraction and quality assessment*

By employing both automatic and manual searches, the initial stage of PRISMA finds possible research to look into. The identification step is followed by the screening phase of the research, which identifies and removes duplicated and unnecessary research. After the qualifying articles are carefully reviewed and their eligibility is determined, the final list of research to be included in our synthesis is determined. We adhered rigorously to the inclusion criteria specified in Table 1 during the screening and eligibility phases. The synthesis did not include any studies that made explicit reference to explainable EDM or LA, except for one workshop paper [12] that coined this term.

According to our collected review studies on explainable EDM and LA, which involved 50 articles gathered from the initial round of automated searches conducted on the six above-mentioned databases and other resources, 49 possibly relevant articles remained after eliminating duplicate publications and scanning titles and abstracts. Following a comprehensive review of all eligible articles, 34 relevant articles were found in the search results; several articles were excluded due to lacking relevance to explainability. 30 publications were further eliminated throughout the extraction process because we could not find enough information in them to address either of our research questions. As a result, four publications [2, 13, 19, 22] were determined to be relevant to the explainability of EDM and LA models.

Additionally, through the following references, based on the snowball method, the authors also conducted manual searches to take into consideration another six articles that broadly address the concept of explainability in EDM and LA [5, 12, 17, 21, 25, 34]. The work by De Laet et al. [12] explores the topic, detailing the outcomes of a workshop that investigated the potential and hurdles of XAI in LA. Conversely, Khosravi et al. [17] propose a framework encompassing six pivotal dimensions pertinent to explainability, tailored for

Table 1: Inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
Published in journals/conferences	Pre-prints and duplicates
Peer reviewed	Review articles, book chapters, magazines, and editorials
Published from 2000 to January 2024	Not relevant to research objectives
Written in English	Studies were published before 2000

the examination, design, and advancement of explainable AI tools in educational settings. The review by Shahiri et al. [34] identified several commonly used observed variables in the prediction process for educational models. Bond et al., Masruroh et al. and Namoun et al. [5, 21, 25] evaluated the most often utilized metrics for the educational models' predictions. As a result, a final sample of 10 publications (that were published in established publication channels [32]) was then included in the umbrella review.

3. DATA ANALYSIS AND RESULTS

When making predictive models, a significant tradeoff arises between achieving high accuracy and maintaining transparency. As accuracy improves, the comprehensibility and transparency of the model tend to decrease for humans. To overcome this challenge, tools and explainable models can be used to make models more comprehensible. To address the research questions, we carried out a thorough qualitative analysis of the studies selected from three perspectives: (i) data types fed into the educational models; (ii) EDM and LA models and their applications in various educational tasks; and (iii) metrics used for measuring model performance, explainability, and fairness.

Result on RQ1: What is the influence of different data types and models on the explainability of EDM and LA models?

Data in the realm of education can be categorized into several types, such as observed and latent, according to its format, content, and structure. Several student-oriented features, including "achievement", "engagement", "participation", "satisfaction", "motivation", and "reflection" were shown to be reachable through using LA techniques in the review by Aldowah et al. [2], which comprised 402 publications from 2000 to 2017. The explainability of models may be directly enhanced by observed variables, which are typically easier to comprehend. The explainability of latent variables in education, such as learning motivation, reflection, and satisfaction, presents challenges. Extra processing and interpretation are needed to get useful insights from that kind of data. Models should highlight the significance that categorical variables play in decision-making by illustrating how they affect predictions.

Model interpretability is improved when categorical impacts are communicated transparently. Student data variables like gender (such as male, female, or other), grade levels (such as elementary, middle school, or high school), or student status (such as enrolled, graduated, or dropped out) are the most commonly used categorical data when it comes to data for such forecasts. In the review by Shahiri et al. [34] from full-text research papers from 2002 to 2015, the formers highlighted that the cumulative grade point average, demographics, external assessments, high school background, and social interaction network are the criteria that have been employed most frequently in predicting students' performance using LA. Moreover, after their research, they concluded that "the result on prediction accuracy depends on the attributes or features that were used during the prediction process" [34].

The majority of research focuses on predicting student performance by using classification models, particularly for outcomes such as attrition, desertion, and crucially, dropout or failure risk. For this, especially tree-structured algorithms were preferred as they are intrinsically explainable while also showing relatively high prediction performance [2, 13, 19]. Li et al. [19] discussed the benefits of decision trees (DT), pointing out that they do not require data normalization and are easy to handle, understand, and analyze in the context of noisy data. Moreover, they are generally easy to explain and illustrate to domain experts. Additionally, in a study conducted by Masruroh et al. [21], a total of 21 models were employed. Among these, it was highlighted that the DT emerged as the most effective technique for academic predictive data mining.

