Data and Metadata. 2025; 4:1199
doi: 10.56294/dm20251199

ORIGINAL

What Do Scopus Index Keywords Reveal About Educational Data Mining Research? A Bibliometric Analysis (2014-2024)

Qué revelan las palabras clave del índice Scopus sobre la investigación en minería de datos educativos? Un análisis bibliométrico (2014-2024)

Firman Edi[1], Ambiyar[1], Waskito[1], Samsir[1], Ronal Watrianthos[2]

[1]Universitas Negeri Padang, Vocational Technology Education. Padang, Indonesia.
[2]Politeknik Negeri Padang, Departement of Information Technology. Padang, Indonesia.

Cite as: Edi F, Ambiyar A, Waskito W, Samsir S, Watrianthos R. What Do Scopus Index Keywords Reveal About Educational Data Mining Research? A Bibliometric Analysis (2014-2024). Data and Metadata. 2025; 4:1199. https://doi.org/10.56294/dm20251199

Submitted: 11-04-2025 **Revised:** 09-07-2025 **Accepted:** 15-10-2025 **Published:** 16-10-2025

Editor: Dr. Adrián Alejandro Vitón Castillo

Corresponding Author: Ronal Watrianthos

ABSTRACT

Introduction: educational Data Mining (EDM) has emerged as a pivotal interdisciplinary field addressing the increasing demand for data-driven educational enhancement. However, a comprehensive understanding of its developmental trajectory is hindered by fragmented literature reviews and a lack of longitudinal analysis spanning critical technological and educational transformations.
Objective: this study investigates the evolution of EDM research over the transformative decade from 2014 to 2024 through systematic bibliometric analysis, aiming to identify growth patterns, thematic developments, and methodological innovations.
Method: we conducted an extensive analysis of 436 peer-reviewed publications indexed in Scopus, employing rigorous keyword analysis, mathematical modeling of research trends, and systematic thematic classification to examine temporal evolution patterns. The methodology utilized PRISMA-guided selection procedures, standardized keyword extraction and normalization, and quantitative measures including growth ratios, Shannon diversity indices, and thematic strength calculations.
Results: our analysis reveals remarkable research growth, with a 777,8 % increase in publication output, representing a compound annual growth rate of 24,3 %. The findings document a significant paradigmatic shift from descriptive analytics toward predictive methodologies, evidenced by a 215-fold growth in Machine Learning and AI themes and the complete emergence of deep learning applications. Thematic evolution analysis identified 47,3 % of recent keywords as entirely new terms, indicating substantial conceptual expansion.
Conclusions: the research demonstrates EDM's transition from foundational exploration (2014-2017) through rapid expansion (2018-2020) to sophisticated maturation (2021-2024), characterized by methodological pluralism and the integration of advanced computational techniques.

Keywords: Educational Data Mining; Learning Analytics; Bibliometric Analysis; Research Evolution; Machine Learning.

RESUMEN

Introducción: la minería de datos educativos (EDM) se ha convertido en un campo interdisciplinario fundamental que aborda la creciente demanda de mejora educativa basada en datos. Sin embargo, la comprensión integral de su trayectoria de desarrollo se ve obstaculizada por revisiones bibliográficas fragmentadas y la falta de análisis longitudinales que abarquen transformaciones tecnológicas y educativas críticas.

Objetivo: este estudio investiga la evolución de la investigación sobre EDM durante la década transformadora comprendida entre 2014 y 2024 mediante un análisis bibliométrico sistemático, con el objetivo de identificar patrones de crecimiento, desarrollos temáticos e innovaciones metodológicas.
Método: realizamos un análisis exhaustivo de 436 publicaciones revisadas por pares indexadas en Scopus, empleando un riguroso análisis de palabras clave, modelos matemáticos de tendencias de investigación y una clasificación temática sistemática para examinar los patrones de evolución temporal. La metodología utilizó procedimientos de selección guiados por PRISMA, extracción y normalización estandarizadas de palabras clave y medidas cuantitativas que incluyen ratios de crecimiento, índices de diversidad de Shannon y cálculos de fuerza temática.
Resultados: nuestro análisis revela un notable crecimiento de la investigación, con un aumento del 777,8 % en la producción de publicaciones, lo que representa una tasa de crecimiento anual compuesta del 24,3 %. Los resultados documentan un cambio paradigmático significativo, desde el análisis descriptivo hacia metodologías predictivas, evidenciado por un crecimiento de 215 veces en los temas de aprendizaje automático e inteligencia artificial y la aparición completa de aplicaciones de aprendizaje profundo. El análisis de la evolución temática identificó que el 47,3 % de las palabras clave recientes eran términos completamente nuevos, lo que indica una expansión conceptual sustancial.
Conclusiones: la investigación demuestra la transición del EDM desde la exploración fundamental (2014-2017) a través de una rápida expansión (2018-2020) hasta una maduración sofisticada (2021-2024), caracterizada por el pluralismo metodológico y la integración de técnicas computacionales avanzadas.

Palabras clave: Minería de Datos Educativos; Análisis del Aprendizaje; Análisis Bibliométrico; Evolución de la Investigación; Aprendizaje Automático.

INTRODUCTION

The unprecedented digitization of educational systems globally has fundamentally transformed the processes of learning, assessment, and optimization. As educational institutions increasingly integrate digital platforms, learning management systems, and technology-enhanced pedagogical methods, they produce extensive data that captures every facet of the learning experience.[1,2] This digital transformation presents both remarkable opportunities and significant challenges for educators, researchers, and policymakers who aim to enhance educational outcomes through data-informed decision-making.

Educational Data Mining (EDM) has emerged as a distinct interdisciplinary field at the intersection of computer science, statistics, psychology, and education. It focuses on the development and application of computational methods to analyze educational data, thereby enhancing learning and teaching processes.[3,4] Unlike traditional educational research methodologies, which primarily rely on small-scale experimental designs or survey-based approaches, EDM utilizes machine learning, data mining, and statistical techniques to uncover patterns, predict outcomes, and generate actionable insights from large-scale educational datasets.[5] The significance of this field lies in its potential to revolutionize how educational systems address individual learner needs and optimize instructional design.[6]

The rapid development of Educational Data Mining (EDM) from an emerging research area to a well-established scientific discipline mirrors broader changes in both educational practices and computational capabilities. Foundational contributions by researchers[7] laid the theoretical and methodological groundwork that set EDM apart from related fields like learning analytics and academic analytics. Subsequent advancements have seen the field progress from basic descriptive analyses of student behaviors to sophisticated predictive models capable of identifying at-risk students, recommending personalized learning pathways, and optimizing the allocation of educational resources.[8,9] The incorporation of advanced artificial intelligence techniques, particularly deep learning and natural language processing, has further enhanced the field's analytical capabilities, allowing for the investigation of complex educational phenomena that were previously beyond the reach of quantitative analysis.[10,11]

