Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 16 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i16.62224

PAPER

The Role of Synthetic Data in Educational Research: A Systematic Review

Ziyaeddin Halid İpek,
Erol Poyraz,
Gökhan Yigit(✉)

Kafkas University,
Kars, Turkey

gyigit@kafkas.edu.tr

ABSTRACT

The primary aim of this study is to provide a comprehensive and structured synthesis of existing research to understand how synthetic data is conceptualized, generated, and utilized within educational contexts. By analyzing 29 peer-reviewed articles, the research identifies seven primary dimensions of application: privacy and data sharing, data augmentation, NLP/text generation, predictive modeling, pedagogical design, methodological analysis, and synthetic data in mobile, interactive, and adaptive learning systems. A significant finding is the increasing integration of artificial intelligence (AI) and machine learning technologies, such as generative adversarial networks (GANs) and large language models (LLMs), which are now central to generating high-fidelity artificial records and augmenting qualitative datasets. Across these analytical, predictive, and pedagogical domains, synthetic data offers a viable response to persistent challenges related to data scarcity, privacy constraints, and limited data accessibility in education. The findings indicate a growing reliance on synthetic generation as an emerging methodological response to data-intensive demands. While synthetic data supports advanced modeling, adaptive learning systems, and instructional design, its epistemological legitimacy and methodological robustness remain contingent on rigorous validation practices. The study concludes that the field currently lacks standardized validation protocols, particularly regarding subgroup equity and fairness. Establishing transparent, equity-aware frameworks remains essential for the future integration of synthetic data into applied educational systems.

KEYWORDS

synthetic data, artificial intelligence (AI), educational research, systematic literature review (SLR), learning analytics, data privacy, data augmentation, generative AI

1 INTRODUCTION

In education, the use of data-driven approaches has expanded substantially over the past decade. The measurement of student performance and the optimization of instructional processes increasingly rely on large-scale educational datasets [1].

İpek, Z. H., Poyraz, E., Yigit, G. (2026). The Role of Synthetic Data in Educational Research: A Systematic Review. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(16), pp. 4–23. https://doi.org/10.3991/ijim.v20i16.62224

Article submitted 2026-05-04. Revision uploaded 2026-06-02. Final acceptance 2026-06-19.

As learning analytics and educational data mining become a central issue to institutional decision-making and pedagogical innovation, the demand for rich and multi-dimensional student data continues to prevail [2]. This growing reliance on data-intensive approaches has, in turn, heightened the importance of access to high-quality and comprehensive datasets.

Despite the necessity of data for educational research, access remains heavily restricted by ethical and institutional policies [3]. Sensitive sectors like healthcare and education often withhold datasets to prevent privacy violations [4], creating a tension between data protection and the data availability required for scientific innovation [5]. The rise of AI and machine learning has further intensified this conflict, as these technologies demand large-scale, complex data for effective assessment [6].

To address these constraints, synthetic data has emerged as a promising solution, replicating the statistical properties of original datasets while protecting privacy [7, 8]. Studies demonstrate that synthetic student records retain predictive utility with reduced disclosure risks [9, 10], positioning the technology as a key enabler for privacy-conscious learning analytics. Specifically, research using generative adversarial networks (GANs) and language models shows that synthetic data can preserve predictive performance and serve as shareable benchmarks for collaborative development when authentic records are unavailable [11, 12]. Consequently, validation should prioritize equity-focused diagnostics, such as subgroup parity, alongside traditional accuracy metrics.

Education data is governed by strict privacy frameworks, incentivizing the development of privacy-preserving release methods that must be evaluated against real-world learning analytics tasks [13]. However, diverse datasets remain scarce due to institutional silos and proprietary restrictions, which threatens the external validity and generalizability of educational research [14]. To overcome these limitations, synthetic data generated via AI and statistical modeling has emerged as a promising alternative. By replicating the relational patterns of authentic student records without exposing identifiable information, synthetic datasets facilitate robust research, benchmarking, and methodological experimentation [15].

The rapid expansion of interactive and mobile learning technologies introduces unique complexities to this data landscape, as these modern ecosystems continuously harvest highly granular, time-stamped, and location-aware user metrics [67]. Because mobile learning relies heavily on continuous streams of contextual data such as touch gestures, micro-interactions, device telemetry, and intermittent connectivity patterns, the privacy risks associated with traditional data sharing are exponentially magnified. In this context, synthetic data generation becomes vital, offering a secure pipeline to model intricate user-system dynamics without compromising sensitive contextual or behavioral markers [68]. By artificially manufacturing diverse, high-fidelity interaction logs, developers and researchers can train adaptive algorithms, rigorously stress-test mobile interfaces across varied hardware profiles, and simulate user pathways across fragmented digital environments. Consequently, the intersection of synthetic data and mobile learning not only bypasses strict regulatory bottlenecks but also provides the scalable infrastructure necessary to design next-generation, responsive educational frameworks [69].

To summarize, the literature suggests that synthetic student data are capable of maintaining meaningful predictive effectiveness for specific educational tasks when appropriate generation methods and frameworks are achieved [9, 10]. On the other hand, synthetic data generation also should be implemented with transparency to lower the risks of overreliance on synthetic outputs, especially regarding equity across student subgroups [6, 15]. If these methodological, ethical, and also validation standards are rigorously paid attention to and implemented, synthetic data generation becomes an effective mechanism for strengthening research capacity, enabling replication in learning analytics, advancing educational tool development, and enhancing educator training [9, 13].

As synthetic data applications diversify across the AI pipeline, the resulting fragmentation makes coherent insights difficult to achieve [16, 17]. Competing priorities of fairness, privacy, and fidelity further reinforce the need for a comprehensive synthesis [4, 18]. Systematic reviews serve as a critical tool in this evolving domain, consolidating findings and clarifying frameworks while offering a replicable basis for identifying research gaps [19, 22]. Adopting this rigorous approach is essential for establishing an analytically coherent understanding of the field.

This review is also relevant to research on interactive and mobile learning, as such systems are increasingly underpinned by learner-generated traces, contextual information, adaptive feedback mechanisms, and predictive modeling approaches. Within this framework, synthetic data offer significant methodological advantages by enabling privacy-preserving experimentation, strengthening learner modeling capabilities, supporting data augmentation strategies, and allowing the simulation of complex interaction patterns in situations where authentic learner data are inaccessible or restricted due to ethical and regulatory constraints.

To guide the synthesis of the existing literature and establish a clear analytical framework, the study was organized around several key areas of inquiry. These areas reflect both the growing use of synthetic data in education and the ongoing discussions about privacy, ethics, and methodological innovation. Structuring the review in this way helps to both present the current state of the field and identify its main themes, conceptual approaches, and emerging directions. Accordingly, the study was guided by the following research questions:

1. What major application areas of synthetic data are emphasized in the educational research literature (e.g., data sharing, analytics, assessment, policy analysis)?
2. Which thematic priorities and topical foci are most prominent in studies that explore synthetic data in educational contexts?
3. How are the existing studies conceptually framed, particularly regarding theoretical perspectives, ethical considerations, and technological assumptions underlying synthetic data use in education?
4. How does the integration of synthetic data in interactive and mobile learning frameworks alter or enhance the design, testing, and delivery of responsive, context-aware educational technologies?