Shahiri et al. [34] pointed out that Neural Network (NN) models had the highest prediction accuracy followed by DT. However, while NNs had a higher predictive power in that study, they are generally known as opaque prediction models, lacking the self-explanatory nature of DTs, and therefore requiring post-hoc explanation techniques for better understandability. Moreover, some reviews highlighted the Support Vector Machine (SVM) approach: Through using a SVM, the only linear model that can classify data that is not linearly separable [4], it is possible, for example, to sort applications related to admission/timetabling, career paths/placement, and student happiness with a higher accuracy rate, in addition to predicting student success and preventing attrition [5, 34]. Nevertheless, in comparison to DTs and other tree-based (ensemble) methods, which are explainable despite being non-linear [6], SVMs are presumed less explainable. Thus, there is usually a trade-off between performance and explainability, especially when the given data is only non-linearly separable [15, 30, 31]. Currently, there is a lack of consensus among EDM and LA studies to determine which algorithms work best with regard to both objectives (i.e., performance and explainability). However, using state-of-the-art deep learning techniques for high predictive performance in combination with post-hoc explainability techniques, such as SHapley Additive exPlanations (SHAP), can yield countless benefits [17].

Result on RQ2: What metrics are suitable for evaluating the explainability of these models?

None of the articles in our corpus mentioned metrics to measure the quality or performance of explanations. However, one article [12] indirectly mentioned that such metrics are needed by emphasizing that the main opportunity lies in finding ways "related to the evaluation of the impact of the explanations" of educational models. In fact, the lack of reporting metrics to measure explanation quality was also highlighted in a recent review article outside the educational bubble: According to Nauta et al. [26], only one in three XAI studies evaluates explanations with anecdotal evidence, and only one in five studies evaluate explanations with users.

Nevertheless, other metrics were discussed. More specifically, two of the articles in our corpus examined educational model performance metrics [13, 19], and two articles discussed fairness metrics [17, 22]. Several works point out that predictive performance, explainability, and fairness are interwoven and that it is important to optimize for all of these objectives [3, 11, 17, 18, 28]. Khosravi et al. [17] presented research suggesting how certain user attributes affect how explanations impact students' learning and how they perceive the adaptive hints offered by the educational system. This research offers perspectives on how personalization metrics could enhance the effectiveness of explanations in this context.

Li et al. [19] found that the influential factors impacting retention metrics were summarised into seven categories: behavioral engagement, student personal characteristics, student enrollment properties, prior academic performance, academic engagement, current academic performance, and course design. Memarian et al. [22] conducted a study to investigate that the models' interpretability, fairness, and transparency are greatly impacted by the representation that is used. As well, the study highlights the importance of fairness in particular, pointing out that it is the most emphasized phrase throughout the examined research. It has the ability to prevent discrimination and bias, as well as ensure learners, educators, and other stakeholders equitable access to educational opportunities. By placing significant emphasis on transparency and accountability, XAI creates a learning environment where stakeholders can comprehend and have trust in their decision-making processes.

4. DISCUSSION

From an initial set of 49 LA and EDM review/secondary studies, we identified 10 studies explicitly discussing explainability. By reviewing these existing reviews, we found that DT-based models that are intrinsically interpretable are often preferred in educational contexts, where clear insights are crucial.

The interpretation of machine learning models is significantly influenced by the nature of educational data. Our review of the literature revealed that observed variables may directly improve the actionability of models. Conversely, latent variables are often easier to understand but necessitate advanced techniques for meaningful interpretation, which makes feature and model selection even more crucial for producing insightful results.

In our corpus, no article addressed any metrics that evaluate the quality or efficacy of explanations. While several articles discussed different model performance and fairness metrics, no such discussion existed for explainability. We see this, that is, the development and use of appropriate explanation metrics as a main realm for future work. Only with such metrics can we effectively evaluate and compare explainability in LA and EDM models.

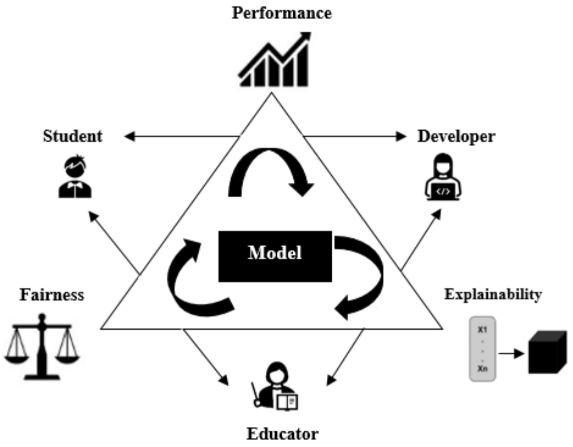

Figure 2: Interplay of Explainability-Performance-Fairness.