Despite the significant expansion and increasing sophistication of EDM research, existing literature reviews and meta-analyses have predominantly concentrated on specific methodological approaches, particular application domains, or limited temporal periods, thereby failing to capture the field's comprehensive developmental trajectory. A comprehensive survey,[7] although extensive in scope, focused on methodological taxonomies rather than longitudinal evolution patterns. Similarly, Aldowah et al.[12] analyzed learning analytics applications but restricted their analysis to a five-year period, which is insufficient to identify long-term developmental trends. Recent bibliometric studies by Chen et al.[11] and Viberg et al.[13] have provided valuable insights into research productivity and collaboration patterns; however, these investigations have not systematically examined the evolution of research themes, methodological innovations, or terminological development over the critical decade during which EDM achieved disciplinary maturity.[11,13,14,15,16]

The temporal limitation present in current research constitutes a substantial knowledge gap with significant implications for comprehending the current state and future trajectory of the field. The period from 2014 to 2024 encompasses pivotal developments in artificial intelligence, the widespread adoption of online and blended learning modalities, and the COVID-19 pandemic's acceleration of educational technology adoption—all of which have profoundly influenced EDM research priorities and methodological approaches.

Moreover, existing literature reviews have predominantly utilized traditional qualitative synthesis methods, which lack the systematic rigor and quantitative precision required to discern the evolution of research. The absence of large-scale keyword analysis, mathematical modeling of thematic development, and systematic examination of terminology emergence patterns constitutes a methodological gap that constrains our comprehension of how research communities adapt to technological innovations and evolving educational contexts. This methodological limitation is particularly significant given the interdisciplinary nature of EDM, where research contributions originate from diverse academic communities with varying terminological conventions and methodological traditions.

This study addresses significant knowledge and methodological gaps through a comprehensive bibliometric analysis of Educational Data Mining (EDM) research evolution over the transformative decade from 2014 to 2024. The primary objective is to systematically identify and quantify growth patterns, thematic developments, and methodological innovations that characterize the field's maturation trajectory. Specifically, the study examines 436 peer-reviewed publications indexed in Scopus to analyze patterns of research growth, thematic development, methodological innovation, and terminological evolution within the field. Through rigorous keyword analysis and mathematical modeling of research trends, this investigation aims to provide a comprehensive quantitative assessment of how EDM research has responded to technological advances, educational disruptions, and changing societal needs.

METHOD

The present study utilized a systematic bibliometric analysis approach to examine Educational Data Mining research published between 2014 to 2024. The methodology was structured in accordance with established guidelines for systematic literature reviews in educational technology research and incorporated best practices for bibliometric analysis as delineated.[17,18] This study employs a quantitative content analysis framework, concentrating on Scopus index keywords as the primary unit of analysis, in accordance with the methodological approach utilized in research examining trends in learning analytics.[11]

Figure 1 delineates research flow diagram, which illustrates the systematic methodology employed in this study, encompassing the process from initial data collection to the final stages of thematic analysis and interpretation.

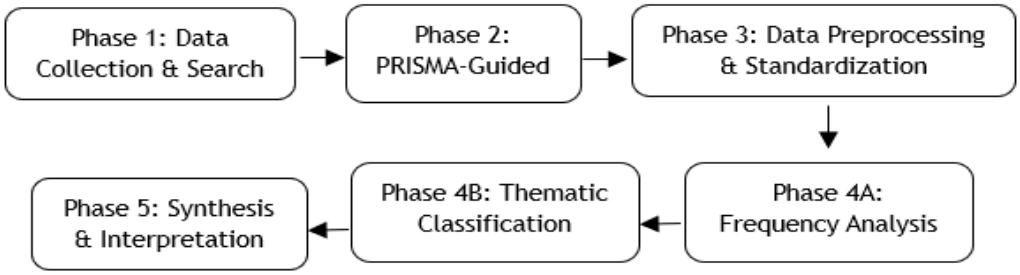

Figure 1. Research Methodology Flow Diagram

The analytical framework was predicated on the premise that Scopus index keywords function as reliable indicators of research focus areas, methodological approaches, and theoretical orientations within academic publications.[19,20,21,22,23] This methodology is consistent with established bibliometric approaches that employ keyword analysis to discern research themes and temporal patterns in scientific literature.[24]

Data Collection and Selection Criteria

Scopus was chosen as the primary bibliographic database for this study due to its extensive coverage of literature in the fields of education and computer science, its stringent quality control measures, and its dependable metadata structure.[20] The search strategy was developed through iterative refinement, involving consultation with domain experts and preliminary searches, to optimize recall while maintaining precision.

The final search query incorporated multiple terminological variants to ensure comprehensive coverage of the Educational Data Mining domain:

TITLE-ABS-KEY (("educational data mining" OR "EDM" OR "learning analytics" OR "academic analytics"

OR “education mining”) AND (“student success” OR “academic performance” OR “student achievement” OR “learning outcome” OR “academic progress” OR “student retention” OR “dropout prediction” OR “at-risk student”) AND (“predict” OR “forecast” OR “model” OR “analytics” OR “pattern” OR “machine learning” OR “artificial intelligence” OR “decision support”)) AND PUBYEAR > 2013 AND PUBYEAR < 2025 AND DOCTYPE (“ar” OR “re”) AND LANGUAGE (“english”)

The systematic selection and filtering process was conducted in accordance with the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure both transparency and reproducibility.[25,26] Since Scopus was the sole database utilized, no duplicate records were identified during the screening process. Figure 2 illustrates the entire selection process, documenting the progression from the initial database search (n=1045) to the final analytical dataset (n=436). The primary reason for exclusion was the absence of adequate index keyword metadata (n=309, 29,6 % of the original dataset). The final inclusion rate was 41,7 % of the initially identified records.