2 RESEARCH METHOD

This study was designed as a systematic literature review (SLR) to provide a comprehensive and structured synthesis of the existing research on synthetic data in education. Systematic reviews follow a transparent and replicable process for identifying, selecting, and analyzing relevant studies, thereby reducing potential bias and increasing the reliability of findings. Compared to traditional narrative reviews, SLRs rely on clearly defined research questions, explicit inclusion and exclusion criteria, and documented screening procedures [23].

The review was conducted using the Web of Science (WoS) Core Collection databases to ensure broad coverage of high-quality, peer-reviewed publications across educational sciences and interdisciplinary research domains [24, 25]. WoS was selected due to its rigorous indexing standards, citation tracking capabilities, and multidisciplinary scope, which make it suitable for evidence synthesis studies in education and social sciences. In line with established guidelines for systematic reviews [26], the study followed a structured protocol to enhance transparency and replicability. This design enabled a consistent mapping of conceptual trends, methodological approaches, and emerging debates concerning the use of synthetic data in educational research.

2.1 Identification of sources

The identification of relevant studies was conducted in line with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework (see Figure 1) [23], ensuring a transparent and systematic selection process. The WoS database was selected as the primary data source because of its strong academic reputation and its frequent use in systematic review research [24, 25, 27]. Rather than limiting the search to a predefined set of high-impact journals, a comprehensive keyword-based search strategy was employed to broaden the scope of the review.

The search terms were developed through an iterative review of the conceptual and methodological literature on synthetic data generation and its applications in educational research. The initial keyword set comprised core terms including "synthetic data," "data synthesis," "data generation," and "artificial data." These records were subsequently screened and evaluated according to predefined criteria, leading to a final sample of 29 relevant peer-reviewed articles. The search was restricted to publications indexed in WoS between March 2000 and March 2026, thereby covering a 26-year period.

The search string was constructed using Boolean operators:

("synthetic data" OR "artificial data" OR "generated data" OR "simulated data" OR "data augmentation")
AND
("education" OR "educational research" OR "learning analytics" OR "educational data mining"
OR "mobile learning" OR "educational app")

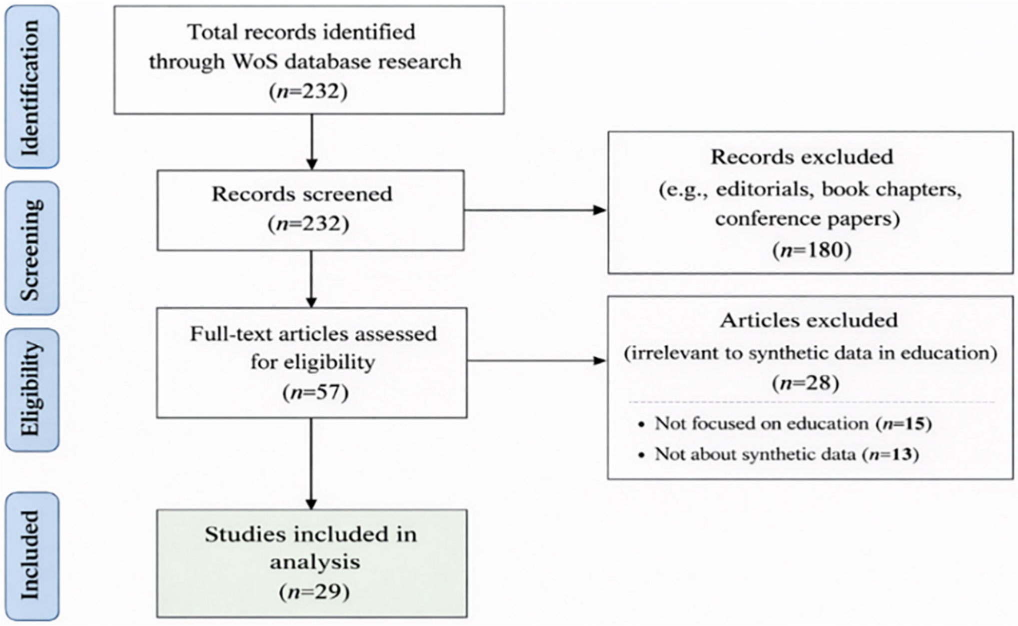

Fig. 1. PRISMA diagram shows the process of review and inclusion of resources (between the years 2000–2026)

2.2 Inclusion and exclusion criteria

The selection process followed a structured protocol aligned with PRISMA guidelines, applying the following specific criteria to the 232 initial records retrieved from the WoS Core Collection:

Inclusion Criteria:

Document Type: Only peer-reviewed academic articles were included to ensure the methodological robustness and scientific reliability of the findings.
Thematic Scope: Studies must explicitly engage with synthetic data generation, artificial data structures, or data augmentation specifically within educational research contexts.
Topical Relevance: Articles were required to address one or more of the study's core dimensions, such as privacy, predictive modeling, pedagogical design, or NLP applications in education.
Database Indexing: Only publications indexed within the WoS database were considered, utilizing its rigorous standards for multidisciplinary research.
Temporal Range: Studies must have been published between March 2000 and March 2026, covering a 26-year period of technological evolution.

Exclusion Criteria:

Publication Type: Non-research publications, including editorials, book reviews, and trade magazine articles, were excluded.
Focus Mismatch: Studies that discussed synthetic data or AI in general terms without a primary application or substantial focus on educational settings were removed.
Source Restriction: Research appearing in journals or conference proceedings not indexed in WoS was excluded from this specific review scope.
Methodological Insufficiency: Documents that lacked clear descriptions of their synthetic data generation techniques or validation logic were excluded during the qualitative content analysis stage.

The synthesis required broad generalizations across a diverse range of generation techniques, which may overlook the specific technical nuances of individual approaches like statistical modeling or generative architecture. As identified in the discussion, the findings are constrained by a broader field-wide limitation: the current absence of standardized, equity-aware validation protocols for synthetic datasets in educational contexts.

2.3 Data extraction and analysis

After applying the predefined inclusion and exclusion criteria, a total of 232 articles were retrieved from the WoS database in plain text format for further examination. The dataset included comprehensive bibliographic information such as authorship details, institutional affiliations, article titles, abstracts, and keywords. Following the retrieval process, the records were carefully reviewed and cleaned to ensure consistency and relevance. This stage involved standardizing terminology, merging synonymous expressions related to synthetic data, and excluding non-research publications that did not meet the scope of the review.

The thematic framework for this systematic review was constructed through an inductive and iterative coding process, utilizing qualitative content analysis to identify recurring patterns within the 29-article corpus. Initial codes were generated by scrutinizing the substantive content of each study across core dimensions, including primary research objectives, methodological architectures, educational application domains, and ethical considerations. While the primary assignment of studies to a single thematic category was operationalized through abstract-level thematic analysis based on their dominant research aims and methodological contributions, the final classification was cross-verified against the full texts of the articles to ensure structural integrity. To mitigate interpretive bias, a collaborative verification procedure was implemented where an independent second reviewer examined the full corpus and resolved any ambiguous classifications. The robustness of this process was further substantiated by an inter-coder reliability assessment using the Miles and Huberman (1994) formula, which yielded an 87% agreement rate, thereby confirming the internal consistency and academic rigor of the qualitative synthesis.

