

Applications of Machine Learning in Accounting: Current Trends and Future Perspectives

Federico Chmet

Università del Piemonte Orientale

Davide Calandra

Silvana Secinaro

Università di Torino

Ahmed A. Elamer

Alfaisal University

ABSTRACT: Machine learning (ML) is increasingly used to detect fraud, predict distress, analyze disclosures, and support audit work. Prior reviews map applications or methods but rarely connect accounting problems to ML design, validation, and governance choices. We reviewed 273 peer-reviewed Web of Science Core Collection articles published from 1992 to 2025, combining bibliometric science mapping with an antecedents-decisions-outcomes (ADO) synthesis enriched by a theory-context-method (TCM) lens. The literature clusters around distress and bankruptcy prediction, reporting quality, and fraud, bridged by audit analytics and textual analysis. ML adoption is driven mainly by assurance pressure, unstructured information, and digital process data; however, many studies rely on accuracy-oriented metrics, nontemporal validation, and explainability that is reported rather than tested. We show how accounting conditions should discipline ML evaluation and propose a research agenda for cost-aware, temporally robust, and governable ML in accounting.

Data Availability: Data are available on request from the authors.

Keywords: machine learning; accounting; auditing; bibliometric analysis; ADO-TCM framework.

I. INTRODUCTION

Machine learning (ML) is now important for accounting both as a practical tool for dealing with large, heterogeneous evidence and as a scholarly object for understanding how data-driven models affect accounting judgment, assurance, and reporting quality. Enterprise systems generate granular transaction traces, control

The authors received no financial support for the research, authorship, and/or publication of this article. The authors declared no potential conflicts of interest.

Federico Chmet, Università del Piemonte Orientale, Dipartimento di studi per l'economia e l'impresa, Novara, Italy; Davide Calandra and Silvana Secinaro, Università di Torino, Scuola di Management ed Economia, Dipartimento di Management "Valter Cantino," Torino, Italy; Ahmed A. Elamer, Alfaisal University, College of Business, Accounting, Riyadh, Saudi Arabia.

Supplemental material is available online, as linked in the text.

Any use of generative artificial intelligence (AI) or AI tools by the authors has been disclosed in the "[Declaration of Generative AI and AI-Assisted Technologies](#)" section.

Editor's note: Accepted by Senior Editor Tawei Wang.

Copyright © 2026 Federico Chmet, Davide Calandra, Silvana Secinaro, and Ahmed A. Elamer. *Journal of Information Systems* published by the American Accounting Association. This is an open access article licensed under Creative Commons Attribution CC-BY 4.0, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited.

Submitted: July 2025

Accepted: June 2026

Early Access: August 2026

logs, and operational metadata, whereas corporate reporting combines quantitative statements with narrative disclosures, audit reports, and other text-intensive artifacts. In this environment, ML, which is a data-driven subset of artificial intelligence, has been adopted to support tasks in which the scale, heterogeneity, and nonlinearity of signals challenge conventional analytical approaches (Jordan and Mitchell 2015; Bhimani and Willcocks 2014). Within accounting research, ML applications have expanded from early predictive uses (e.g., distress and fraud screening) to a broader range of applications, including misstatement detection, audit analytics, and textual analysis of disclosures and audit communications (Perols 2011; Hayes and Boritz 2021; Bao, Ke, Li, Yu, and Zhang 2020).

Simultaneously, accounting settings impose constraints that complicate the direct translation of predictive performance into decision value. ML outputs often support professional judgment under asymmetric error costs, temporal dependence, and accountability requirements that emphasize documentation, reproducibility, and reviewability, especially when models inform audit planning, exception triage, or reporting quality assessments (Bertomeu 2020; Zhang, Cho, and Vasarhelyi 2022). These constraints are becoming more salient as ML is applied to unstructured disclosures and process-level data (e.g., deep learning for restatement classification, natural language processing (NLP)-based interpretation of key audit matters, process mining combined with deep learning for suspicious transaction detection) (Hayes and Boritz 2021; Liu, Yen, and Wu 2022; Wang, T. Chiu, V. Chiu, and Vasarhelyi 2025).

As the literature has expanded, synthesizing what we know has become increasingly difficult. Prior reviews often organize evidence either by topical domains (e.g., fraud detection, distress prediction, and textual analysis) or by families of methods (e.g., neural networks, ensembles, and NLP), which is informative but can fragment the accumulation of knowledge (Amani and Fadlalla 2017; Kureljusic and Karger 2024). Topic- or method-centric syntheses are useful, but they often leave one question underdeveloped: whether ML design and validation choices fit the accounting problem being addressed. This matters because fraud detection, distress prediction, disclosure analytics, and audit planning impose different requirements on error costs, validation, documentation, and reviewability (Bertomeu 2020; Hardies, Ohlrogge, Mentens, and Vandennieuwenhuysen 2024). Without a linkage-based synthesis, there is a risk of conflating “algorithmic accuracy” with accounting value and of overlooking the governance conditions required for models to be auditable and useful in practice.

Two limitations in existing reviews motivate our study. First, prior syntheses clarify where ML is used in accounting, but they provide less evidence on whether ML design choices fit the accounting problems they address. This distinction matters because the same performance metric or validation design does not carry the same meaning across fraud detection, distress prediction, disclosure analytics, and audit planning. Fraud and misstatement settings involve asymmetric error costs; distress and disclosure settings often involve temporal dependence; and audit settings require documentation and reviewability. A review of this literature, therefore, needs to connect the accounting conditions to the ML decision and the claimed outcome. Second, theory is often used to motivate the empirical setting but less often to discipline ML design and evaluation choices, such as cost-sensitive loss functions, validation windows, threshold selection, or explainability artifacts.

To address this gap, we combine an antecedents-decisions-outcomes (ADO) synthesis with a theory-context-method (TCM) lens (Paul, Parthasarathy, and Gupta 2017; Paul and Benito 2018). ADO is needed because it links the accounting conditions that motivate ML adoption to the design and validation decisions through which ML is implemented and to the outcomes that are claimed. TCM is needed because it shows how this evidence is produced: which theories frame the problem, which accounting contexts are examined, and which methods support the claims reported. Used together, ADO and TCM allow us to move beyond a descriptive inventory of applications and toward a synthesis of whether the ML choices are aligned with accounting problems.

This framing leads directly to two research questions. The first maps the structure of the field and how evidence is produced, whereas the second examines whether accounting antecedents are matched by appropriate ML design, validation, and outcome claims. Accordingly, we ask:

RQ1: What intellectual structure, application domains, and theory-context-method patterns characterize ML research in accounting?

RQ2: How do accounting antecedents shape ML design and validation decisions, and how do these decisions condition reported outcomes and future research gaps?

We answer RQ1 through bibliometric science mapping and TCM coding. The bibliometric analysis identifies the field’s thematic structure and research fronts, whereas the theory, context, and method codes show how evidence is produced across accounting settings. We answer RQ2 through the ADO synthesis, which links adoption antecedents to design and validation decisions and the outcomes those decisions condition. This sequencing ensures that the descriptive map of the field is used as the basis for a more diagnostic assessment of alignment between accounting problems and ML choices.

Our protocol-driven search and screening were conducted on the Web of Science Core Collection (WoSCC), yielding a final sample of 273 peer-reviewed journal articles published between 1992 and 2025. The selection trajectory is

documented in a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-style flow diagram, and screening decisions are logged record-by-record using standardized exclusion rationales.

Our review contributes in four ways. First, it provides a transparent map of ML research in accounting, clarifying the field's core themes and the bridging role of audit analytics and textual analysis across application domains. Second, it links adoption antecedents (e.g., assurance/compliance pressure and unstructured information growth) to recurring design and validation decisions (e.g., task framing under rare events and asymmetric costs; representation choices for structured and unstructured data; temporally robust evaluation; explainability artifacts as assurance-relevant documentation) and to the outcomes these choices condition (Perols 2011; Hayes and Boritz 2021; Bao et al. 2020). Third, by reporting theory and method patterns through TCM, it identifies where theoretical framing is present but not yet tightly coupled with ML design and governance requirements in accountability-sensitive settings (Zhang et al. 2022; Janvrin, Raschke, Schneider, and Hsieh 2025). Finally, it distills a forward agenda emphasizing cost-aware evaluation, human-algorithm teaming, and documentation/provenance practices required for reproducible and governable ML in accounting.