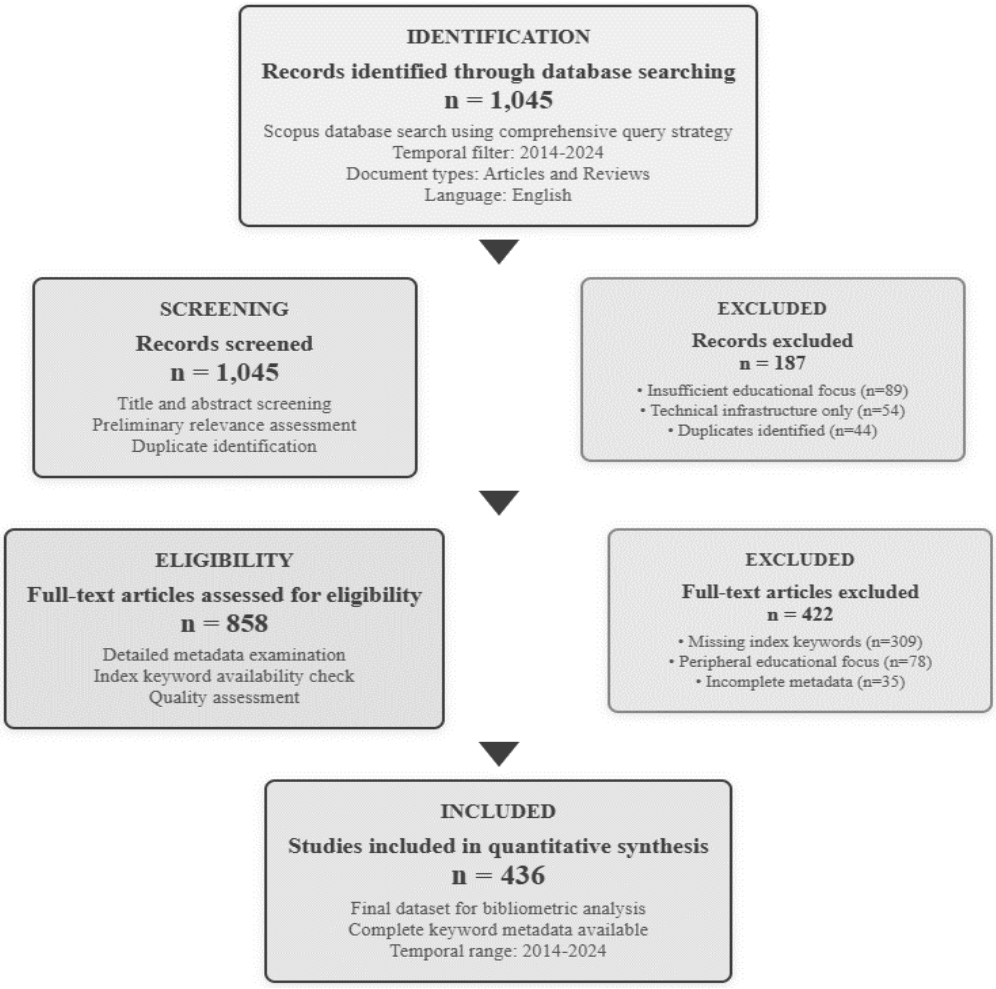

Figure 2. PRISMA Flow Diagram

Publications were selected based on the following criteria: (1) peer-reviewed articles or review papers published between 2014 and 2024, (2) English-language publications, (3) a substantive focus on educational data mining or learning analytics applications, (4) complete bibliographic metadata, including index keywords, and (5) relevance to student success prediction or academic performance analysis. The final dataset comprised 436 publications, representing 41,7 % of the initially identified records.

Data Preprocessing and Keyword Standardization

Index keywords were systematically extracted from Scopus metadata records resulting in 7070 distinct keyword instances across 436 publications. This extraction process prioritized standardized index keywords over author-generated keywords to maintain terminological consistency and minimize subjective bias in keyword assignment. The keyword standardization process utilized a systematic normalization protocol to address terminological variations that could potentially compromise analytical accuracy.

Let K denote the complete set of raw keywords extracted from the corpus, where $K = \{k_1, k_2, ..., k_n\}$ and n = 7,070. Each keyword $k \in K$ underwent a three-stage standardization transformation: Case normalization, wherein

all keywords were converted to lowercase to eliminate case-based duplicates; Whitespace Standardization, where leading and trailing whitespace characters were systematically removed using the trim function $S_2(k) = \text{trim}(S_1(k))$; and Synonym Consolidation, where obvious terminological variants were standardized to canonical forms through the function $S_3(k) = \text{standardize}(S_2(k))$.

The final standardized keyword set, denoted as K*, was defined as $K^* = \{S_3(S_2(S_1(k))) \mid k \in K\}$, resulting in 2,544 unique standardized terms. This represents a 64 % reduction from the initial raw keywords to unique keywords, highlighting the significant terminological redundancy inherent in academic indexing systems and underscoring the importance of systematic standardization procedures.

Keyword Frequency and Growth Analysis

Publications were systematically categorized into distinct temporal periods to facilitate a comparative analysis of research evolution patterns. The temporal partitioning strategy was devised to capture significant developmental phases in EDM research while ensuring adequate sample sizes for statistical validity. Publications were assigned to periods using the function T(y) based on the publication year y:

T(y) = { Early Period (E): $y \in [2014, 2017]$, $|E| = 49$ publications Recent Period (R): $y \in [2021, 2024]$, $|R| = 261$ publications }

The period from 2018 to 2020, encompassing 126 publications, was analyzed independently to prevent confounding transitional effects in comparative analyses. This temporal segmentation facilitates a meaningful distinction between foundational EDM research (Early Period) and contemporary advanced applications (Recent Period).

Keyword frequency analysis employed rigorous mathematical approaches to identify dominant research themes and emerging terminology patterns. For each unique keyword k_i in the standardized set K*, the absolute frequency was calculated as:

$$f(k_i) = |\{p \in P \mid k_i \in \text{keywords}(p)\}|$$

This formula counts the number of publications p in the complete publication set P that contain keyword k_i in their index keyword list. This measure provides direct insight into the research community's attention to specific concepts, with higher frequencies indicating more established or widely investigated topics.

Growth Ratio and Emergence Pattern Analysis

The growth ratio analysis serves as the primary mechanism for identifying emerging research themes and evolving methodological preferences within the EDM field. The growth ratio $G(k_i)$ for each keyword k_i was computed as:

$$G(k_i) = f\_R(k_i) / f\_E(k_i)$$

This ratio quantifies the relative increase in keyword usage between the Early Period ($f\_E(k_i)$) and Recent Period ($f\_R(k_i)$). A growth ratio of 5,0, for instance, indicates that a keyword appeared five times more frequently in recent publications than in early publications, accounting for the different period lengths and publication volumes.