To enhance the credibility and reduce potential interpretive bias in the classification process, a second reviewer with expertise in educational technology and artificial intelligence (AI) independently examined the full corpus of 29 included articles. The second reviewer assessed the appropriateness of category assignments across the dimensions and raised concerns where classifications were ambiguous or insufficiently justified. This collaborative verification procedure strengthened the internal consistency of the classification framework and minimized the risk of idiosyncratic interpretation on the part of a single researcher [21]. To enhance the trustworthiness of the coding process, inter-coder reliability was assessed following the procedure suggested by Miles and Huberman (1994). Two researchers independently coded a subset of the data and compared their coding decisions. Reliability was calculated using the formula: Reliability = Agreements (87 codes)/(Agreements (87) + Disagreements (13)) × 100. The resulting agreement rate was 87%, exceeding the 80% threshold recommended by Miles and Huberman, indicating satisfactory coding consistency.

3 RESULTS

The initial findings of the study show the distribution of the articles. Within this distribution, 29 academic articles are examined in this study after the selection process based on the inclusion and exclusion criteria. The studies are presented in Table 1 according to subject category, article title, author and year, education field, and the journal in which they were published.

Table 1. Classification of 29 peer-reviewed academic articles

Subject Category	Article Title	Author and Year	Education Field	Journal
Privacy/Access Expansion	Ensuring privacy through synthetic data generation in education	Liu, Q. et al. (2025)	Educational Technology/ Educational Data Mining	*British Journal of Educational Technology*
	The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data	Bonnéry, D., et al. (2019)	Education Policy/ Higher Education	*Journal of Research on Educational Effectiveness*

(Continued)

Table 1. Classification of 29 peer-reviewed academic articles *(Continued)*

Subject Category	Article Title	Author and Year	Education Field	Journal
Data Augmentation and Imbalanced Data	Coping with unbalanced designs of generalizability theory: G String V	Teker, G. T. (2019)	Educational Measurement and Assessment	*International Journal of Assessment Tools in Education*
	Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning	Chachou., et al. (2024)	Computer Science Education/ Programming Education	*Computers and Education: AI*
	Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative AI	Martin, P. P., and Graulich, N. (2024)	Science Education (Chemistry Education)	*Computers and Education: AI*
	Data augmentation for sparse multidimensional learning performance data using generative AI	Zhang, L., et al. (2025)	Intelligent Tutoring Systems/ Learning Science	*IEEE Transactions on Learning Technologies*
	Educational data augmentation in physics education research using ChatGPT	Kieser, F., et al. (2023)	Physics Education Research	*Physical Review Physics Education Research*
NLP, Text Generation, and Qualitative Data	Leveraging AI-generated synthetic data to train natural language processing models for qualitative feedback analysis	Fuchs, S., et al. (2025)	Engineering Education/ STEM Education	*Journal of Engineering Education*
	Investigating evidence-oriented generation of synthetic text data with a generative large language model (LLM) in science education	Stanja, J., et al. (2025)	Science Education (Biology Education)	*International Journal of Science Education*
	Leveraging open-source LLMs for data augmentation in hospital staff surveys: a mixed-methods study	Ehrett, C., et al. (2024)	Health Professional Education/Medical Education	*JMIR Medical Education*
	Affordances and limitations of using LLMs to generate qualitative data about mental health perceptions in engineering	Sanders, J., et al. (2026)	Engineering Education	*Journal of Engineering Education*
	Leveraging LLM respondents for item evaluation: A psychometric analysis	Liu, Y. T., et al. (2025)	Educational Measurement/ Higher Education	*British Journal of Educational Technology*
	Automatic classification of Chinese programming MOOC reviews using fine-tuned BERTs and GPT-augmented data	Chen, X. L., et al. (2025)	MOOC/Online Learning/ Computer Science Education	*Educational Technology and Society*
Predictive Modeling and Learning Analytics	Predictive model to analyze real and synthetic data for learners' performance prediction using regression techniques	Shabnam, A. S. J., et al. (2025)	Online Learning/ Higher Education	*Online Learning*
	Advancing student outcome predictions through GANs	Farhood, H., et al. (2024)	Educational Technology/ AI in Education	*Computers and Education: AI*
	Integrating AI and learning analytics for data-driven pedagogical decisions and personalized interventions in education	Sajja, R., et al. (2025)	Higher Education/ AI in Education	*Technology, Knowledge and Learning*
	Temporal learner modeling through integration of neural and symbolic architectures	Hooshyar, D. (2024)	Intelligent Tutoring Systems/ Adaptive Learning	*Education and Information Technologies*
	Comparing harm done by mobility and class absence: Missing students and missing data	Dunn, M., et al. (2003)	K-12 Education/ Urban Education	*Journal of Educational and Behavioral Statistics*
	Enhanced predictive performance: A comparative analysis of ML and DL models using augmented LMS interaction data	Fahd, K., and Miah, S. J. (2025)	Higher Education/E-Learning/LMS	*Contemporary Educational Technology*
	Contrastive personalized exercise recommendation with reinforcement learning	Wu, S. Y., et al. (2024)	AI in Education/ Intelligent Tutoring and Recommendation Systems	*IEEE Transactions on Learning Technologies*

(Continued)

Table 1. Classification of 29 peer-reviewed academic articles *(Continued)*

Subject Category	Article Title	Author and Year	Education Field	Journal
Educational Applications and Pedagogical Design	Enriching data science and health care education: Application and impact of synthetic data sets through the Health Gym project	Nicolas et al. (2024)	Health Data Science Education/Medical Education	*JMIR Medical Education*
	An introduction to modeling through a microbial interaction application	Zama, F. (2024)	Mathematics and STEM Education (Higher Education)	*International Journal of Mathematical Education in Science and Technology*
Methodological Analysis and Systematic Review	Data fusion with international large-scale assessments: A case study using the OECD PISA and TALIS surveys	Kaplan, D., and McCarty, A. T. (2013)	Educational Assessment/ International Comparative Education	*Large-Scale Assessments in Education*
	Use of GANs in educational technology research	Bethencourt-Aguilar, A., et al. (2023)	Educational Technology/ Teacher Education	*Journal of New Approaches in Educational Research*
	Mapping the landscape of generative AI in learning analytics: A SLR	Misieju, K., et al., (2025)	Learning Analytics/ Educational Technology	*Journal of Learning Analytics*
Synthetic Data in Mobile, Interactive, and Adaptive Learning Systems	Capturing the process of students' AI interactions when creating and learning complex network structures	López-Pernas, S., et al. (2025)	Computer Science Education/ Network Science Education	*IEEE Transactions on Learning Technologies*
	AI-driven knowledge discovery: Developing a human-machine collaborative framework for learning Japanese sentence patterns	Liu, J. (2025)	Language Education/Second Language Acquisition (Japanese)	*Education and Information Technologies*
	The impact of using artificial data in undergraduate statistics students' projects due to COVID-19 lockdowns	Dunn, P. K. (2024)	Statistics Education/ Higher Education	*International Journal of Mathematical Education in Science and Technology*
	Machine learning's model-agnostic interpretability on the prediction of students' academic performance in video-conference-assisted online learning during the COVID-19 pandemic	Miranda, E., et al., (2024)	Online Learning/Higher Education (COVID-19 Context)	*Computers and Education: AI*

Table 1 presents a systematic thematic classification of 29 peer-reviewed academic articles retrieved from the WoS database, all of which engage with synthetic data generation and its applications within educational research contexts. The 29 articles are grouped under seven thematically defined subject categories – (1) Privacy and Data Sharing, (2) Data Augmentation and Imbalanced Data, (3) NLP, Text Generation, and Qualitative Data Production, (4) Predictive Modelling and Learning Analytics, (5) Educational Applications and Pedagogical Design, (6) Methodological Analysis and Systematic Review and (7) Synthetic Data in Mobile, Interactive, and Adaptive Learning Systems. Each article reflects a distinct research orientation within the broader synthetic data literature. Classification was determined through abstract-level thematic analysis with primary category assignment based on each study's dominant research aim and methodological contribution. The corpus spans a wide range of educational fields including engineering education, health data science education, physics education, language learning, computer science education, and K-12 urban education and draws from 22 distinct peer-reviewed journals by reflecting the interdisciplinary nature of synthetic data research in education. Publication years range from 2003 to 2026, though the overwhelming majority of articles (n = 24) were published between 2023 and 2026.