The rest of the paper is structured as follows. Section II presents the background and positioning. Section III describes the review protocol and data collection. Section IV reports the bibliometric mapping results. Section V presents the ADO-TCM synthesis. Section VI discusses the implications, limitations, and forward agenda.

II. BACKGROUND AND POSITIONING

Machine learning (ML) has moved from early classification tools for distress and fraud to a set of deployable analytics supporting accounting and assurance tasks where signals are high-dimensional, heterogeneous, and often unstructured. Foundational studies applied neural networks to financial distress and fraud risk (Coats and Fant 1993; Yang, M. Platt, and H. Platt 1999), whereas more recent work demonstrates how modern ML pipelines improve screening and prioritization for misstatements and fraud under rare events and asymmetric costs (Perols 2011; Bertomeu 2020; Parker, Jiang, Cho, and Vasarhelyi 2025; Bao et al. 2020). In Accounting Information Systems (AIS) outlets, the stream has broadened to disclosure and audit-text analytics and process-aware audit applications, including restatement classification and sentiment extraction from audit communications, and deep-learning process mining for suspicious transaction detection (Hayes and Boritz 2021; Duan, Vasarhelyi, and Codesso 2025; Wang et al. 2025).

Existing reviews document this expansion, but they usually organize the evidence by application domains or method families. This is useful for mapping the field, yet it makes the fit between accounting conditions and ML choices less visible. The key unresolved issue is therefore not whether ML is used in accounting, but whether design, validation, and explanation practices are aligned with the decision setting in which the model is expected to operate.

Accordingly, our review complements prior work by combining bibliometric science mapping with an antecedents-decisions-outcomes synthesis enriched by a theory-context-method lens (Paul et al. 2017; Paul and Benito 2018). Bibliometrics identifies the field's structure, whereas ADO-TCM connects adoption drivers, design and validation choices, reported outcomes, and the evidentiary base supporting them. This positioning shifts the review from a descriptive catalog of applications toward a diagnostic assessment of whether ML methods are appropriate for the accounting problems they address.

In this synthesis, ADO is used to link accounting antecedents, ML decisions, and reported outcomes rather than merely to classify papers. Antecedents include conditions such as fraud risk, assurance pressure, unstructured information, and process digitalization. Decisions capture task framing, model design, validation, thresholding, and explainability choices. Outcomes refer to predictive lift, audit-effort allocation, exception triage, and reviewability. This structure makes it possible to assess alignment: for instance, fraud and misstatement settings imply asymmetric error costs, so evidence based only on accuracy or area under the curve (AUC) provides a weaker account of accounting decision value than evidence using cost-sensitive evaluation.

TCM plays a complementary role by documenting how the evidence is produced: which theories frame the problem, which accounting contexts are examined, and which methods support the claims. Together, ADO and TCM allow us to identify dominant applications as well as mismatches, such as nontemporal validation in time-dependent settings or explainability artifacts reported without evidence of reviewer use.

III. METHODOLOGY

Research Design

We conducted a hybrid review that combines performance bibliometrics with a framework-based qualitative synthesis. The design follows the structured literature review (SLR) tradition in accounting: a predefined protocol, a

delimited corpus, comparable evidence extraction, and triangulation between quantitative mapping and qualitative coding to improve transparency and replicability (Massaro, Dumay, and Guthrie 2016). ADO structures the substantive synthesis by linking adoption conditions, ML design and validation decisions, and reported outcomes, whereas TCM documents how the evidence is produced in terms of theory, context, and method (Paul et al. 2017; Paul and Benito 2018). Thus, TCM is used as a complementary review lens rather than as a new conceptual contribution.

To strengthen transparency, we specified the protocol *ex ante* and reported study selection through a PRISMA-style flow diagram (Page et al. 2021). The design mirrors recent hybrid reviews that combine bibliometric mapping with structured coding to move from field description to critical synthesis and agenda development (Calandra, Wang, Cane, and Alfiero 2023).

The methodology proceeds in four stages: corpus definition and screening; search-string and eligibility specification; bibliometric mapping of the field's intellectual and thematic structure; and article-level ADO-TCM coding for the framework-based synthesis.

Data Collection

We assembled the dataset from the WoSCC, using a single source to ensure replicability, stable metadata, and a controlled peer-reviewed journal corpus. This choice is appropriate for bibliometric mapping because co-citation, keyword, source, and country analyses are sensitive to duplicates, inconsistent indexing, and heterogeneous metadata across databases. WoSCC also supports transparent filtering by document type, subject category, language, and index and allows consistent export into Bibliometrix/Biblioshiny (Aria and Cuccurullo 2017). To capture relevant accounting information systems research, we included both the Social Sciences Citation Index (SSCI) and the Emerging Sources Citation Index (ESCI) and retained the WoS categories Business, Business Finance, and Management (De Filippo and Gorraiz 2020; Prancutè 2021; Martín-Martín, Orduna-Malea, Thelwall, and Delgado López-Cózar 2018). We also checked the corpus against key accounting and AIS outlets, including the *International Journal of Accounting Information Systems*, the *Journal of Emerging Technologies in Accounting*, the *Journal of Information Systems*, and major accounting journals such as *The Accounting Review*, *Contemporary Accounting Research*, *Accounting Horizons*, and the *Journal of Accounting Research*. We therefore treat WoSCC as a high-quality and reproducible corpus for mapping peer-reviewed ML accounting research, while recognizing that it does not cover the full interdisciplinary literature indexed elsewhere.

Procedure

The search strategy was designed to identify machine learning studies with an explicit accounting, auditing, or accountability focus. We searched WoS Topic fields, which include title, abstract, author keywords, and Keywords Plus, using the following query:

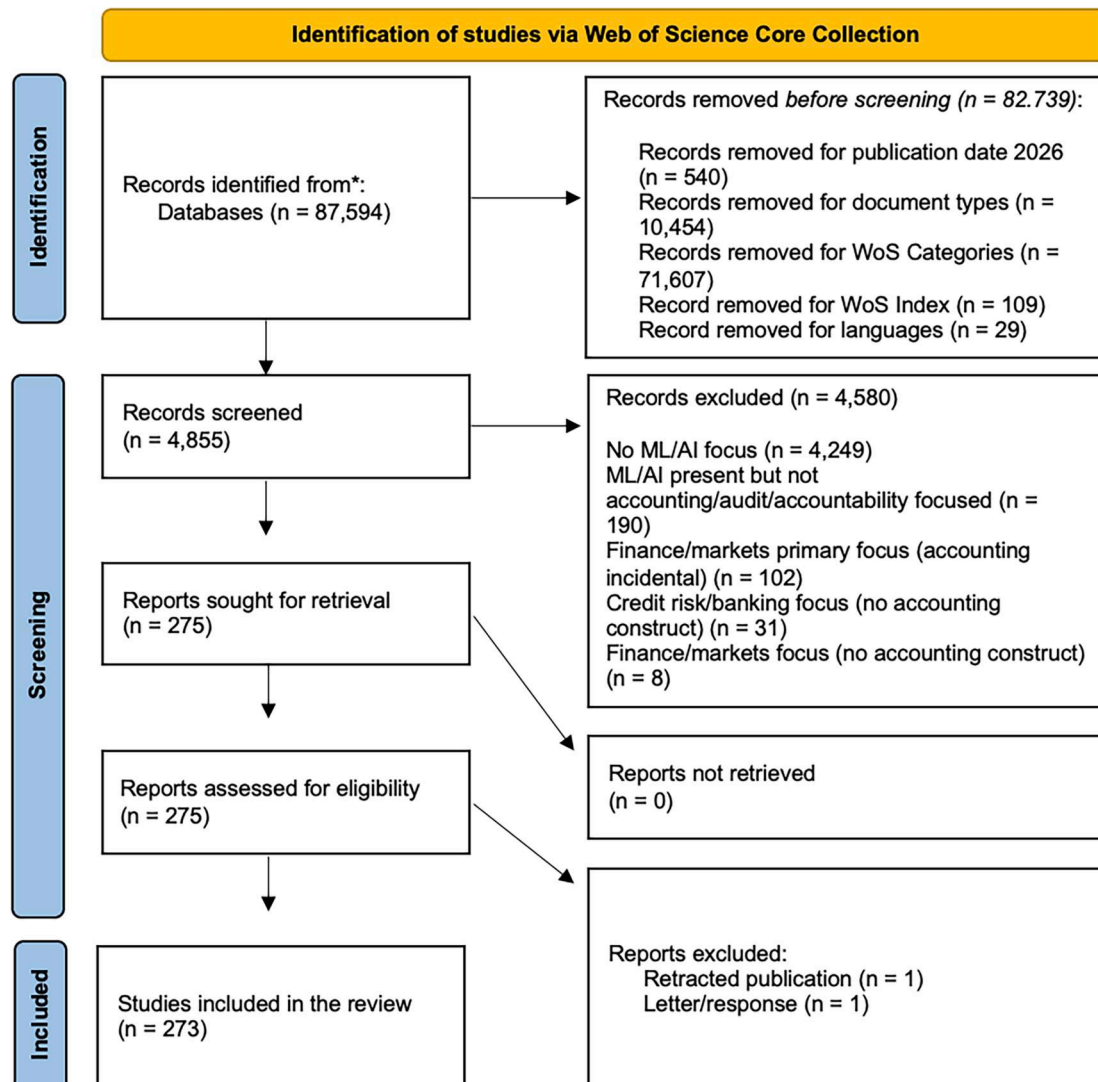
TS= (("machine learning" OR "deep learning" OR "supervised learning" OR "unsupervised learning" OR "reinforcement learning" OR "neural network*" OR "random forest*" OR "support vector*" OR "gradient boosting" OR "xgboost" OR "forecast*" OR "predict*") AND ("accounting" OR "auditing" OR "accountability" OR "accountant*" OR "auditor*"))