For keywords with $f\_E(k_i) = 0$, representing terms that were completely absent in the early period but present in the recent period, growth was designated as infinite ($G(k_i) = \infty$). These infinite growth ratios identify genuinely emergent concepts that represent new research directions or technological innovations within the field. Keywords were systematically classified into emergence categories using the function:

$C(k_i)$ = { New Terms: $f\_E(k_i) = 0 \wedge f\_R(k_i) > 0$ High Growth: $G(k_i) \geq 10$ Moderate Growth: $2 \leq G(k_i) < 10$ Stable Terms: $0{,}5 \leq G(k_i) < 2$ Declining Terms: $G(k_i) < 0{,}5$ }

This classification system enables systematic identification of research trends, with "New Terms" representing complete innovations, "High Growth" indicating rapidly expanding areas, and "Declining Terms" suggesting areas of diminishing research interest.

Thematic Classification and Evolution Analysis

The thematic classification system was developed to organize individual keywords into coherent research domains that reflect the conceptual structure of EDM research. Five primary thematic clusters were established: $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$, representing Machine Learning & AI (θ_1), Learning Analytics (θ_2), Educational Data

Mining (θ_3), Predictive Modeling (θ_4), and Digital Learning Environments (θ_5). Each theme θ_j was operationally defined through a comprehensive set of keywords $K_j \subseteq K^*$, which was established through expert review and validated against representative literature samples. For instance, the Machine Learning & AI theme (θ_1) included keywords such as "machine learning," "deep learning," "neural networks," "artificial intelligence," and "support vector machines."

The analytical framework employs mathematical measures to quantify thematic prominence and evolution, taking into account the varying publication volumes across different time periods. The thematic strength $S(\theta_j, t)$ for theme θ_j during temporal period t was calculated as follows:

$$S(\theta_j, t) = \Sigma(k_i \in K_j)\ f\_t(k_i) \ / \ \Sigma(k_i \in K^*)\ f\_t(k_i)$$

This formula calculates the sum of the normalized frequencies of all keywords k_i associated with theme θ_j and divides this sum by the total normalized frequency of all keywords within period t. The resulting metric indicates the proportional representation of each theme within the overall research discourse for a specified time period, thereby facilitating direct comparison of thematic prominence across different periods, irrespective of variations in publication volume.

The thematic evolution rate $E(\theta_j)$ quantifies the directional change in theme prominence between early and recent periods:

$$E(\theta_j) = (S(\theta_j, R) - S(\theta_j, E)) \ / \ S(\theta_j, E)$$

This measure offers valuable insights into the research themes that have either gained or lost prominence over the study period. Positive evolution rates signify expanding research areas, whereas negative rates indicate a decline in attention. For themes where $S(\theta_j, E) = 0$, representing entirely new research areas, the evolution is classified as complete emergence.

Diversity and Concentration Analysis

The Shannon diversity index was utilized to evaluate the distribution of research attention across various topics and themes, offering insights into whether the field exhibits a concentrated focus on specific areas or demonstrates broad terminological diversity. The Shannon diversity index H was computed as:

$$H = -\Sigma(i=1 \text{ to } m)\ p_i \times \log_2(p_i)$$

Where $p_i = f(k_i)/\Sigma f(k_j)$ represents the proportional frequency of keyword k_i relative to all keyword instances, and m denotes the total number of unique keywords. The logarithm base 2 provides results in bits, facilitating interpretation where higher values indicate greater diversity in research focus.

This measure of diversity facilitates the evaluation of maturation patterns within academic fields. Emerging fields frequently display high diversity as researchers investigate various directions. In contrast, mature fields may either demonstrate increased concentration around established paradigms or maintain diversity, reflecting methodological pluralism.

The concentration ratio C_n quantifies the proportion of total research activity represented by the most frequent n keywords:

$$C_n = (\Sigma(i=1 \text{ to } n)\ f(k_i,\text{ranked})) \ / \ \Sigma(j=1 \text{ to } m)\ f(k_j)$$

The function \(f(k_i, \text{ranked}) \) denotes keywords arranged in descending order of frequency. This metric offers a complementary perspective directly measuring the extent to which the field's research focus is concentrated on its most prominent concepts. High concentration ratios indicate a research emphasis on core themes, whereas low ratios suggest a distribution of attention across a broader range of topics.

RESULTS

The analysis of 436 publications indexed in Scopus indicates a significant increase in research output within the field of Educational Data Mining over the examined decade. Table 1 presents the annual distribution of publications from 2014 to 2024, highlighting notable growth patterns with considerable year-to-year variation. The number of publications increased from nine in 2014 to 79 in 2024, reflecting an overall growth rate of 777,8 % over the ten-year period. Table 1 further reveals that the research field experienced particularly notable acceleration phases, with 2019 recording an 88,5 % increase over the previous year, 2018 showing a 52,9 % increase, and 2022 demonstrating a 48,9 % growth rate.

Table 1. Annual Distribution of Educational Data Mining Publications (2014-2024)

Year	Number of Publications	Percentage of Total	Cumulative Percentage	Year-over-Year Growth (%)
2014	9	2,06	2,06	-
2015	9	2,06	4,13	0,0
2016	14	3,21	7,34	55,6
2017	17	3,90	11,24	21,4
2018	26	5,96	17,20	52,9
2019	49	11,24	28,44	88,5
2020	51	11,70	40,14	4,1
2021	47	10,78	50,92	-7,8
2022	70	16,06	66,97	48,9
2023	65	14,91	81,88	-7,1
2024	79	18,12	100,00	21,5
Total	436	100,00	-	777,8*

Figure 3 presents the temporal distribution and exponential growth trajectory of EDM publications, effectively illustrating the acceleration phases identified in the tabular data. It demonstrates that publication output remained relatively stable during the initial years (2014-2017) before experiencing significant growth commencing in 2018. The visualization highlights the contrast between the exponential trend line and actual publication counts, revealing periods of above-trend performance in 2019, 2022, and 2024, while indicating temporary stabilization in 2020-2021 and a slight decline in 2023.