Finally, the findings obtained within the scope of the research are divided into 7 dimensions. The use of synthetic data has been used for different purposes in the literature. These purposes have been evaluated under seven main dimensions.

3.1 Privacy and data sharing

The Privacy and Access Expansion dimension addresses the critical tension between the escalating demand for high-quality educational datasets and the strict legal and security constraints surrounding data disclosure. The primary objective of the research in this domain is to resolve the inability to share real-world datasets by developing secure, synthetic alternatives that statistically mimic relational variables without representing actual individuals. In studies of this dimension, Liu et al. [28] highlight that traditional anonymization techniques often fail to meet privacy standards; henceforth, synthetic data can be a good option for preserving privacy. Bonnéry et al. [29] approach the issue from a policy perspective, and provides a practical template for administrative agencies to bypass the legal and security barriers that currently restrict access to sensitive data for research and policy evaluation. Together, these studies demonstrate that synthetic data generation serves as a robust strategic framework for ensuring data security while facilitating the broad dissemination of information across institutional levels.

3.2 Data augmentation and imbalanced data

The dimension of Data Augmentation and Imbalanced data focuses on technical methodologies designed to overcome the limitations of sparse, skewed, or insufficient datasets in educational research. The primary objective within this domain is to enhance the statistical reliability and predictive accuracy of models by synthetically addressing data irregularities rather than relying on traditional, resource-intensive collection methods. The literature in this dimension approaches the data enrichment from different perspectives. For example, Teker [30] uses the G String V software to utilize synthetic data for balancing designs with generalizability theory. On the other hand, Martin and Graulich [31] evaluate the efficacy of using chatbots to paraphrase existing student responses or generate entirely new synthetic data. The findings of studies conducted within this domain suggest that transitioning from the constraints associated with real-world data toward the adoption of generative and algorithmic tools represents a promising approach to constructing more effective and educationally oriented models.

3.3 NLP, text generation, and qualitative data production

Classroom transcripts and written student feedback are among the most informative data sources in educational research and among the hardest to reuse, precisely because they are full of identifying detail. LLMs have become a practical workaround for this. Fuchs et al. [45] found that synthetic feedback transcripts produced by locally hosted models pushed NLP classification accuracy to around 81%, a meaningful result for contexts where real student text is simply unavailable. Stanja et al. [32] take a different route through evidence-oriented prompting, generating textual data from educational frameworks rather than from any student input at all, which solves the privacy problem but introduces another: text built from theoretical constructs does not necessarily reflect how students actually write or speak. Both approaches are useful, and both carry limits that the field has not fully worked through. The deeper issue, as Misiejuk et al. [33] point out, is that there are no standardized frameworks for evaluating whether synthetically generated text is actually good enough to use. In quantitative work, a poorly fitted distribution is a technical problem. In qualitative research, getting student discourse wrong affects what researchers conclude about learning. That distinction matters, and it means transparency about how synthetic text was generated and validated is not a formality but a basic requirement of methodological credibility.

3.4 Predictive modelling and learning analytics

The primary objective of Predictive Modelling and Learning Analytics dimension is to overcome the constraints of data scarcity and privacy by expanding synthetic and hybrid datasets to model learning for predicting students' learning performances and improve analytical precision. Regarding this dimension, Shabnam et al. [34], for instance, provide an empirical comparison between real, synthetic, and hybrid datasets using a suite of regression-based machine learning algorithms. On the other hand, Sajja et al. [35] emphasize the role of integrating AI with learning analytics to facilitate data driven pedagogies. With these predictive models, educators can implement personalized implementations that are specifically utilized to predict the individual learners' performances. Studies in this dimension highlight a significant shift toward using Generative AI as a cornerstone for measurable, privacy-preserving educational analytics that do not sacrifice predictive utility.

3.5 Educational applications and pedagogical design

Educational Applications and Pedagogical Design dimension mainly focuses on integrating synthetic data into instructional content, applied student projects, and curriculum evaluation. In this dimension, maintaining pedagogical continuity and enhancing domain-specific knowledge construction is emphasized especially when traditional data collection is hindered by external constraints or resource limitations. For example, Dunn [36] investigates the projects of students enrolled at introductory statistics courses, including data of students and AI generated data after the COVID lockdown. The study found that while artificial data was less likely to contain outliers or necessitate complex analyses like t-tests, it allowed students to continue practicing core research competencies. For another point, Liu [37] aimed to develop a model to assist learners of Japanese as a second language since they found some sentence patterns difficult. For this reason, the researcher developed a human-machine framework combining AI techniques to create large number of samples since it is challenging, time consuming, and expensive to collect human annotated samples. Although there are some differences among the human and AI generated samples, this experiment demonstrated how synthetic data can facilitate the acquisition of complex patterns in Japanese. Therefore, synthetic data is not merely a technical solution to the problems but can be used as a versatile pedagogical tool.

3.6 Methodological analysis and systematic review

The Methodological Analysis and Systematic Review dimension focuses on the rigorous evaluation of synthetic data generation techniques and their validity within educational research frameworks. The primary objective is to establish standardized methodologies for data fusion, verify the equivalence between artificial and authentic datasets, and provide a critical view of the transformation of generative AI in learning analytics. In this respect, Kaplan and McCarty [38] examine the challenge of missing information in international exams such as PISA and TALIS and they propose synthetic data to merge disparate data sets gathered from these international exams into a single, usable file. This methodological advancement allows for a more holistic analysis of educational systems, diminishing missing information. On the other hand, Bethencourt-Aguilar et al. [39] investigate the reliability of GANs by comparing synthetic samples against real data concerning university teachers' self-perceptions on technological pedagogical content knowledge (TPACK). The study confirms that synthetic data can produce qualitatively similar results to real datasets, identifying consistent teaching profiles across 150 tests. This validates that synthetic data can provide statistical integrity and relational structures of the original sources and provide evidence for the adoption of synthetic methodologies in systematic reviews.