The search was run in January 2026, but the analytical window was closed on December 31, 2025. We excluded 2026 records to avoid partial-year bias in annual production and trend analyses.

The initial query returned 87,594 records. We then applied sequential filters: publication years excluding 2026 ($n = 87,054$); Articles and Early Access only ($n = 76,600$); WoS subject categories Business, Business Finance, and Management ($n = 4,993$); SSCI and ESCI indices ($n = 3,573 + 1,311 = 4,884$); and English language only ($n = 4,855$).

Two authors independently screened titles and abstracts against the review scope: ML applied to accounting, auditing, accountability problems, constructs, or settings. Disagreements were resolved through discussion and, when needed, third-reader arbitration. Records retained after title and abstract screening underwent full-text eligibility checks. At this stage, we excluded studies using ML solely for nonaccounting financial-market prediction, nonscholarly sources, and papers that did not critically assess ML methods. Screening decisions were logged record by record in a structured spreadsheet using standardized decision codes and exclusion rationales. The final analytical sample includes 273 papers for bibliometric and framework-based qualitative synthesis. The broad initial query, including terms such as "forecast*" and "predict*," was intentionally inclusive to avoid missing early or methodologically diverse studies that did not use the label "machine learning." Online Appendix Table A1 reports the exclusion rationales applied during screening.

FIGURE 1
Review Workflow and Study Selection Process



Source: Adapted from Page et al. (2021). This work is licensed under CC BY 4.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

* Consider, if feasible to do so, reporting the number of records identified from each database or register searched (rather than the total number across all databases/registers).

(The full-color version is available online.)

Figure 1 summarizes the selection and analytical workflow, linking PRISMA-style screening to the bibliometric outputs, article-level ADO-TCM coding, and framework-based synthesis.

Studies were eligible if they explicitly addressed accounting, auditing, or accountability questions; applied or critically assessed ML as part of the methodological framework; were peer-reviewed journal articles in English; and reported sufficient methodological detail for bibliometric extraction and qualitative coding. We excluded conference proceedings, books and chapters, letters without empirical or theoretical substance, retracted publications, and studies in which ML was peripheral to accounting. Online Appendix Table A1 summarizes the exclusion rationales.

Bibliometric Analysis

For the bibliometric strand, we used Bibliometrix/Biblioshiny (Aria and Cuccurullo 2017) to extract and analyze standard fields (sources, authors, citations, countries, keywords) and to produce science-mapping outputs (e.g., co-citation networks to identify intellectual foundations; and keyword co-occurrence maps to delineate thematic clusters). Parameter settings and thresholds (e.g., minimum co-occurrence frequency) were held constant across analyses to preserve comparability; detailed specifications are provided in the [Online Appendix](#). In addition, we used the “Biblio AI” extension, powered by Google Gemini (Gemini 2.5 Flash model), only as an auxiliary tool for initial keyword-cluster interpretation and contextual summarization. Final topic labels and trend groups were not taken directly from the AI output. They were manually reviewed, harmonized, and aligned with prior accounting and ML review categories, distinguishing substantive accounting domains from generic field descriptors and method labels. The bibliometric analysis provides the field-level map used to answer RQ1 and to identify the thematic clusters examined more diagnostically in the subsequent ADO-TCM synthesis.

Framework-Based Coding

For the qualitative strand, we used an ADO-TCM coding scheme. For each article, we coded antecedents, decisions, and outcomes, and recorded its theory, context, and method to characterize how the evidence was produced. The ADO fields captured adoption drivers, design and validation choices, and reported outcomes; the TCM fields captured theoretical lens, empirical setting, data source, methodological design, ML task, validation protocol, and use of explainable AI (XAI). Two researchers independently coded the articles using a shared codebook. Initial agreement was 91.6 percent across 1,638 article-level ADO-TCM coding decisions,¹ with 138 disagreements. Disagreements were resolved through discussion; 17 cases (1.0 percent) required third-reader arbitration before reaching consensus.

The move from coding to synthesis followed five steps. First, each article was assigned article-level ADO codes to identify the accounting antecedent, the ML design or validation decision, and the reported outcome. Second, each article was coded for TCM to record the theoretical framing, accounting context, and methodological design supporting the evidence. Third, after coder reconciliation, we constructed a consensus coding matrix and grouped repeated ADO combinations within each context. Fourth, we assessed alignment between antecedents and decisions; for example, we assessed whether fraud and misstatement studies used cost-sensitive metrics, whether time-dependent accounting tasks used temporal validation, and whether explainability artifacts supported reviewer judgment rather than merely describing tasks model outputs. Fifth, we aggregated these patterns into the synthesis reported in [Section V](#), distinguishing recurring alignments, misalignments, and future research gaps.

The ADO structure supports a cumulative narrative that links why ML is adopted in accounting (A), how researchers and practitioners design and validate ML systems (D), and what theoretical and practical impacts ensue (O). This framework-based synthesis is consistent with prior knowledge-mapping work that embeds ADO within hybrid reviews to identify robust gaps and future research pathways and aligns with recent accounting SLR guidance (Massaro et al. 2016). The resulting ADO-TCM matrix provides the bridge between the descriptive bibliometric results in [Section IV](#) and the diagnostic synthesis in [Section V](#).

The complete article-level coding matrix, including year, authors, source, DOI, theoretical lens, and methodological classification, is reported in [Online Appendix Table A2](#).

IV. RESULTS: BIBLIOMETRIC ANALYSIS

The bibliometric analysis maps the corpus’s growth, venues, author structure, influential contributions, thematic evolution, conceptual clusters, and collaboration geography. These results answer the mapping component of RQ1 and provide the field-level structure interpreted through ADO-TCM coding in [Section V](#).