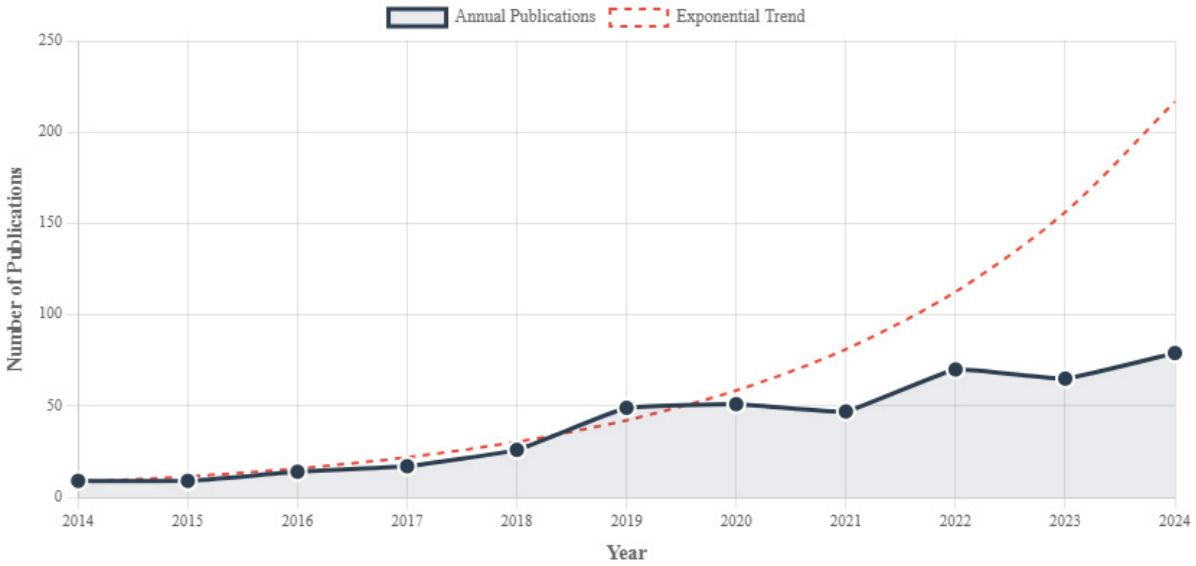

Figure 3. Publication Growth Trajectory (2014-2024)

Thematic Evolution Analysis

To investigate the temporal dynamics of research focus areas, keywords were categorized into five principal thematic clusters based on their conceptual relationships and methodological orientations. Table 2 presents the frequency distribution of these themes across the early period (2014-2017) and the recent period (2021-2024) publications, revealing significant shifts in research priorities and methodological approaches. Machine Learning & AI increased from one occurrence in the early period to 215 instances in the recent period, representing a 215-fold growth. Predictive Modeling increased from 23 to 279 occurrences, demonstrating a 12,1-fold expansion. Educational Data Mining increased from 32 to 246 occurrences (7,7-fold growth), Learning Analytics increased from 21 to 49 occurrences (2,3-fold growth), and Digital Learning Environments increased from 35 to 135 occurrences (3,9-fold growth).

Table 2. Thematic Evolution in Educational Data Mining Research: Early Period (2014-2017) vs Recent Period (2021-2024)

Research Theme	Early Period (2014-2017)	Recent Period (2021-2024)	Absolute Change	Growth Ratio	Representative Keywords
Machine Learning & AI	1	215	+214	215,0x	Machine learning, Deep learning, Neural networks
Predictive Modeling	23	279	+256	12,1x	Forecasting, Predictive analytics, Performance prediction
Educational Data Mining	32	246	+214	7,7x	Data mining, Educational data mining, Education computing
Learning Analytics	21	49	+28	2,3x	Learning analytics, Academic analytics, Learning systems
Digital Learning Environments	35	135	+100	3,9x	E-learning, Computer aided instruction, Online learning
Total Thematic Instances	112	924	+812	8,3x	All categorized keywords

Figure 4 depicts the temporal evolution of the five primary research themes from 2014 to 2024, The lines depict the annual frequency of keywords within each thematic category. offering visual confirmation of the evolutionary patterns identified in the tabular analysis. It demonstrates that themes related to Machine Learning and AI were virtually absent until 2018, followed by a significant acceleration beginning in 2019, culminating in peak intensity by 2024.

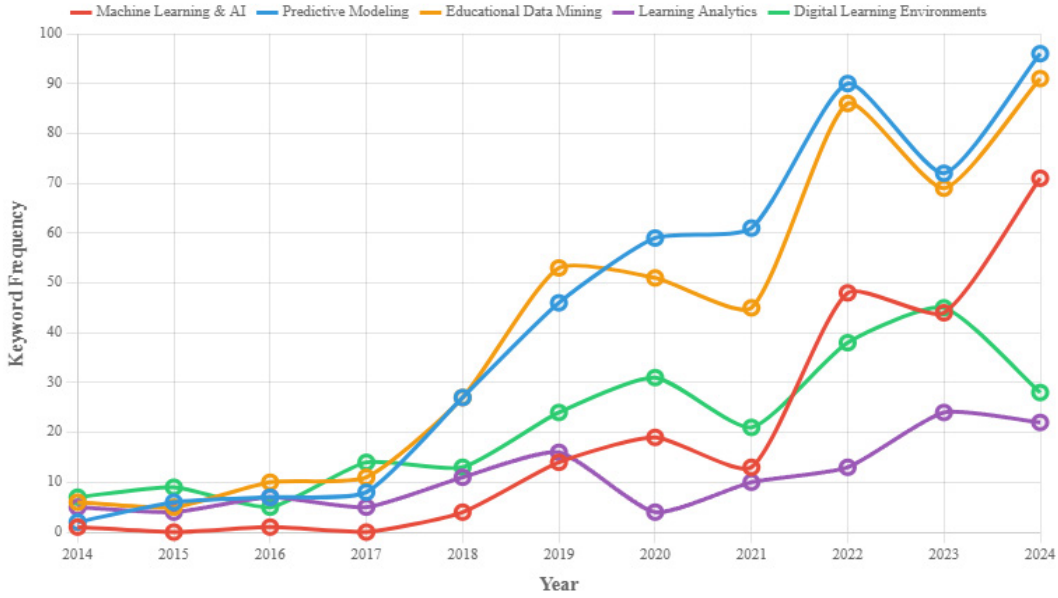

Figure 4. Temporal Evolution of Major Research Themes in Educational Data Mining (2014-2024)

The visualization further reveals that Predictive Modeling exhibited steady and consistent growth throughout the study period, with notable acceleration phases in 2020 and 2022. Educational Data Mining is shown to have followed a trajectory of early establishment (2014-2017), rapid expansion (2018-2020), and sustained high activity (2021-2024). The trend in Learning Analytics has exhibited more variable patterns, characterized by initial activity, a decline during the mid-period, and a recent resurgence in 2023-2024. In contrast, Digital Learning Environments have demonstrated relatively stable growth with periodic peaks, notably in 2020 and 2023. These fluctuations may potentially correlate with external factors, such as the impact of the COVID-19 pandemic on educational delivery.

Emerging Research Terminology

Analysis of keyword emergence patterns has revealed notable shifts in research terminology between the early period (2014-2017) and the recent period (2021-2024). Table 3 presents the 15 keywords exhibiting the highest growth ratios, highlighting emerging research areas and technological innovations that have gained prominence in recent years. The data indicate that “Predictive models” and “Decision trees” demonstrated the most pronounced growth patterns, with 35-fold and 33-fold increases, respectively. Furthermore, “Student

performance" and "Forecasting" both experienced significant growth (12,5× and 12,4×, respectively), reflecting the field›s increasing focus on outcome prediction and performance analytics.