3.7 Synthetic data in mobile, interactive, and adaptive learning systems

The "Synthetic Data in Mobile, Interactive, and Adaptive Learning Systems" dimension highlights how artificial and advanced data techniques are increasingly utilized to predict student outcomes and optimize human-machine collaboration within intelligent educational ecosystems. A major focus within this domain is leveraging data analytics and machine learning to build transparent, highly responsive predictive models that power adaptive systems. For instance, Miranda et al. [62] deployed machine learning models on student performance data in video-conference-assisted online environments, emphasizing that model-agnostic interpretability is crucial for generating transparent, actionable insights into student success. When combined with collaborative frameworks like the one developed by Liu [37]—which uses an AI-driven knowledge discovery model to optimize language acquisition—these findings underscore the growing potential of using advanced data frameworks. Together, they demonstrate how artificial data structures and intelligent modeling can move educational technology past static content delivery toward highly responsive, privacy-preserving, and adaptive learning environments.

Within this theme, four focal studies—including Miranda et al. (2024) on interpretable machine learning in online environments and Liu (2025) on human-machine language acquisition frameworks—directly affect mobile and interactive learning by embedding active, intelligent data structures into student-facing systems. Conversely, the remaining 25 studies indirectly support these environments by establishing the essential, privacy-preserving infrastructure behind the scenes. By focusing on the generative robustness, validation, and statistical privacy of synthetic samples, this larger body of indirect research provides the foundational data pipelines required to safely train and scale adaptive learning systems.

4 DISCUSSION

The findings of this study indicate a clear shift in educational data practices from traditional anonymization approaches toward synthetic data generation as a response to persistent data privacy constraints. This shift is increasingly evident in data-driven educational research, where privacy concerns limit both the availability and analytical depth of real datasets. As Liu et al. [40] highlight, conventional anonymization techniques are often insufficient for supporting advanced educational research, as they reduce data utility while still failing to fully resolve privacy risks. In contrast, synthetic data enables the generation of artificial but statistically consistent datasets that preserve analytical value without exposing identifiable individual information [11, 28].

Building on the implications of this shift, the findings further emphasize the persistent tension between data privacy and research utility in educational sciences [11, 41]. To address this tension, distinct strategies, particularly mirroring and augmentation approaches, have been developed. As discussed earlier, the synthetic data produced with this approach matches the real data by approximately 85 percent [34]. Considering these findings, it can be said that synthetic data can be used as an alternative tool with different mirroring strategies such as digital twins compared to real data [41]. Therefore, the mirroring approach can generate synthetic data that can be used as an alternative to real data while protecting privacy.

The findings also demonstrate a complementary pattern in research on the augmentation approach. In contrast to mirroring strategies, augmentation extends synthetic data generation by addressing missing or incomplete information while producing more heterogeneous datasets [42]. This increased diversity enhances the analytical depth of datasets and improves their potential for generating more generalizable educational insights [31]. From a methodological perspective, this suggests that augmentation does not merely reconstruct existing data structures but actively enriches them by introducing variability that is often underrepresented in real-world educational datasets. Such variability is particularly valuable in contexts characterized by small, imbalanced, or fragmented data, where traditional analytical approaches may struggle to capture meaningful patterns [43, 44].

Extending this discussion to qualitative data contexts, particularly text-based data, recent studies highlight the role of LLMs in addressing data scarcity and privacy concerns. In educational research, synthetic data generation offers a promising solution for augmenting small datasets in text classification tasks, particularly in contexts where real data for NLP training is limited or difficult to obtain [45]. For instance, open-source LLMs have been used for data augmentation to enhance the performance of text classification [46]. Similarly, the use of synthetic feedback transcripts generated by locally hosted models has been shown to improve NLP model performance, increasing classification accuracy up to 81% [45]. These findings collectively suggest that within qualitative educational data contexts, synthetic data approaches are not only used for data augmentation but are also evolving toward more structured and theory-informed text generation strategies. Instead of using student data, based on prompts, educational theories can be created with high quality datasets, including domain specific concepts [33]. However, existing weaknesses in generative models persist. For instance, Fuchs et al. [45] found that synthetic transcripts may include extraneous details or unable to capture the specific nuances of instructor-dominant discourse.

Synthetic data has also become a key enabler of predictive modelling in educational research under conditions of data scarcity and privacy constraints. Findings demonstrate that synthetically generated datasets can approximate the statistical properties of authentic educational records with sufficient fidelity to support regression-based and deep learning models for student performance prediction [11, 34, 47]. This addresses a long-standing structural limitation in learning analytics related to restricted access to large-scale, representative, and ethically shareable datasets, positioning synthetic data as a methodological response to systemic data governance constraints rather than a technical convenience [10, 13]. Building on this foundation, recent studies extend the application of synthetic data by integrating it into adaptive and dynamic learning analytics frameworks. Beyond simple dataset substitution, synthetic data is embedded within real-time and temporally sensitive analytical systems to enable personalized educational decision-making [35, 48]. This development aligns with the broader shift in learning analytics toward continuous, individualized feedback and intervention mechanisms rather than retrospective, population-level analysis [1, 2, 8]. Accordingly, the predictive utility of synthetic data should be understood as an outcome of carefully designed generation and validation pipelines, with increasing evidence supporting its integration into adaptive educational infrastructures [49].

This evolution toward adaptive educational infrastructures is increasingly manifested through mobile learning architectures, where synthetic data serves as a critical asset for optimizing interactive user experiences. In mobile-centric learning analytics, real-time data scarcity and latency often hinder the responsiveness and pedagogical value of interactive mobile interfaces [63, 71]. By embedding synthetic data generation pipelines directly within mobile application frameworks, developers and researchers can simulate high-fidelity user interactions and complex mobile device telemetry. This enables the rigorous testing, evaluation, and refinement of context-aware, interactive mobile learning tools without compromising student privacy or straining limited device network bandwidth [64, 70]. Consequently, as emphasized in recent frameworks for adaptive mobile educational applications, the utility of synthetic data extends beyond static backend analysis, directly shaping the deployment of highly responsive, interactive mobile technologies that provide individualized, immediate scaffolding to learners on the move [63, 65, 72].

While the predictive modeling literature treats synthetic data mainly as analytical infrastructure, a distinct strand of the corpus conceptualizes it as an active pedagogical resource. These studies challenge the assumption that artificially generated data is a methodological compromise, something adopted only when real data are unavailable. Synthetic learning environments carry instructional affordances that are worth designing around deliberately. Dunn [36] provides empirical support for this argument by demonstrating that, during COVID-19 disruptions that limited students' access to real fieldwork data, learners who engaged with synthetically generated datasets were still able to develop core statistical reasoning competencies. The findings suggest that the pedagogical value of such environments lies less in the epistemic authenticity of the data source and more in the depth and quality of the analytical engagement they facilitate. This resonates with constructivist learning theory, which positions active knowledge construction through engagement with structured problems rather than the nature of the material itself as the primary driver of conceptual understanding [2].

Beyond its function as analytical infrastructure in predictive modeling, the findings suggest that synthetic data can also be understood as a pedagogically meaningful construct that actively shapes learning processes. Liu [37] similarly shows that synthetically produced sentence samples can scaffold complex grammatical acquisition in second language learning, a finding that extends the pedagogical relevance of synthetic data well beyond quantitative disciplines and into areas where fine-grained linguistic modeling is required. Purpose-built synthetic environments have also demonstrated capacity for domain-specific skill development in health data science and STEM education [50, 51]. In both contexts, the controllable and designable character of synthetic environments appears to offer a kind of pedagogical precision that naturalistic datasets cannot reliably provide. This aligns with emerging scholarship on the instructional affordances of AI-generated content in applied educational settings [52], and it invites a reorientation in which synthetic data is no longer seen as a substitute for authentic experience but as a distinct mode of knowledge engagement with advantages of its own.