Overview of the Dataset

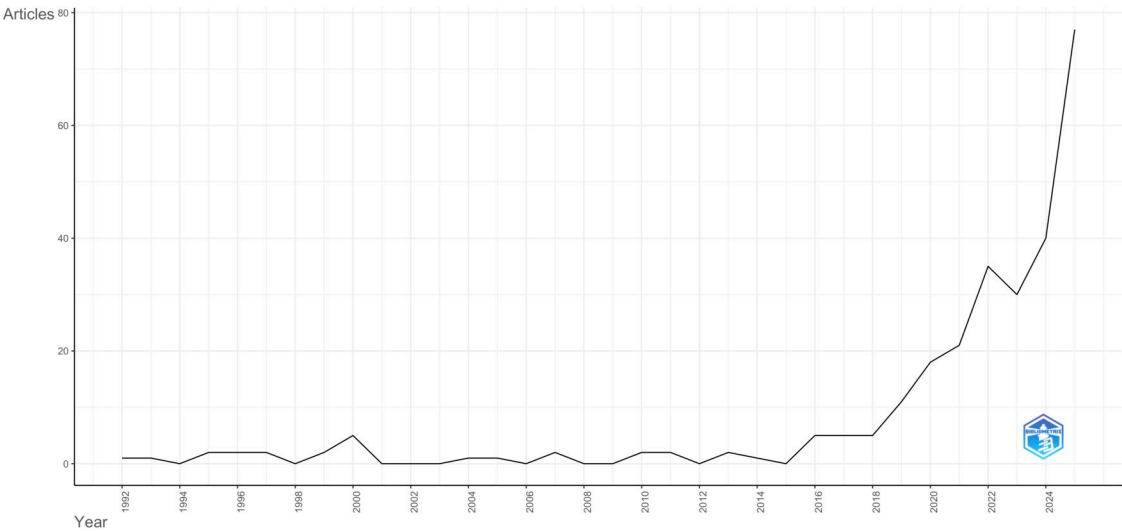
[Table 1](#) summarizes the dataset: 273 documents published between 1992 and 2025 across 125 sources. The annual growth rate of 14.07 percent confirms sustained expansion of ML accounting research ([Figure 2](#)), whereas the average of 21.16 citations per document indicates solid academic impact in a relatively young field. Collaboration is also prominent: 753 authors contributed to the corpus, with 3.04 coauthors per paper and 26.37 percent international collaboration. This profile indicates a sufficiently large and interdisciplinary corpus for examining how ML is used across accounting and assurance problems.

¹ The 1,638 coding decisions are calculated as 273 articles multiplied by six article-level coding fields: antecedents, decisions, outcomes, theory, context, and method. The count refers to the coding fields compared between the two independent coders before reconciliation; it does not double-count the two coders’ entries.

TABLE 1
Main Information

Description	Results
Main information about data	
Timespan	1992–2025
Sources (journals)	125
Documents	273
Annual growth rate, %	14.07
Document average age	5.22
Average citations per document	21.16
Document contents	
Keywords Plus (ID)	649
Authors' keywords (DE)	1,000
Authors	
Authors	753
Authors of single-authored docs	41
Authors' collaboration	
Coauthors per document	3.04
International coauthorships, %	26.37
Document types	
Article	250
Article; early access	23

FIGURE 2
Annual Scientific Production



Source: Authors' elaboration using Bibliometrix.
(The full-color version is available online.)

TABLE 2
Source Analysis

Sources	Articles
<i>International Journal of Accounting Information Systems</i>	18
<i>Journal of Emerging Technologies in Accounting</i>	13
<i>Journal of Information Systems</i>	12
<i>Accounting and Finance</i>	7
<i>The Accounting Review</i>	7
<i>Intelligent Systems in Accounting Finance and Management</i>	7
<i>Accounting Horizons</i>	6
<i>Journal of Accounting Literature</i>	6
<i>Contemporary Accounting Research</i>	5
<i>Decision Sciences</i>	5

Source Analysis

Table 2 shows that AIS-oriented outlets are the main publication hubs: the *International Journal of Accounting Information Systems*, the *Journal of Emerging Technologies in Accounting*, and the *Journal of Information Systems* lead the corpus. ML work also appears in mainstream accounting journals and decision science outlets, indicating diffusion beyond the AIS. This venue pattern matters because the evidence base spans implementation-oriented research as well as studies on reporting quality, fraud, and distress prediction. A complete journal list is reported in [Online Appendix Table A3](#).

Authors' Analysis

Author productivity is moderately concentrated (Table 3). M. A. Vasarhelyi is the most prolific contributor, followed by Z. Zhang and H. K. Duan; however, the top ten authors account for only 38 publications, or 13.9 percent of the corpus. The field therefore has a visible core of repeat contributors without single-author dominance.

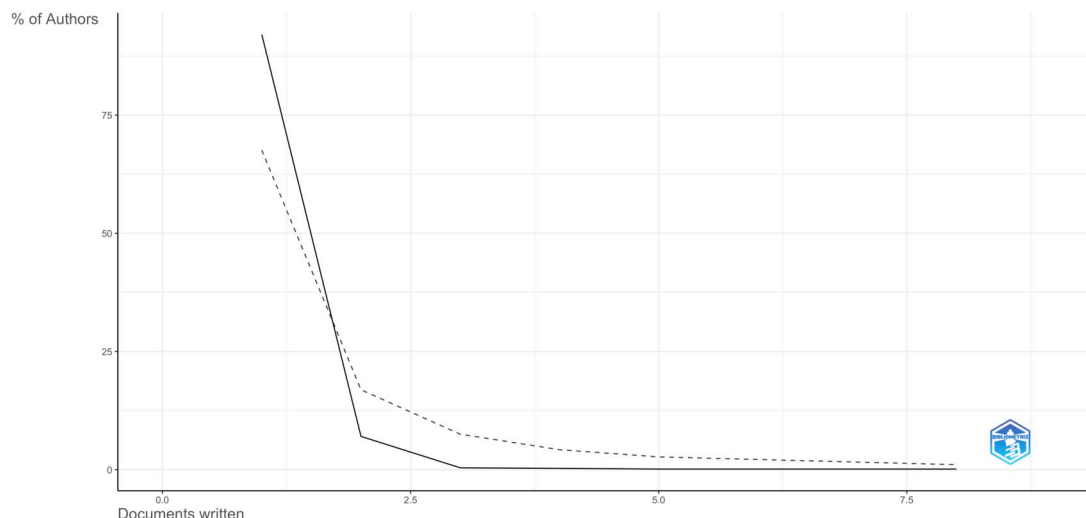
The productivity distribution also aligns with Lotka's law (Figure 3): a small set of authors publishes repeatedly, whereas most contribute occasionally. This pattern suggests a field that is consolidating around a recognizable core while remaining open to varied theoretical, contextual, and methodological choices.

TABLE 3
Most Relevant Authors

Paper	Articles	Global Bibliometric Variables of the Authors		
		Number of Publications	Citations	H-Index
Vasarhelyi, M. A.	12	311	11,821	53
Zhang, Z.	5	8	50	4
Duan, H. K.	4	22	258	7
Cho, S.	3	32	379	7
Gu, H.	3	8	71	2
Wu, Y.	3	281	4,239	33
Alam, P.	2	87	1,421	19
Bansal, M.	2	20	265	11
Bertomeu, J.	2	91	1,668	22
Cao, J.	2	45	581	13

Source: Authors' elaboration using Bibliometrix and OpenAlex (date of analysis February 8, 2026).

FIGURE 3
Author Productivity through Lotka's Law



Source: Authors' elaboration using Bibliometrix.
 (The full-color version is available online.)

Citations Analysis

Table 4 shows that the most cited contributions combine predictive applications with governance and professional concerns. Influential work includes ML-based fraud and misstatement detection (Bao et al. 2020; Bertomeu, Cheynel, Floyd, and Pan 2021), ethical implications of AI-enabled auditing (Munoko, Brown-Liburd, and Vasarhelyi 2020), human-algorithm collaboration (Grønsund and Aanestad 2020), and early neural-network applications to distress and fraud risk (Coats and Fant 1993; Green and Choi 1997; Yang et al. 1999). This citation structure shows that the field's influential work already links prediction with usability, defensibility, and governance in accounting settings.

Trend Topics

Figure 4 reports thematic evolution using harmonized article-level topic groups rather than raw keyword labels. We excluded generic descriptors such as “accounting,” “machine learning,” “artificial intelligence,” and “algorithm(s),” and merged overlapping method labels such as “neural network,” “artificial neural network,” and “deep learning.” The resulting groups distinguish substantive domains: distress and going-concern prediction; fraud, misstatement, and reporting quality; audit analytics and assurance; textual disclosure and NLP; earnings management and governance; AIS, process mining, and continuous monitoring; and management accounting, education, and adoption.

The trend analysis shows a shift from early distress, bankruptcy, and neural-network forecasting work toward fraud, reporting quality, audit analytics, and textual disclosure. Since 2020, the field has become more explicitly connected to assurance and reporting problems, especially through unstructured disclosures and process-level data. These domains anchor the ADO-TCM synthesis in Section V.