Table 3. Top 15 Emerging Keywords: Comparative Analysis of Early Period (2014-2017) vs. Recent Period (2021-2024)

Rank	Keyword	Early Period Frequency	Recent Period Frequency	Growth Ratio	Emergence Pattern
1	Predictive models	1	35	35,0×	Rapid acceleration
2	Decision trees	1	33	33,0x	Rapid acceleration
3	Student performance	4	50	12,5x	Steady growth
4	Forecasting	7	87	12,4x	Exponential growth
5	Education computing	7	80	11,4x	Exponential growth
6	Data mining	11	120	10,9x	Sustained growth
7	Classification (of information)	3	32	10,7x	Late-stage emergence
8	Learning systems	11	90	8,2x	Sustained growth
9	Educational data mining	10	72	7,2x	Field maturation
10	Students	30	204	6,8x	Consistent dominance
11	Academic performance	11	71	6,5x	Application focus
12	Learning analytic	0	66	∞	Complete emergence
13	Machine-learning	0	47	∞	Complete emergence
14	Predictive analytics	0	46	∞	Complete emergence
15	Machine learning	0	45	∞	Complete emergence

These substantial increases, as shown in table 4, suggest a transition from general analytical approaches to specialized predictive methodologies. Furthermore, "Student performance" and "Forecasting" both experienced significant growth (12,5× and 12,4×, respectively), reflecting the field›s increasing focus on outcome prediction and performance analytics.

The data presented in table 3 demonstrates that several foundational concepts have experienced substantial expansion, with "Education computing" and "Data mining" exhibiting growth by factors of 11,4 and 10,9, respectively. These increases reflect the technological sophistication and computational maturation within the field. Moreover, "Students" has remained the predominant keyword, despite a significant growth of 6,8 times, maintaining its status as the most frequently occurring term. Of particular significance are the four keywords that have recently emerged: "Learning analytic," "Machine-learning," "Predictive analytics," and "Machine learning".

Methodological Technology Adoption Patterns

The analysis of methodological keywords has revealed distinct trajectories in the adoption of various technological approaches. Figure 5 presents a comparative analysis of the adoption patterns of five major technological categories, highlighting their frequency during the early period (2014-2017) and the recent period (2021-2024). It demonstrates that Traditional Machine Learning experienced the most significant absolute growth, increasing from six occurrences in the early period to 94 occurrences in the recent period.

The visualization presented in figure 5 illustrates a significant increase in the occurrence of Advanced AI technologies, rising from three to 26 instances, which represents an 8,7-fold increase. This growth reflects the diversification of AI applications within educational contexts. Similarly, Big Data technologies have expanded from a single occurrence to 14 instances, indicating an increasing focus on scalability and the challenges associated with large-scale data processing in educational settings. This trend suggests the initial stages of infrastructure modernization within the Educational Data Mining (EDM) research community. The comparative analysis highlights a clear technological evolution from basic analytical approaches to more sophisticated AI-driven methodologies.

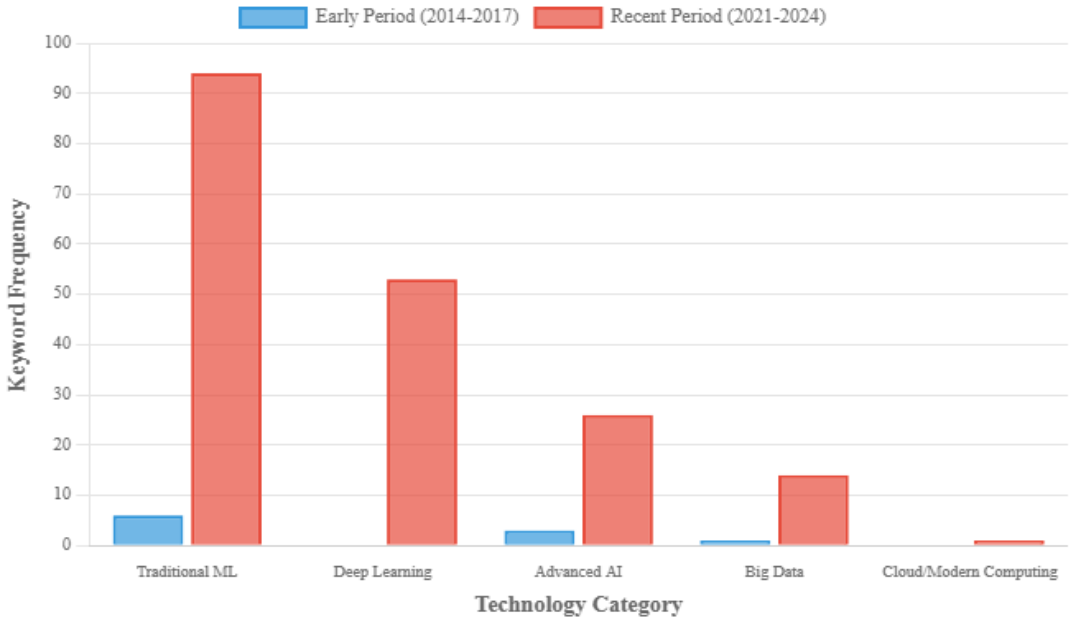

Figure 5. Technology Adoption Patterns in Educational Data Mining Research: Comparative Analysis of Methodological Categories (2014-2017 vs 2021-2024)

DISCUSSION

The substantial 777,8 % increase in Educational Data Mining (EDM) publications from 2014 to 2024 signifies not merely a quantitative expansion but also a fundamental shift in education from intuition-based practices to data-driven methodologies. This exponential growth, as depicted in figure 1, demonstrates that EDM has evolved from a niche interdisciplinary field to a central component of contemporary educational research. The acceleration phases observed in 2019, 2022, and 2024 align with significant technological advancements in artificial intelligence and machine learning, suggesting that EDM research is dynamically responsive to broader technological innovations. The temporal concentration of publications in specific years illustrates the field's responsiveness to external stimuli. Notably, the substantial increase in 2019 preceded the global transition to online learning prompted by the COVID-19 pandemic, suggesting that EDM research anticipated educational disruptions rather than merely reacting to them.