More specifically, we suggest employing metrics that concentrate on model-based explanations, utilizing the task model or newly built models to evaluate the quality of explanations. An understanding of the simplicity and comprehensibility of both local and global explanations may be gained by taking into consideration factors such as model size, runtime operation counts, major effect complexity, and interaction intensity. Metrics like effective complexity, mutual information, and continuity can also be used to further improve the assessment process by ensuring that explanations are clear, informative, and consistent in various contexts. For post-hoc explanations, metrics such as robustness, stability, and fidelity [29] can be employed. This will make it easier for model developers to reliably measure and compare the explanation quality—and to transparently communicate this to educators and other domain experts.

Additionally, we recommend that the educational community adopt a methodological shift to produce accurate and fair AI to assist learning and reduce bias, as well as incorporate a perception analysis for various demographics to analyze the individual shift in perception under different scenarios. For this, it is also important to have reliable metrics for the explanations, as explanations should be of similar quality for all subgroups to ensure fairness [3, 11].

Figure 2 represents our view of the interplay of the three critical metrics (i.e., *explainability*, *fairness*, and *performance*) that should be balanced when developing and employing educational models. Achieving this balance and incorporating the interests and needs of all concerned parties (such as students, educators, as well as model developers) is essential for building trust, enhancing learning outcomes, and promoting equity in the educational domain.

5. CONCLUSION

In this study, we conducted an umbrella review (i.e., a review of reviews) on explainability in LA and EDM studies. We examined the explainability findings of these studies, particularly with regard to used data types and employed metrics. We found that latent variables (such as *active reading* [20]) are easier to explain to domain experts, while observed variables typically provide more actionability. More specifically, it might be easier to comprehend AI predictions and their explanations based on latent variables (such as, *actively reading students perform better*); but to provide recommendations or actions for better performance, we should know from which measurements the latent variables were constructed. Simpler and less transformed data types (such as *uses annotations, answers quiz linked to the content* [20] instead of *actively reading*) directly tell the associated actions to design recommendations and draw insightful conclusions.

Several of the LA and EDM review articles emphasized that the most used algorithms are supervised machine learning models. Regarding the explainability, especially the benefits of DT and other tree-based algorithms were highlighted, such as being intrinsically explainable and understandable to domain experts while simultaneously also showing relatively high prediction performance. Some highly accurate models, such as deep neural networks, are inherently opaque. However, it is an ongoing attempt to develop methods to improve the interpretability of such complex models without compromising their performance. One possibility of doing that is through the use of post-hoc explainability methods, which some of the more recent studies employed.

Ongoing challenges persist in ensuring user understanding and trust, adapting to dynamic educational contexts, addressing ethical concerns, and overseeing regulatory compliance. Model performance, explainability, and fairness are intertwined (see Figure 2), yet current studies in LA and EDM lack a systematic approach to identify the most effective algorithms or optimize for these objectives simultaneously. To overcome these challenges, collaboration among educators, specialists, and AI practitioners is essential. Developing explainable LA and EDM systems become imperative to enhance prediction performance, ensuring transparency, user understanding, model and outcome explainability, and addressing ethical concerns within educational settings. Finally, metrics are needed to not only reliably measure performance and fairness, but also explainability in LA and EDM models.

6. ACKNOWLEDGMENTS

This work was supported by the Academy of Finland (project no. 356314).

7. ACRONYMS

AI Artificial Intelligence

DT Decision Tree

EDM Educational Data Mining

FATE Fairness, Accountability, Transparency, Ethical

LA Learning Analytics

NN Neural Network

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analysis

SHAP SHapley Additive exPlanations

SLR Systematic Literature Review

SVM Support Vector Machine

XAI Explainable Artificial Intelligence

8. REFERENCES

[1] A. Aileen Shibani Michael Xavier, R. Ratnavel, S. Selvaraj, F. Mattins, and D. Chinnappa. Explainable models for feedback design: An argumentative writing example. In *The 16th International Conference on Educational Data Mining*, 2023.

[2] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, 37:13–49, 2019.

[3] A. Balagopalan, H. Zhang, K. Hamidieh, T. Hartvigsen, F. Rudzicz, and M. Ghassemi. The road to explainability is paved with bias: Measuring the fairness of explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1194–1206, 2022.