Conceptual Structure

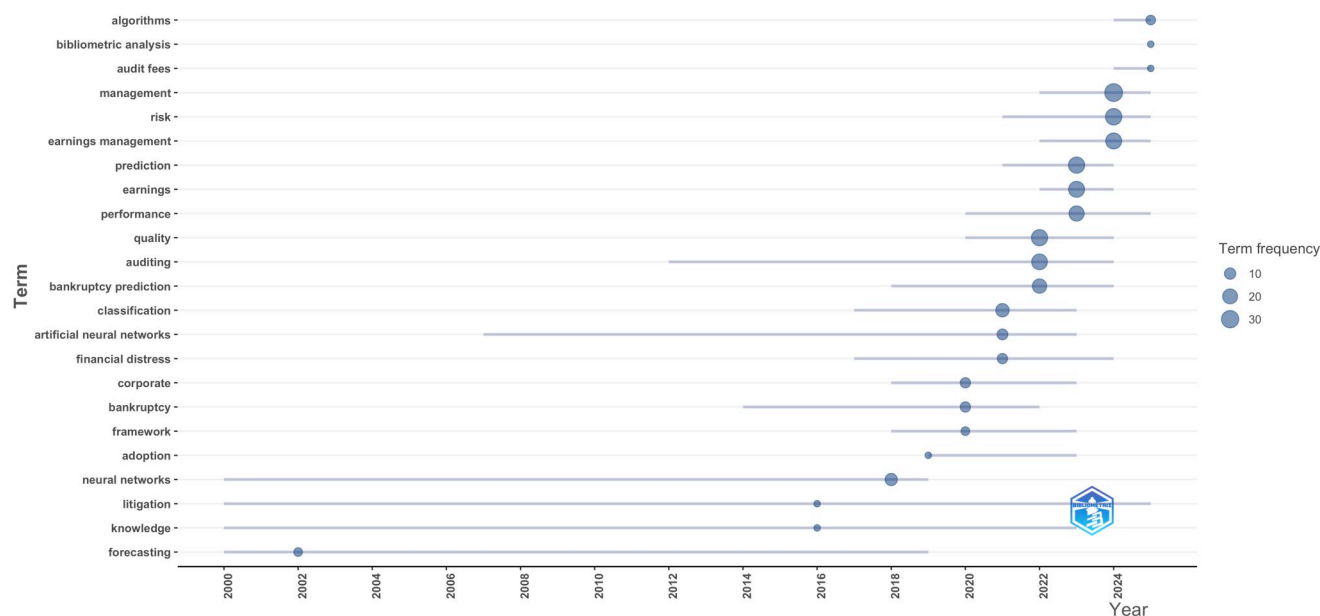
Figure 5 maps the field through an all-keywords co-occurrence network. “Machine learning” is the central hub, with “artificial intelligence” as a secondary hub. The main communities capture AI-enabled accounting and auditing systems, predictive modeling and risk assessment, reporting outcomes and governance mechanisms, and disclosure/textual analysis. The structure is modular but interconnected through auditing, risk, and prediction, showing the diffusion of ML across assurance, distress prediction, reporting quality, and disclosure analytics.

TABLE 4
Most Global Cited Documents

Author(s)/Year/Journal	Title and DOI	Total Citations	TC per Year	Normalized TC
Bao et al. (2020). <i>Journal of Accounting Research</i>	Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach (https://doi.org/10.1111/1475-679X.12292)	241	34.43	3.40
Munoko et al. (2020). <i>Journal of Business Ethics</i>	The ethical implications of using artificial intelligence in auditing (https://doi.org/10.1007/s10551-019-04407-1)	226	32.29	3.19
Gronlund and Aanestad (2020). <i>The Journal of Strategic Information Systems</i>	Augmenting the algorithm: Emerging human-in-the-loop work configurations (https://doi.org/10.1016/j.jsis.2020.101614)	225	32.14	3.17
Coats and Fant (1993). <i>Financial Management</i>	Recognizing financial distress patterns using a neural network tool (https://www.jstor.org/stable/3665934)	193	5.68	1.00
Perols (2011). <i>Auditing: A Journal of Practice & Theory</i>	Financial statement fraud detection: An analysis of statistical and machine learning algorithms (https://doi.org/10.2308/ajpt-50009)	186	11.63	1.82
Yang et al. (1999). <i>Journal of Business Research</i>	Probabilistic neural networks in bankruptcy prediction (https://doi.org/10.1016/S0148-2963(97)00242-7)	155	5.54	1.91
Dong, Liao, and Zhang (2018). <i>Journal of Management Information Systems</i>	Leveraging financial social media data for corporate fraud detection (https://doi.org/10.1080/07421222.2018.1451954)	142	15.78	2.73
Brown, Crowley, and Elliott (2020). <i>Journal of Accounting Research</i>	What are you saying? Using topic to detect financial misreporting (https://doi.org/10.1111/1475-679X.12294)	141	20.14	1.99
Bertomeu et al. (2021). <i>Review of Accounting Studies</i>	Using machine learning to detect misstatements (https://doi.org/10.1007/s11142-020-09563-8)	139	23.17	4.25
Green and Choi (1997). <i>Auditing: A Journal of Practice & Theory</i>	Assessing the risk of management fraud through neural network technology (DOI not available)	136	4.53	1.81

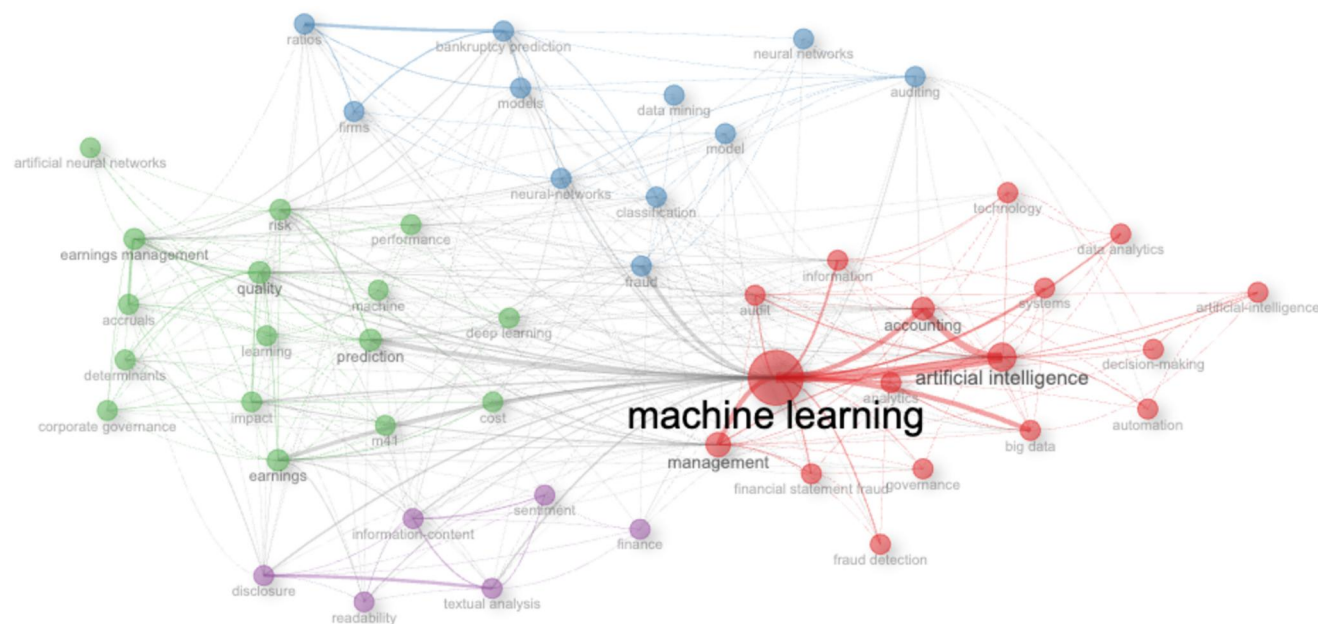
Source: Authors' elaboration using Bibliometrix.