As the application of synthetic data expands across both analytical and pedagogical domains, increasing attention is directed toward its methodological validation and epistemological legitimacy as a central area of inquiry. Across these thematic strands, a growing body of work moves beyond instrumental use and instead critically examines the methodological foundations, validation logic, and disciplinary boundaries of synthetic data. This line of research is epistemologically significant, as it addresses the conditions under which synthetically generated datasets can be considered valid substitutes for authentic data sources, with direct implications for the credibility and robustness of findings across the field. Kaplan and McCarty [38] provide an early and influential contribution by demonstrating that synthetic imputation can serve as a structural integrator across incompatible large-scale assessment instruments, thereby enabling cross-survey analyses that conventional data linkage approaches have difficulty achieving. This suggests that synthetic data should not be understood solely as a mechanism for replicating individual datasets but also as a tool for constructing analytical bridges across institutionally and methodologically heterogeneous data systems, such as PISA and TALIS. Extending this validation-oriented perspective, Bethencourt-Aguilar et al. [39] provide empirical evidence that GAN-generated educational profiles can achieve distributional equivalence with real data across repeated simulation cycles. These findings offer concrete support for claims regarding the statistical fidelity of synthetic data in educational research and reinforce the growing emphasis on rigorous pre-use validation frameworks in the literature [4, 7]. Despite advances in validation studies, the literature consistently identifies a critical unresolved issue that continues to shape methodological debates in the field. Misiejuk et al. [33] note that although generative AI models in learning analytics have become increasingly sophisticated, the corresponding validation frameworks have not developed at a comparable pace. This gap is particularly salient in relation to subgroup equity, as distortions in the distributional properties of marginalized student populations may lead predictive models and pedagogical applications to reproduce or even amplify existing inequalities. Such concerns extend foundational discussions on the epistemic status of synthetic datasets and highlight the need for evaluation protocols that move beyond statistical fidelity to include equity-sensitive validity across diverse educational populations [4, 7]. However, the absence of standardized validation protocols addressing these dimensions remains one of the most significant limitations in the field.

Overall, the findings of this study indicate that synthetic data has evolved from a niche methodological solution for privacy protection into a central component of contemporary educational data science. Across analytical, predictive, and pedagogical domains, it offers a viable response to persistent challenges related to data scarcity, privacy constraints, and limited data accessibility in education. The synthesis of the literature suggests that while synthetic data supports advanced modeling, adaptive learning systems, and instructional design, its epistemological legitimacy and methodological robustness remain contingent on rigorous validation practices [52]. In particular, concerns regarding statistical fidelity and equity-sensitive representativeness continue to shape its responsible use in educational research [4, 7]. At the same time, the increasing integration of synthetic data into learning analytics and educational decision-making systems highlights its growing role as both an analytical infrastructure and a pedagogical resource. Taken together, these findings underscore that the future development and broader adoption of synthetic data in education will depend on the establishment of standardized, transparent, and equity-aware validation frameworks that ensure both methodological rigor and ethical integrity in applied contexts.

5 LIMITATIONS

This study provides an important contribution to the growing research on synthetic data generation in education by systematically reviewing a wide range of studies indexed in the WoS database. The findings offer a broad overview of the field's development and highlight key methodological and conceptual patterns. Nevertheless, as with all systematic reviews, several limitations should be acknowledged. First, although careful attention was paid to the screening, coding, and categorization processes, minor errors may still remain due to the complexity of the topic and the large number of documents analyzed. Ambiguous descriptions or incomplete information in the original studies may also have influenced some classifications. Second, the review includes only studies published in journals indexed in WoS. As a result, relevant research appearing in other well-regarded journals, particularly those focusing on AI, data science, or educational technology, could not be included. This is a common limitation of database-based reviews, but it nonetheless narrows the scope of the study. Finally, some generalizations were unavoidable when synthesizing diverse approaches to synthetic data generation—from statistical methods to GAN-based and LLM-driven techniques. Future research may benefit from narrowing the focus to a single methodological category or a specific educational context to allow for deeper and more fine-grained analysis.

6 REFERENCES

[1] P. Long and G. Siemens, "Penetrating the fog: Analytics in learning and education," *Italian Journal of Educational Technology*, vol. 22, no. 3, pp. 132–137, 2014.

[2] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning Analytics: From Research to Practice*, J. A. Larusson and B. White, Eds., 2014, pp. 61–75. https://doi.org/10.1007/978-1-4614-3305-7_4

[3] A. Hansen and D. Machin, *Media and Communication Research Methods*. London: Bloomsbury Publishing, 2018.

[4] S. M. Bellovin, P. K. Dutta, and N. Reitinger, "Privacy and synthetic datasets," *Stanford Technology Law Review*, vol. 22, no. 1, pp. 1–52, 2018. https://law.stanford.edu/wp-content/uploads/2019/01/Bellovin_20190129.pdf

[5] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare," *Nature Biomedical Engineering*, vol. 5, pp. 493–497, 2021. https://doi.org/10.1038/s41551-021-00751-8

[6] A. Koul, D. Duran, and T. Hernandez-Boussard, "Synthetic data, synthetic trust: Navigating data challenges in the digital revolution," *The Lancet Digital Health*, vol. 7, no. 11, p. 100924, 2025. https://doi.org/10.1016/j.landig.2025.100924

[7] J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. New York, NY: Springer Science & Business Media, 2011. https://doi.org/10.1007/978-1-4614-0326-5

[8] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, and G. Epelde, "Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing," *JMIR Medical Informatics*, vol. 8, no. 7, p. e18910, 2020. https://doi.org/10.2196/18910

[9] D. Iloh, O. T. Olayinka, F. Iyere, and S. Chilakala, "Generative private synthetic student data for learning analytics: An empirical study," *IEEE Access*, vol. 13, pp. 177345–177351, 2025. https://doi.org/10.1109/ACCESS.2025.3619091

[10] C. Zhan, O. B. Deho, X. Zhang, S. Joksimovic, and M. de Laat, "Synthetic data generator for student data serving learning analytics: A comparative study," *Learning Letters*, vol. 1, p. 5, 2023. https://doi.org/10.59453/KHZW9006

[11] H. Farhood, I. Joudah, A. Beheshti, and S. Muller, "Advancing student outcome predictions through generative adversarial networks," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100293, 2024. https://doi.org/10.1016/j.caeai.2024.100293

[12] M. Khalil, F. Vadiee, R. Shakya, and Q. Liu, "Creating artificial students that never existed: Leveraging large language models and CTGANs for synthetic data generation," in *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 2025, pp. 439–450. https://doi.org/10.1145/3706468.3706523

[13] M. E. Gursoy, A. Inan, M. E. Nergiz, and Y. Saygin, "Privacy-preserving learning analytics: Challenges and techniques," *IEEE Transactions on Learning Technologies*, vol. 10, no. 1, pp. 68–81, 2016. https://doi.org/10.1109/TLT.2016.2607747

[14] B. Williamson, *Big Data in Education: The Digital Future of Learning, Policy and Practice*. London: SAGE Publications, 2017. https://doi.org/10.4135/9781529714920

[15] A. Kieser, S. Müller, and P. Weber, "Synthetic datasets for reproducibility in educational research: A methodological framework," *Educational Research Review*, vol. 39, p. 100521, 2023. https://doi.org/10.1016/j.edurev.2023.100521

[16] S. Kapania, S. Ballard, A. Kessler, and J. W. Vaughan, "Examining the expanding role of synthetic data throughout the AI development pipeline," in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2025, pp. 45–60. https://doi.org/10.1145/3715275.3732005

[17] S. I. Nikolenko, *Synthetic Data for Deep Learning*. Cham: Springer, 2021. https://doi.org/10.1007/978-3-030-75178-4

[18] K. Kesgin, "FairSYN-Edu a diffusion-based model for fair and private educational data synthesis," *Discovery Education*, vol. 4, pp. 336–345, 2025. https://doi.org/10.1007/s44217-025-00743-9

[19] B. Kitchenham, "Procedures for performing systematic reviews," Keele, UK, Keele University, vol. 33, pp. 1–26, 2004.