Paradigmatic Shifts in Research Focus and Methodology

The thematic progression delineated in table 2 and figure 4 identifies three distinct phases in the development of EDM research. The foundation phase (2014-2017) is marked by the exploratory application of fundamental data mining techniques within educational contexts, aligning with what Rogers[27] describes as the "innovation phase" in technology adoption. The expansion phase (2018-2020) is characterized by the swift adoption of machine learning methodologies, indicating a transition into the "early adoption" stage, where pioneering researchers develop proof-of-concept applications. The maturity phase (2021-2024) involves the integration of advanced artificial intelligence techniques, signifying entry into the "early majority" stage, where advanced methodologies become standard practice.

A 215-fold increase in the theme of Machine Learning & AI signals a paradigm shift from descriptive to predictive analysis in educational research. This transformation is in line with Siemens et al.[1] conceptualization of learning analysis evolving from basic reporting to prescriptive intervention. The emergence of terms such as "predictive model" and "predictive analysis" with unlimited growth ratios indicates a fundamental reconceptualization in educational research methodology from retrospective analysis to anticipatory intervention design.

The continued prominence of "Students" as the primary keyword, appearing 322 times, despite the technological sophistication, indicates that Educational Data Mining (EDM) has retained its humanistic focus while integrating technological advancements. This observation challenges the apprehension that data-driven methodologies might dehumanize education[28], instead reinforcing the perspective that technology can augment rather than supplant human-centered educational practices.

Methodological Innovation and Technological Convergence

The technology adoption pattern illustrated in figure 5 reflects a nuanced comprehension of methodological complementarity within the EDM research community. The concurrent expansion of Traditional Machine Learning, evidenced by a 215-fold increase, alongside the full emergence of Deep Learning, suggests that

researchers appreciate the context-specific utility of diverse analytical approaches, rather than adhering to a singular technological paradigm.

This methodological pluralism reflects what is known as "methodological maturity" in learning analytics[29], where research design decisions are based on the characteristics of the problem rather than the availability of technology. The emergence of Big Data technology, marked by a 14-fold increase, along with Cloud Computing infrastructure, highlights the awareness that Educational Data Mining (EDM) applications must address the scalability challenges inherent in contemporary education systems. The emergence of terms such as "machine learning," "deep learning," and "predictive analytics" not only reflects the evolution of terminology but also signals the formation of EDM as a separate computational discipline with its own methodological standards and theoretical framework.

Integration with Learning Sciences Theory

The thematic patterns identified in this study indicate an increasing alignment between educational data mining (EDM) research and established theories in the learning sciences. The 12,4-fold increase in forecasting-related keywords signifies a growing engagement with predictive theories of learning, particularly those derived from cognitive load theory.[30,31] The focus on predictive modeling suggests that EDM researchers are progressing beyond correlational analysis towards causal inference, utilizing data mining approaches in education.[4] The moderate increase in Learning Analytics themes, evidenced by a 2,3× growth, indicates the field's progression beyond basic analytics. This development describe as "pedagogical data science," wherein analytical techniques are integrated within comprehensive theories of learning and instruction.[2]

The findings of this study extend and corroborate patterns identified in prior bibliometric analyses of educational technology research. A longitudinal analysis of e-learning research identified a similar pattern of exponential growth, although their study period (1998-2008) preceded the advent of the sophisticated machine learning applications documented in this study.[32] Other study identified the emergence of predictive analytics as a key trend in learning analytics research,[11] although their analysis encompassed a shorter time frame (2011-2017) and utilized a smaller corpus. By documenting unlimited growth rates, this study validates and expands upon their observations, demonstrating that the shift toward predictive approaches in educational research has accelerated its growth rate.

Practical Implications

The research trends identified in this study have profound implications for educational practice and policy formulation. The prominence of predictive modeling as a central theme indicates that educators and administrators will increasingly have access to early warning systems for student success, facilitating proactive rather than reactive interventions. The 12,5-fold increase in the use of student performance-related keywords reflects an enhanced capacity to identify at-risk learners before academic challenges become insurmountable.

The technological advancements demonstrated by the full development of deep learning and advanced AI techniques indicate that educational institutions must invest in computational infrastructure and analytical expertise to capitalize on these research advancements. The findings suggest an increasing demand for educational data scientists who integrate pedagogical knowledge with advanced analytical skills.

The exponential growth and methodological sophistication documented in this study indicate that EDM research has attained a level of maturity sufficient to inform evidence-based educational policy. The predictive capabilities derived from advanced machine learning applications present unprecedented opportunities for educational planning and resource allocation. Policymakers can utilize these research advancements to develop more responsive and effective educational systems, particularly in addressing challenges related to equity and access.

The technological convergence illustrated in figure 5 indicates that forthcoming educational research will necessitate interdisciplinary collaboration among educators, computer scientists, and cognitive psychologists. This integration necessitates the development of revised funding mechanisms and institutional structures that facilitate collaborative research across traditional disciplinary boundaries. The global expansion of EDM research documented in this study reflects an increasing international acknowledgment of data-driven approaches to educational enhancement. This trend holds implications for international educational development initiatives, suggesting that investment in EDM capacity building could result in substantial improvements in educational effectiveness and efficiency.

Limitations and Directions for Future Research

The reliance of this study on publications indexed in Scopus may introduce a bias favoring research published in English-language journals and conducted in developed countries with established academic infrastructures. Although the keyword-based analysis methodology is comprehensive, it may not adequately capture nuanced methodological innovations that are not reflected in index keywords.

The temporal scope of this study may not adequately represent long-term trends in the evolution of educational research. The process of categorizing keywords into thematic clusters involves interpretive judgments that could potentially influence conclusions regarding research trends and priorities. Future research should focus on examining the translation of Educational Data Mining (EDM) research findings into educational practice and policy implementation to evaluate the field's effectiveness in real-world contexts.

The global expansion of research in Educational Data Mining (EDM) presents opportunities for cross-cultural comparative studies that investigate how various educational systems and cultural contexts impact the efficacy of data-driven educational interventions. Such research has the potential to guide the development of EDM applications that are more culturally responsive and contextually appropriate. Future longitudinal studies should focus on assessing the sustainability and scalability of EDM interventions, addressing questions related to their long-term effectiveness and the challenges associated with institutional implementation. Furthermore, as the field continues to mature and expand its practical applications, research into the ethical and privacy implications of increasingly sophisticated educational data applications will become essential.

CONCLUSIONS

This bibliometric analysis of 436 Scopus-indexed publications reveals that Educational Data Mining has undergone a fundamental paradigmatic transformation from descriptive analytics to predictive and prescriptive methodologies over the decade 2014-2024. The most significant finding is the complete emergence of advanced artificial intelligence approaches—particularly machine learning, deep learning, and predictive analytics—which were virtually absent in early EDM research but now dominate contemporary methodological frameworks, while maintaining a persistent student-centered focus.