FIGURE 4
Trend Topics



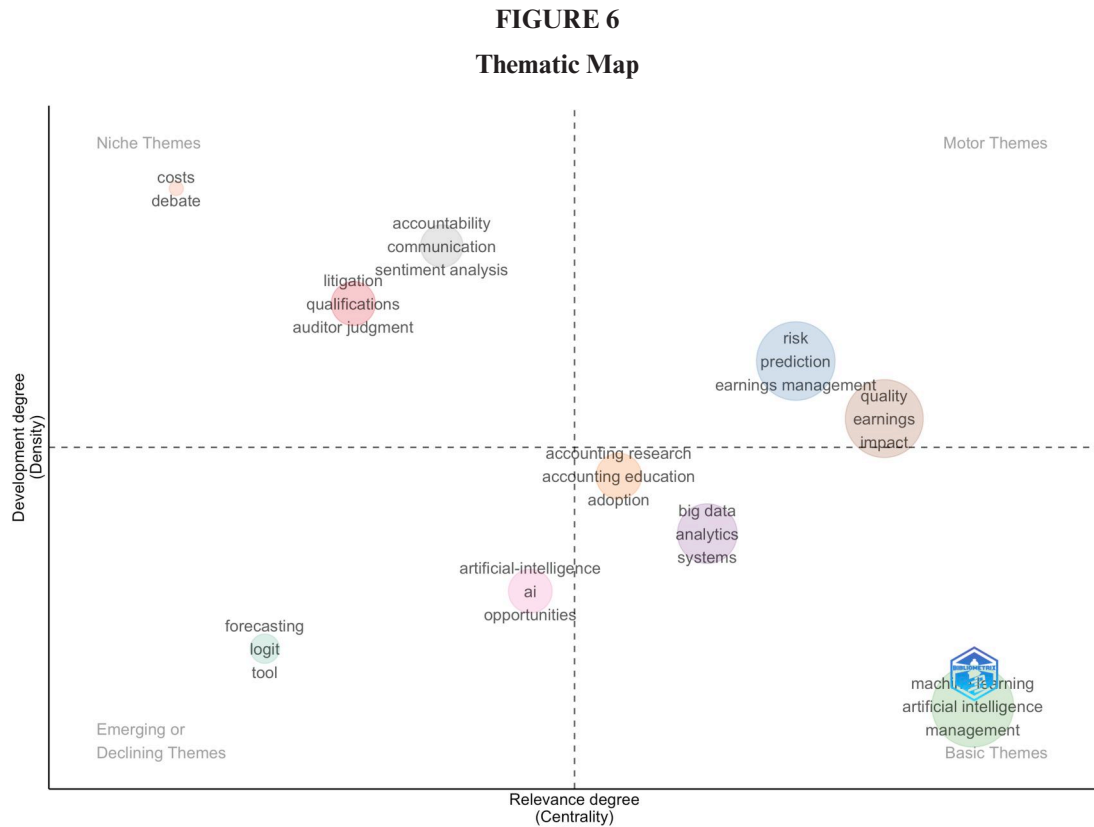
Source: Authors' elaboration using Bibliometrix.
(The full-color version is available online.)

FIGURE 5
Co-Occurrence Network



Source: Authors' elaboration using Bibliometrix.

Each dot (node) is a keyword, and the dot size is proportional to the keyword's occurrence frequency in the corpus (i.e., how often the term appears). The lines (edges) represent keyword co-occurrences (two terms appearing together in the same paper's keyword set). (The full-color version is available online.)



Source: Authors' elaboration using Bibliometrix.
(The full-color version is available online.)

Figure 6 shows that predictive risk assessment and reporting outcomes are the main motor themes. AI/ML, Big Data, analytics, and systems operate as basic themes, whereas accountability, sentiment and communication, auditor judgment, litigation, and cost debates remain more peripheral. This pattern indicates that governance and judgment issues are present but not yet fully integrated with the core prediction agenda.

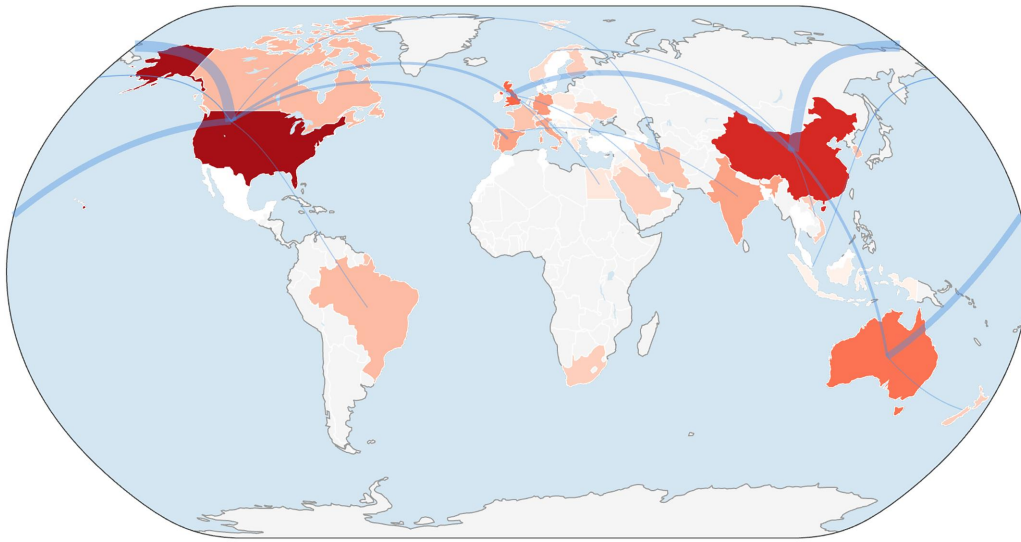
Figure 7 maps cross-country coauthorships. The United States is the main hub, with strong links to China, the United Kingdom, Canada, Australia, and Western Europe. China is the second pole, whereas India, Japan, South Korea, Brazil, and South Africa appear as secondary contributors. Parts of Africa, Central Asia, and Latin America remain less visible, providing a boundary condition for the TCM synthesis.

Together, the bibliometric results show that ML accounting research has expanded rapidly, is concentrated in AIS and accounting outlets, and is organized around prediction, fraud and reporting quality, audit analytics, textual disclosure, and governance. Section V builds on this map by using ADO-TCM coding to explain why ML is adopted, how design choices are made, and when reported outcomes are decision-relevant.

V. RESULTS: ADO-TCM SYNTHESIS

Building on the bibliometric mapping, we now trace how accounting conditions drive ML adoption and shape methodological choices. The bibliometric analysis identified the main application domains and thematic clusters; the ADO-TCM synthesis explains how those domains operate at the article level. The ADO coding identifies the accounting antecedents associated with each study, the ML design and validation decisions used in response, and the outcomes claimed. The TCM coding records the theoretical lenses, empirical contexts, and methods through which this evidence is produced. The purpose of this section is therefore to move from mapping the field to explaining the links between accounting problems, ML choices, and reported value.

FIGURE 7
Countries' Collaboration World Map



Source: Authors' elaboration using Bibliometrix.
 (The full-color version is available online.)

Antecedents (A): Why is Machine Learning Adopted in Accounting?

Across the 273 reviewed studies, four antecedent clusters drive ML adoption in accounting. Each cluster exerts a distinct pressure on the design choices analyzed in the “[Decisions \(D\): How Is ML Designed and Validated in Accounting?](#)” subsection, and identifying these conditional pathways is necessary to move beyond a descriptive taxonomy of “drivers.” The synthesis presented here specifies which antecedent conditions tend to elicit which methodological responses, and why these alignments matter for the outcomes reported in the literature.

The first cluster concerns assurance and compliance pressures, which dominate the audit subset of the corpus (113 studies, 41.4 percent of the sample). When risk indicators are numerous, heterogeneous, and time-varying, ML is mobilized to operationalize complex risk profiles and improve the *ex ante* targeting of audit effort ([Parker et al. 2025](#); [Wang et al. 2025](#); [Wu and Dull 2025](#)). This antecedent generates an asymmetric error-cost structure, since false negatives in fraud detection carry materially higher costs than false positives. In principle, this should push design toward cost-sensitive evaluation, but the empirical pattern indicates otherwise. Although 21 percent of fraud-focused studies adopt rare-event handling techniques such as synthetic minority over-sampling technique (SMOTE) or class weighting, only 3 percent of audit-context studies and none of the fraud-context studies adopt explicitly cost-sensitive evaluation metrics. Recognition of data scarcity has not yet translated into evaluations aligned with decision costs, a failure we examine in detail in the decisions subsection. This first cluster, therefore, shows how an accounting condition, assurance pressure, creates a methodological requirement that is only partially reflected in current ML evaluation practices.