[4] K. P. Bennett and C. Campbell. Support vector machines: hype or hallelujah? *ACM SIGKDD explorations newsletter*, 2(2):1–13, 2000.

[5] M. Bond, H. Khosravi, M. De Laat, N. Bergdahl, V. Negrea, E. Oxley, P. Pham, S. W. Chong, and G. Siemens. A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour, dec 2024.

[6] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

[7] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar. A review of trustworthy and explainable artificial intelligence (xai). *IEEE Access*, 2023.

[8] B. K. Charters and S. Guidelines for performing systematic literature reviews in software engineering. *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*, 1(October):1–54, 2007.

[9] T. K. Chiu, Q. Xia, X. Zhou, C. S. Chai, and M. Cheng. Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 4:100118, 2023.

[10] C. Conati, O. Barral, V. Putnam, and L. Rieger. Toward personalized xai: A case study in intelligent tutoring systems. *Artificial intelligence*, 298:103503, 2021.

[11] J. Dai, S. Upadhyay, U. Aivodji, S. H. Bach, and H. Lakkaraju. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 203–214, 2022.

[12] T. De Laet, M. Millecamp, T. Broos, R. De Croon, K. Verbert, and R. Duorado. Explainable learning analytics: challenges and opportunities. In *Companion Proceedings of the 10th International Conference on Learning Analytics & Knowledge LAK20 Society for Learning Analytics Research (SoLAR)*, pages 500–510, 2020.

[13] C. F. de Oliveira, S. R. Sobral, M. J. Ferreira, and F. Moreira. How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review. *Big Data and Cognitive Computing*, 5(4), dec 2021.

[14] M. J. Grant and A. Booth. A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, 26(2):91–108, 2009.

[15] V. Heilala, P. Jääskelä, T. Kärkkäinen, and M. Saarela. Understanding the study experiences of students in low agency profile: Towards a smart education approach. In *International conference on smart Information & communication Technologies*, pages 498–508. Springer, 2019.

[16] M. Hoq, P. Brusilovsky, and B. Akram. Analysis of an explainable student performance prediction model in an introductory programming course. *International Educational Data Mining Society*, 2023.

[17] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gašević. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074, 2022.

[18] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 131–138, 2019.

[19] C. Li, M. Li, C. L. Huang, Y. T. Tseng, S. H. Kim, and S. Yeom. Educational Data Mining in Prediction of Students' Learning Performance: A Scoping Review. In *IFIP Advances in Information and Communication Technology*, volume 685 AICT, pages 361–372. Springer Science and Business Media Deutschland GmbH, 2023.

[20] R. Majumdar, K. Takami, and H. Ogata. Learning with explainable ai-recommendations at school: Extracting patterns of self-directed learning from learning logs. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 245–249. IEEE, 2023.

[21] S. U. Masruroh, D. Rosyada, Zulkifli, Sururin, and N. A. R. Vitalaya. Adaptive Recommendation System in Education Data Mining using Knowledge Discovery for Academic Predictive Analysis: Systematic Literature Review. *2021 9th International Conference on Cyber and IT Service Management, CITSM 2021*, pages 1–6, 2021.

[22] B. Memarian and T. Doleck. Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI), and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, 5:100152, 2023.

[23] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66, 2022.

[24] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, P. Group, et al. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *International journal of surgery*, 8(5):336–341, 2010.

[25] A. Namoun and A. Alshanqiti. Predicting student performance using data mining and learning analytics techniques: A systematic literature review, jan 2021.

[26] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[27] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery*, 88:105906, 2021.

[28] M. Saarela. On the relation of causality- versus correlation-based feature selection on model fairness. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing (SAC '24)*, pages 56–64, 2024.

[29] M. Saarela and L. Geogieva. Robustness, stability, and fidelity of explanations for a deep skin cancer classification model. *Applied Sciences*, 12(19):9545, 2022.

[30] M. Saarela, V. Heilala, P. Jääskelä, A. Rantakaulio, and T. Kärkkäinen. Explainable student agency analytics. *IEEE Access*, 9:137444–137459, 2021.

[31] M. Saarela and S. Jauhiainen. Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3(2):1–12, 2021.

[32] M. Saarela and T. Kärkkäinen. Can we automate expert-based journal rankings? Analysis of the Finnish publication indicator. *Journal of Informetrics*, 14(2):101008, 2020.

[33] A. Saranya and R. Subhashini. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, page 100230, 2023.

[34] A. M. Shahiri, W. Husain, and N. A. Rashid. A Review on Predicting Student's Performance Using Data Mining Techniques. In *Procedia Computer Science*, volume 72, pages 414–422. Elsevier B.V., 2015.