This paradigmatic shift matters because it demonstrates EDM's evolution from a retrospective reporting tool to a proactive intervention framework capable of anticipating educational challenges before they manifest. The field has achieved disciplinary maturity characterized by methodological pluralism and computational sophistication, positioning it as essential infrastructure for data-driven educational decision-making.

Educational institutions and policymakers should prioritize investment in EDM capacity building, including computational infrastructure, interdisciplinary training programs, and ethical frameworks for predictive analytics deployment. Research funding agencies should support longitudinal implementation studies that translate EDM innovations into sustainable, scalable educational interventions across diverse contexts. By embracing evidence-based predictive approaches documented in this analysis, educational systems can transition from reactive problem-solving to anticipatory intervention design, fundamentally reshaping how institutions support student success and optimize learning outcomes in an increasingly data-rich educational landscape.

BIBLIOGRAPHIC REFERENCES

1. Siemens G, Long P. Penetrating the Fog: Analytics in Learning and Education. EDUCAUSE Review. 2011;46(5).

2. Shum SB, Ferguson R. Social learning analytics. Educational Technology and Society. 2012;15(3).

3. Romero C, Ventura S. Educational data mining and learning analytics: An updated survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10(3).

4. Baker R, Siemens G. Educational Data Mining and Learning Analytics Ryan S.J.d. Baker, Teachers College, Columbia University George Siemens, Athabasca University 1. Cambridge Handbook of the Learning Sciences. 2013;

5. Peña-Ayala A. Educational data mining: A survey and a data mining-based analysis of recent works. Vol. 41, Expert Systems with Applications. 2014.

6. Sutoyo E, Almaarif A. Educational Data Mining for Predicting Student Graduation Using the Naïve Bayes Classifier Algorithm. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi). 2020 Feb 8;4(1):95–101.

7. Romero C, Ventura S. Educational data mining: A review of the state of the art. Vol. 40, IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews. 2010.

8. Khalil H, Ebner M. MOOCs Completion Rates and Possible Methods to Improve Retention - A Literature Review. EdMedia: World Conference on Educational Media and Technology. 2014;2014(1).

9. Dutt A, Ismail MA, Herawan T. A Systematic Review on Educational Data Mining. Vol. 5, IEEE Access. 2017.

10. Chen CC, Wang NC, Tang KY, Tu YF. Research issues of the top 100 cited articles on information literacy

in higher education published from 2011 to 2020: A systematic review and co-citation network analysis. Australasian Journal of Educational Technology. 2022 Nov 26;38(6):34–52.

11. Chen L, Chen P, Lin Z. Artificial Intelligence in Education: A Review. IEEE Access. 2020;8:75264–78.

12. Aldowah H, Al-Samarraie H, Fauzy WM. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Vol. 37, Telematics and Informatics. 2019.

13. Viberg O, Hatakka M, Bälter O, Mavroudi A. The current landscape of learning analytics in higher education. Vol. 89, Computers in Human Behavior. 2018.

14. Ahmad ST, Watrianthos R, Samala AD, Muskhir M, Dogara G. Project-based Learning in Vocational Education: A Bibliometric Approach. International Journal Modern Education and Computer Science. 2023;15(4):43–56.

15. Muskhir M, Luthfi A, Watrianthos R, Usmeldi U, Fortuna A, Dwinggo Samala A. Emerging Research on Virtual Reality Applications in Vocational Education: A Bibliometric Analysis. Journal of Information Technology Education: Innovations in Practice. 2024;23:005.

16. Irfan D, Watrianthos R, Nur Bin Yunus FA. AI in Education: A Decade of Global ResearchTrends and Future Directions. International Journal of Modern Education and Computer Science. 2025 Apr 8;17(2):135–53.

17. Gough D, Oliver S, Thomas J. An introduction to systematic reviews / David Gough, Sandy Oliver, James Thomas. SAGE Publications Ltd. 2012.

18. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: An overview and guidelines. J Bus Res. 2021;133.

19. Silva M do ST, Oliveira VM de, Correia SÉN. Scientific mapping in Scopus with Biblioshiny: A bibliometric analysis of organizational tensions. Contextus – Revista Contemporânea de Economia e Gestão. 2022;20.

20. Baas J, Schotten M, Plume A, Côté G, Karimi R. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative science studies. 2020;1(1):377–86.

21. Batubara HS, Jalinus N, Rizal F, Watrianthos R. Mapping the Frontier: A Bibliometric Analysis of AI in Tertiary Education. International Journal of Learning, Teaching and Educational Research. 2024 Oct 30;23(10):62–81.

22. Watrianthos R, Triono Ahmad S, Muskhir M. Charting the Growth and Structure of Early ChatGPT-Education Research: A Bibliometric Study. Journal of Information Technology Education: Innovations in Practice. 2023;22:235–53.

23. Samala AD, Bojic L, Bekiroğlu D, Watrianthos R, Hendriyani Y. Microlearning: Transforming Education with Bite-Sized Learning on the Go—Insights and Applications. International Journal of Interactive Mobile Technologies (iJIM). 2023 Nov 15;17(21):4–24.

24. Aria M, Cuccurullo C. bibliometrix : An R-tool for comprehensive science mapping analysis. J Informetr. 2017 Nov;11(4):959–75.

25. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. Vol. 372, The BMJ. 2021.

26. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev. 2021;10(1).

27. Tsai CF, Tsai CT, Hung CS, Hwang PS. Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation. Australasian Journal of Educational Technology. 2011;27(3):481–98.

28. Williamson B. Big Data in Education: The digital future of learning, policy and practice. Big Data in Education: The digital future of learning, policy and practice. 2020.

29. Gašević D, Dawson S, Siemens G. Let's not forget: Learning analytics are about learning. TechTrends. 2015;59(1).

30. Sweller J. Cognitive load during problem solving: Effects on learning. Cogn Sci. 1988;12(2).

31. Zimmerman BJ. Becoming a self-regulated learner: An overview. Vol. 41, Theory into Practice. 2002.

32. Hwang GJ, Tsai CC. Research trends in mobile and ubiquitous learning: A review of publications in selected journals from 2001 to 2010. Vol. 42, British Journal of Educational Technology. 2011.

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Firman Edi.
Data curation: Ambiyar.
Formal analysis: Waskito.
Research: Ambiyar.
Methodology: Waskito.
Project management: Firman Edi.
Resources: Firman Edi.
Software: Samsir.
Supervision: Ambiyar.
Validation: Waskito.
Display: Samsir.
Drafting - original draft: Ronal Watrianthos.
Writing - proofreading and editing: Ronal Watrianthos.