The second cluster reflects the growth of high-dimensional unstructured information. It emerges as the dominant antecedent in 30 studies on narrative disclosure and audit-text analytics, and it appears pervasively across the 273 studies that incorporate textual or process-trace data. Expanded audit reporting (Key Audit Matters, Critical Audit Matters), MD&A narratives, and process-log artifacts produce signals that traditional analytic procedures cannot consistently extract ([Fulcer et al. 2025](#); [Xie and Li 2026](#); [Hayes and Boritz 2021](#)). This antecedent imposes two specific design pressures. First, it requires representation-learning pipelines such as transfer learning or embeddings. Second, it makes auditability difficult to preserve when intermediate features are not human-interpretable. For example, NLP-based disclosure studies can extract incremental information from narrative text, but this benefit creates an accompanying need to document how textual features are represented and how model outputs can be reviewed. The empirical record shows that explainability artifacts remain underadopted in this stream (3 percent in the text/disclosure subset, against 12 percent in distress prediction), suggesting that the contexts most reliant on opaque representations are also

those least equipped to document model reasoning for assurance purposes. This antecedent, therefore, directly links to later design choices regarding representation, interpretability, and reviewability.

The third cluster centers on early-warning prediction under nonlinear and opaque accounting signals. It characterizes 17 studies in financial distress, bankruptcy prediction, and going-concern contexts (Atman and Lamarque 2025; Lin, Padliansyah, Lu, and Liu 2025; Lee and Tahmoush 2025). This antecedent has been recognized longest in the literature, dating to the foundational work of Coats and Fant (1993) and Yang et al. (1999). Its temporal nature, which involves predicting future states from current accounting data, imposes a clear design requirement: validation must reflect the time-ordered structure of the information flow. This antecedent leads to design decisions around out-of-time validation, rolling windows, and robustness to changes in economic regimes. The empirical pattern is informative. Distress studies show the highest temporal-validation rate in the corpus (12 percent against 1 percent overall), reflecting a partial but uneven recognition that out-of-time generalization is the substantive criterion of interest. Even so, 88 percent of distress studies still rely on cross-sectional or nontemporal validation, despite the regime-shift sensitivity well documented in the corporate-failure literature. This cluster therefore shows that, even in the most mature predictive stream, validation choices remain unevenly aligned with the accounting problem.

The fourth cluster involves process digitalization, continuous-control infrastructures, and accountability or responsible-AI concerns (Xiong, Peng, Xiao, and Yang 2025; Otero and Agu 2025; Ashraf 2025; Munoko et al. 2020). This antecedent is qualitatively different from the previous three because it shifts the focal mechanism from “ML as predictor” to “ML as organizational artifact requiring governance.” It motivates documentation, audit-trail design, and human-algorithm review protocols. For example, process-level audit applications depend not only on whether the model flags suspicious transactions but also on whether the model’s inputs, transformations, thresholds, and reviewer actions can be reconstructed after the fact. As the Outcomes subsection develops, this is also the antecedent most consequential for whether reported predictive gains translate into durable assurance value, since governance shortfalls are repeatedly identified as the binding constraint on deployment. The four clusters are neither mutually exclusive nor uniformly distributed across the corpus. Their substantive relevance lies in what they predict. Each antecedent generates a specific design pressure that, in turn, conditions which design choices are theoretically appropriate and which boundary conditions later govern the realization of outcomes. Therefore, the antecedents show that ML adoption in accounting is not driven solely by technology but by accounting conditions that impose distinct evaluation, validation, and governance requirements. The next subsection explicitly develops these antecedent-decision linkages.

Decisions (D): How is ML Designed and Validated in Accounting?

Design decisions in the corpus are best understood as conditional responses to the antecedent pressures identified above, rather than as a flat enumeration of methodological choices. Three analytical dimensions structure this synthesis. The first concerns how task framing and evaluation respond to asymmetric error costs. The second concerns how validation design responds to the temporal structure of accounting evidence. The third concerns how explainability functions, or fails to function, as an accountability mechanism rather than a descriptive feature. In each case, the relevant question is whether the ML decision matches the accounting condition that motivated adoption.

First, task framing and evaluation often fail to reflect asymmetric error costs. Across the corpus, 112 studies (41.0 percent) report performance using accuracy, F1, AUC, precision, or recall as primary criteria. Only 28 studies (10.3 percent) adopt cost-sensitive evaluation or expected-cost frameworks in which false-negative and false-positive errors carry differential weights aligned with audit or assurance objectives. The disproportion becomes striking when juxtaposed with the antecedent structure. Fraud and misstatement contexts (33 studies) display the highest rate of accuracy-only reporting (36 percent), even though these are settings in which the cost of missing a true fraud is materially asymmetric to the cost of a false alarm. Three illustrative cases stand out among the studies that explicitly engage with this asymmetry.

Wu and Dull (2025) frame misclassification costs explicitly in the nonprofit discontinuance setting, motivated by “donors’ asymmetric losses from false negatives versus false positives.” Tragouda, Doumpos, and Zopounidis (2024) implement cost-sensitive weighting for class imbalance with rolling temporal validation in financial-statement fraud, combining both substantive design responses to the antecedent structure.

Huang and Wang (2023) prioritize recall, precision, and misclassification cost over accuracy in an audit-pedagogy setting and explicitly frame model selection as a professional judgment. These cases stand in contrast to a majority of fraud and audit studies that report area under the curve as the headline metric, even when threshold selection is left implicit. The implication for cumulative knowledge is direct: when evaluation does not embed the cost structure imposed by the antecedent, reported performance is decoupled from the decision value the model is intended to produce. This decoupling offers a structural explanation for why claims of “predictive lift” tend to attenuate when reinterpreted for assurance use.

Thus, the first decision pattern links assurance and fraud antecedents to a recurring methodological gap: evaluation metrics often assess classification performance without accounting for accounting decision value.

Second, the validation design is unevenly aligned with the temporal structure of accounting evidence. The corpus shows a comparable disconnect between antecedent and design. Only 35 studies (12.8 percent) explicitly employ blocked, out-of-time, or rolling-window validation, and 21 studies (7.7 percent) explicitly engage with class imbalance and rare-event modeling. Joint adoption is rarer still. Audit and assurance studies, for which fiscal-year boundaries, regime shifts, and standard cycles create strong temporal dependence, show only a 2 percent adoption rate for temporal validation. The distress and bankruptcy subset, by contrast, shows 12 percent adoption. Distress prediction has historically internalized the temporal logic of the prediction problem, and the study by [Serrano-Cinca, Gutiérrez-Nieto, and Bernate-Valbuena \(2019\)](#) illustrates the canonical practice, namely pairing failed and healthy firms by industry and size for training and then validating out-of-time on a later-period holdout. Audit and assurance studies more frequently rely on stratified random splits, which can lead to data leakage when an engagement-year is shared across the train and test partitions. The empirical pattern thus reveals an asymmetry. Temporal-validation rigor is highest in the antecedent context where its absence has been most repeatedly criticized in the literature ([Bertomeu 2020](#)) but remains lowest in the antecedent contexts (audit, fraud, text and disclosure) where claim defensibility most depends on robust out-of-sample evidence. This second decision pattern links temporal accounting problems to a validation gap: many studies test predictive performance as if accounting observations were exchangeable, even when the task is explicitly time ordered.

Third, explainability is more often reported than evaluated as an accountability mechanism. The corpus contains 22 studies (8.1 percent) that incorporate explainability practices into the model-design stage and 39 studies (14.3 percent) that report explainability artifacts as part of the reported outcome. Treating explainability as a mechanism rather than a feature requires asking what work the artifact does in the assurance chain. The relevant question is whether feature attributions, SHapley Additive exPlanations (SHAP) values, or interpretable surrogates change how a human reviewer evaluates a flagged observation, documents reasoning in working papers, or escalates exceptions. Three patterns from the coded evidence are instructive. [Zhang et al. \(2022\)](#) explicitly frame the design problem as “pairing black-box models with appropriate XAI techniques for both global tool understanding and engagement-level (local) explanations,” coupling time-based splits with under-sampling to address imbalance. This is an integrated treatment of design, validation, and accountability. [Trinkle and Baldwin \(2016\)](#) treat the performance-interpretability tradeoff as a foregrounded design choice in credit decisions that must be explained to regulators and applicants and illustrate accountability as the binding constraint. [Schneider and Brühl \(2023\)](#) add interpretability components (feature importance, functional form, interactions) specifically to explain why the model flags fraud, thereby moving explainability from accuracy supplement to explanatory artifact. Against these examples, however, the broader corpus is characterized by explainability adopted but not evaluated. Feature-importance rankings are reported, but their effect on reviewer decisions, calibration, or reliance is rarely tested. This third decision pattern links accountability antecedents to an explainability gap: many studies produce explanations, but few test whether those explanations improve human review or documentation.