[20] M. Y. Mustafa *et al.*, "A systematic review of literature reviews on artificial intelligence in education (AIED): A roadmap to a future research agenda," *Smart Learning Environments*, vol. 11, pp. 59–83, 2023. https://doi.org/10.1186/s40561-024-00350-5

[21] M. Petticrew and H. Roberts, *Systematic Reviews in the Social Sciences: A Practical Guide*. Chichester, U.K.: John Wiley & Sons, 2008.

[22] H. Snyder, "Literature review as a research methodology: An overview and guidelines," *Journal of Business Research*, vol. 104, pp. 333–339, 2019. https://doi.org/10.1016/j.jbusres.2019.07.039

[23] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Medicine*, vol. 6, no. 7, p. e1000097, 2009. https://doi.org/10.1371/journal.pmed.1000097

[24] M. Esen, M. S. Bellibas, and S. Gumus, "The evolution of leadership research in higher education for two decades (1995–2014): A bibliometric and content analysis," *International Journal of Leadership in Education*, vol. 23, no. 3, pp. 259–273, 2020. https://doi.org/10.1080/13603124.2018.1508753

[25] A. Kuzhabekova, D. D. Hendel, and D. W. Chapman, "Mapping global research on international higher education," *Research in Higher Education*, vol. 56, no. 8, pp. 861–882, 2015. https://doi.org/10.1007/s11162-015-9371-1

[26] D. Tranfield, D. Denyer, and P. Smart, "Towards a methodology for developing evidence-informed management knowledge by means of systematic review," *British Journal of Management*, vol. 14, no. 3, pp. 207–222, 2003. https://doi.org/10.1111/1467-8551.00375

[27] D. Abdullah, M. I. A. Aziz, and A. L. M. Ibrahim, "A 'research' into international student-related research: (Re)visualising our stand?" *Higher Education*, vol. 67, no. 3, pp. 235–253, 2014. https://doi.org/10.1080/13603124.2018.1508753

[28] Q. Liu, R. Shakya, J. Jovanovic, M. Khalil, and J. de la Hoz-Ruiz, "Ensuring privacy through synthetic data generation in education," *British Journal of Educational Technology*, vol. 56, no. 3, pp. 1053–1073, 2025. https://doi.org/10.1111/bjet.13576

[29] D. Bonnéry *et al.*, "The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data," *Journal of Research on Educational Effectiveness*, vol. 12, no. 4, pp. 616–647, 2019. https://doi.org/10.1080/19345747.2019.1631421

[30] G. T. Teker, "Coping with unbalanced designs of generalizability theory: G String V," *International Journal of Assessment Tools in Education*, vol. 6, no. 5, pp. 57–69, 2019. https://doi.org/10.21449/ijate.658747

[31] P. P. Martin and N. Graulich, "Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100265, 2024. https://doi.org/10.1016/j.caeai.2024.100265

[32] J. Stanja, S. Dannemann, J. Krugel, and A. Hoppe, "Investigating evidence-oriented generation of synthetic text data with a generative large language model in science education," *International Journal of Science Education*, pp. 1–23, 2025. https://doi.org/10.1080/09500693.2025.2538834

[33] K. Misiejuk, S. López-Pernas, R. Kaliisa, and M. Saqr, "Mapping the landscape of generative artificial intelligence in learning analytics: A systematic literature review," *Journal of Learning Analytics*, vol. 12, no. 1, pp. 12–31, 2025. https://doi.org/10.18608/jla.2025.8591

[34] A. S. Shabnam, T. Ramachandriah, and M. S. Haladappa, "Predictive model to analyze real and synthetic data for learners' performance prediction using regression techniques," *Online Learning*, vol. 29, no. 1, 2025. https://doi.org/10.24059/olj.v29i1.4390

[35] R. Sajja, Y. Sermet, D. Cwiertny, and I. Demir, "Integrating AI and learning analytics for data-driven pedagogical decisions and personalized interventions in education," *Technology, Knowledge and Learning*, 2025. https://doi.org/10.1007/s10758-025-09897-9

[36] P. K. Dunn, "The impact of using artificial data in undergraduate statistics students' projects due to COVID-19 lockdowns," *International Journal of Mathematical Education in Science and Technology*, vol. 55, no. 8, pp. 1759–1768, 2024. https://doi.org/10.1080/0020739X.2022.2056095

[37] J. Liu, "AI-driven knowledge discovery: Developing a human–machine collaborative framework for learning Japanese sentence patterns," *Education and Information Technologies*, vol. 30, no. 9, pp. 12413–12446, 2025. https://doi.org/10.1007/s10639-024-13267-w

[38] D. Kaplan and A. T. McCarty, "Data fusion with international large scale assessments: A case study using the OECD PISA and TALIS surveys," *Large-Scale Assessments in Education*, vol. 1, no. 1, p. 6, 2013. https://doi.org/10.1186/2196-0739-1-6

[39] A. Bethencourt-Aguilar, D. Castellanos-Nieves, J. J. Sosa-Alonso, and M. Area-Moreira, "Use of generative adversarial networks (GANs) in educational technology research," *Journal of New Approaches in Educational Research*, vol. 12, no. 1, pp. 153–170, 2023. https://doi.org/10.7821/naer.2023.1.1231

[40] Y. Liu, S. Bhandari, and Z. A. Pardos, "Leveraging LLM respondents for item evaluation: A psychometric analysis," *British Journal of Educational Technology*, vol. 56, no. 3, pp. 1028–1052, 2025. https://doi.org/10.1111/bjet.13570

[41] F. S. Bäumer, S. Schultenkämper, M. Geierhos, and Y. S. Lee, "Mirroring privacy risks with digital twins: When pieces of personal data suddenly fit together," *SN Computer Science*, vol. 5, no. 8, 2024. https://doi.org/10.1007/s42979-024-03413-z

[42] T. Brothers, K. Adhikari, M. Ramazani, A. Imtiaz, and M. Al-Mamun, "Synthetic data-augmented machine learning for 30-day readmission prediction in patients with chronic conditions: A retrospective real-world study," *BMJ Open*, vol. 16, no. 4, p. e108273, 2026. https://doi.org/10.1136/bmjopen-2025-108273

[43] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023. https://doi.org/10.1016/j.lindif.2023.102274

[44] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, 2020. https://doi.org/10.1007/s11023-020-09548-1

[45] S. Fuchs, A. Werth, C. Méndez, and J. Butcher, "Leveraging AI-generated synthetic data to train natural language processing models for qualitative feedback analysis," *Journal of Engineering Education*, vol. 114, no. 4, p. e70033, 2025. https://doi.org/10.1002/jee.70033

[46] C. Ehrett, S. Hegde, K. Andre, D. Liu, and T. Wilson, "Leveraging open-source large language models for data augmentation in hospital staff surveys: Mixed methods study," *JMIR Medical Education*, vol. 10, p. e51433, 2024. https://doi.org/10.2196/51433

[47] K. Fahd and S. J. Miah, "Enhanced predictive performance: A comparative analysis of ML and DL models using on augmented LMS interaction data," *Contemporary Educational Technology*, vol. 17, no. 4, p. ep606, 2025. https://doi.org/10.30935/cedtech/17453

[48] D. Hooshyar, "Temporal learner modelling through integration of neural and symbolic architectures," *Education and Information Technologies*, vol. 29, no. 1, pp. 1119–1146, 2024. https://doi.org/10.1007/s10639-023-12334-y

[49] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[50] I. Nicholas *et al.*, "Enriching data science and health care education: Application and impact of synthetic data sets through the health gym project," *JMIR Medical Education*, vol. 10, no. 1, p. e51388, 2024. https://doi.org/10.2196/51388

[51] F. Zama, "An introduction to modelling through a microbial interaction application," *International Journal of Mathematical Education in Science and Technology*, vol. 55, no. 2, pp. 340–351, 2024. https://doi.org/10.1080/0020739X.2023.2249465

[52] D. Surendran, V. Arulkumar, M. Aruna, K. Sangamithrai, and N. Thangadurai, "Improving the quality of education through data analytics and big data contributions," *AIP Conference Proceedings*, vol. 2742, no. 1, p. 020001, 2024. https://doi.org/10.1063/5.0183570

[53] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, 2020. https://doi.org/10.1002/widm.1355

[54] Y. Chachoui, N. Azizi, R. Hotte, and T. Bensebaa, "Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100222, 2024. https://doi.org/10.1016/j.caeai.2024.100222

[55] L. Zhang *et al.*, "Data augmentation for sparse multidimensional learning performance data using generative AI," *IEEE Transactions on Learning Technologies*, vol. 18, pp. 145–164, 2025. https://doi.org/10.1109/TLT.2025.3526582

[56] F. Kieser, P. Wulff, J. Kuhn, and S. Küchemann, "Educational data augmentation in physics education research using ChatGPT," *Physical Review Physics Education Research*, vol. 19, no. 2, p. 020150, 2023. https://doi.org/10.1103/PhysRevPhysEducRes.19.020150

[57] S. López-Pernas, K. Misiejuk, R. Kaliisa, and M. Saqr, "Capturing the process of students' AI interactions when creating and learning complex network structures," *IEEE Transactions on Learning Technologies*, vol. 18, pp. 556–568, 2025. https://doi.org/10.1109/TLT.2025.3568599

[58] J. Sanders, J. Mobley IV, I. Miller, N. W. Sochacka, P. A. Jensen, and K. J. Jensen, "Affordances and limitations of using large language models to generate qualitative data about mental health perceptions in engineering," *Journal of Engineering Education*, vol. 115, no. 1, p. e70037, 2026. https://doi.org/10.1002/jee.70037

[59] X. Chen, H. Xie, D. Zou, L. Xu, and F. L. Wang, "Automatic classification of Chinese programming MOOC reviews using fine-tuned BERTs and GPT-augmented data," *Educational Technology & Society*, vol. 28, no. 1, pp. 230–249, 2025.

[60] M. C. Dunn, J. B. Kadane, and J. R. Garrow, "Comparing harm done by mobility and class absence: Missing students and missing data," *Journal of Educational and Behavioral Statistics*, vol. 28, no. 3, pp. 269–288, 2003. https://doi.org/10.3102/10769986028003269

[61] S. Wu, J. Wang, and W. Zhang, "Contrastive personalized exercise recommendation with reinforcement learning," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 691–703, 2003. https://doi.org/10.1109/TLT.2023.3326449

[62] E. Miranda, M. Aryuni, M. I. Rahmawati, S. E. Hiererra, and A. V. D. Sano, "Machine learning's model-agnostic interpretability on the prediction of students' academic performance in video-conference-assisted online learning during the COVID-19 pandemic," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100312, 2024. https://doi.org/10.1016/j.caeai.2024.100312

[63] S. Papadakis and N. Kalogiannakis, "Mobile educational applications for children: What educators and parents need to know," *International Journal of Mobile Learning and Organisation*, vol. 11, no. 3, pp. 256–277, 2017. https://doi.org/10.1504/IJMLO.2017.085338

[64] S. Papadakis, "Evaluating pre-service teachers' acceptance of mobile devices with regards to their age and gender: A case study in Greece," *International Journal of Mobile Learning and Organisation*, vol. 12, no. 4, pp. 336–352, 2018. https://doi.org/10.1504/IJMLO.2018.095130

[65] A. I. Zourmpakis, M. Kalogiannakis, and S. Papadakis, "Adaptive gamification in science education: An analysis of the impact of implementation and adapted game elements on students' motivation," *Computers*, vol. 12, no. 7, p. 143, 2023. https://doi.org/10.3390/computers12070143

[66] M. B. Miles and A. M. Huberman, *Qualitative Data Analysis: An Expanded Sourcebook*, 2nd Ed. Thousand Oaks, CA: Sage Publications, 1994.

[67] M. K. Savi *et al.*, "A standardised differential privacy framework for epidemiological modeling with mobile phone data," *PLOS Digital Health*, vol. 2, no. 10, p. e0000233, 2023. https://doi.org/10.1371/journal.pdig.0000233

[68] G. Min and J. Oh, "Can synthetic data protect privacy?" *IEEE Access*, vol. 13, pp. 31544–31561, 2025. https://doi.org/10.1109/ACCESS.2025.3542266

[69] M. Miletic, "Synthetic data generation methods for longitudinal and time series health data: A systematic review," *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, pp. 1–14, 2025. https://doi.org/10.1186/s12911-025-03326-8

[70] M. Arones, I. Aybar-Bellido, W. Adauto-Medina, S. Rubiños-Jimenez, and J. A. Arévalo-Tuesta, "Functional validation of a generative AI-based mobile app for assessing speech difficulties in children," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 20, no. 1, pp. 137–159, 2026. https://doi.org/10.3991/ijim.v20i01.57991

[71] B. H. Hayadi and T. Hariguna, "Predictive analytics in mobile education: Evaluating logistic regression, random forest, and gradient boosting for course completion forecasting," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 5, pp. 210–232, 2025. https://doi.org/10.3991/ijim.v19i05.52381

[72] M. Imran, N. Almusharraf, M. S. Abdellatif, and M. Y. Abbasova, "Artificial intelligence in higher education: Enhancing learning systems and transforming educational paradigms," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 18, no. 18, pp. 34–48, 2024. https://doi.org/10.3991/ijim.v18i18.49143

7 AUTHORS

Ziyaeddin Halid İpek is with the Kafkas University, Kars, Türkiye.

Erol Poyraz is with the Kafkas University, Kars, Türkiye.

Gökhan Yigit is with the Kafkas University, Kars, Türkiye (E-mail: gyigit@kafkas.edu.tr).