Three additional design dimensions (data representation, model-family selection, and feature construction) are documented in the corpus but exhibit lower conditional variance across antecedent contexts and are therefore reported here in compressed form. Representation choices increasingly span unstructured data, including NLP and text-mining pipelines in 11 studies and transfer-learning embeddings in disclosure-focused work ([Siano and Wysocki 2021](#)), and process-level data such as deep-learning process mining for suspicious transactions ([Duan et al. 2025](#)). Model families remain dominated by regression-centered designs (102 studies, 37.4 percent) and neural architectures (28 studies, 10.3 percent), with a recurrent performance-auditability tradeoff in which interpretable baselines are retained when reviewability is salient ([Krupa and Minutti-Meza 2022](#)). These additional choices further reinforce the same pattern: ML design in accounting is valuable when it is selected in response to the task’s information structure, error costs, and accountability requirements.

Three propositions summarize the antecedent-decision linkages developed above.

- Proposition 1:** When the antecedent generates asymmetric error costs (fraud, audit, distress), cost-sensitive evaluation is theoretically appropriate but empirically rare; reliance on accuracy-only metrics decouples claimed performance from decision value.
- Proposition 2:** When the antecedent imposes temporal dependence (assurance cycles, distress windows, disclosure regime shifts), blocked or rolling-window validation is appropriate but rare outside the distress subset; nontemporal validation inflates apparent generalization.
- Proposition 3:** When the antecedent demands accountability and reviewability (audit, regulated credit, fraud), explainability must operate as a mechanism whose effect on human review is documented; the corpus more frequently treats explainability as a reported artifact, without evidence of its decision-relevant function.

These propositions clarify how design choices condition the value claims reported in the literature. The next subsection examines how these conditional decisions shape the outcomes reported across the literature.

Outcomes (O): What Machine Learning Delivers in Accounting?

Building on the decision patterns above, the Outcomes analysis examines what ML is reported to deliver when accounting antecedents and methodological choices are aligned. Across the reviewed studies, outcomes concentrate on three recurring value claims: predictive lift, expanded assurance coverage/efficiency, and governable decision support, with clearly stated boundary conditions. First, the most consistent outcome is out-of-sample performance improvement over traditional statistical baselines in high-stakes accounting tasks characterized by nonlinearities and class imbalance. This pattern is evident in fraud and misstatement screening, where ensemble and modern ML pipelines improve identification and prioritization of risky observations (Bao et al. 2020; Bertomeu et al. 2021; Parker et al. 2025). Similar gains are reported when unstructured disclosures are incorporated: transfer-learning/NLP models extract incremental signals from narrative text, improving classification and prediction beyond structured accounting variables alone (Siano and Wysocki 2021; Hayes and Boritz 2021). These outcomes are strongest when ML design responds directly to the information problem, such as when nonlinear risk patterns or text-intensive evidence are present.

Second, a substantial set of papers frames outcomes as audit execution benefits: broader test coverage, faster exception triage, and reallocation of effort from routine sampling to targeted review in journal-entry analytics, process monitoring, or continuous controls (Wei, Cho, Vasarhelyi, and Te-Wierik 2024; Duan et al. 2025). Third, outcomes increasingly include defensibility and transparency through feature attribution and documented validation. Across these outcomes, the boundary condition is consistent: predictive gains do not automatically translate into durable assurance value when temporal dependence, preprocessing choices, leakage safeguards, model-risk controls, or reviewability are underspecified (Zhang et al. 2022; Smith and Lamprecht 2024).

Theories

Theoretical grounding is present but remains thinly distributed across the corpus (Table 5). Of 273 papers, only 94 explicitly state at least one theoretical lens; most adopt a single theory, with occasional multitheory positioning. The most recurrent lenses concentrate around three families. First, incentive- and opportunism-based explanations, agency theory, and fraud-oriented frameworks (notably fraud triangle theory) anchor studies focused on misstatement/fraud risk and governance mechanisms. Second, information and market-interpretation perspectives, mainly signaling theory, are used to frame disclosure-related inference and the informational role of textual/structured reporting signals.

TABLE 5
Papers with at Least One Explicit Theory

Theory Type	Count
Agency theory	8
Fraud triangle theory	7
Signaling theory	7
Technological acceptance model (TAM)	3
Positive accounting theory	3
Actor-network theory	2
Fraud diamond theory	2
Impression management theory	2
Institutional theory	2
Legitimacy theory	2
Real options theory	2
Upper-echelons theory	2
Other single-occurrence theories (long tail; see Online Appendix Table A2 mapping)	55 distinct lenses

Third, a smaller stream relies on technology and sociotechnical adoption lenses (e.g., the Technology Acceptance Model [TAM]) when the focal outcome concerns user acceptance, workflow integration, or professional change. Beyond these recurring frameworks, the field displays a long tail of one-off lenses (e.g., transaction cost, contingency, impression management, actor-network), indicating conceptual pluralism but limited consolidation around a dominant paradigm. Overall, the evidence suggests that theory is still more often employed as contextual framing than as a driver of model design choices (e.g., loss functions, validation, explainability) or as a basis for theorizing accountability-relevant outcomes, leaving room for tighter alignment between theoretical mechanisms and ML design/validation decisions in future research. This theory pattern helps explain why many studies identify useful accounting applications but stop short of theorizing how ML choices should reflect accountability, decision usefulness, or professional judgment. The next subsection reports the context dimension of TCM, showing where these theoretical and methodological choices are empirically concentrated.

Contexts

The context dimension identifies where ML is studied and where the evidence is concentrated. The corpus is dominated by audit and assurance settings, especially fraud, misstatement, anomaly detection, and audit-risk applications. This concentration is consistent with the ADO evidence reported above: assurance contexts give rise to rare-event classification problems, asymmetric error costs, and documentation requirements. A second major context concerns financial accounting and reporting, including distress prediction, bankruptcy, earnings management, disclosure quality, restatements, and textual reporting analysis. These studies often rely on firm-level archival data and are more likely to frame ML as a predictive or classification tool. A smaller but growing stream examines AIS and process-level settings, where enterprise resource planning logs, transaction traces, and continuous-monitoring data make ML relevant for control testing and process-aware auditing. Managerial accounting, public-sector accounting, small and medium enterprises (SMEs), and education remain underrepresented, although recent studies indicate growing interest in these areas. Geographically, the evidence is concentrated in North America, Western Europe, and East Asia, with thinner coverage of Africa, Central Asia, and parts of Latin America. Overall, the context coding shows that the literature is strongest where large archival, textual, or process data are available, whereas contexts requiring field access, organizational process evidence, or public accountability data remain underdeveloped. This contextual concentration matters for interpreting outcomes: ML evidence is most developed where data are accessible and machine-readable, but weaker where accounting work depends on organizational processes, professional interaction, and public accountability settings that are harder to observe. The next subsection, therefore, examines whether the methods used in the corpus are sufficient to support both predictive and behavioral claims.

Methods

Methodologically, the literature is dominated by empirical quantitative designs, with a clear prevalence of regression-based setups and predictive benchmarking, reflecting the field's emphasis on out-of-sample performance and risk-focused inference (Table 6). In our sample, 102 papers use regression-centered empirical designs, often as a baseline or as the main analytic approach in fraud/misstatement and distress-related settings. A second, sizeable stream relies on neural-network architectures (28 papers), which are typically motivated by nonlinearities and high-dimensional inputs and increasingly include unstructured data. Complementing these approaches, NLP/text-mining pipelines appear in 11 papers, signaling a shift toward disclosure- and narrative-intensive accounting tasks where representation learning provides an incremental signal. Beyond these core empirical traditions, we observe 43 mixed-method contributions and 32 conceptual papers, whereas qualitative designs remain comparatively rare (case studies: seven; interviews: five), typically used to unpack implementation, professional change, and governance mechanisms around ML adoption. Finally, the evidence base includes a limited set of experimental studies (seven), suggesting that causal identification of how decision-makers use (or misuse) ML outputs remains underdeveloped relative to predictive modeling work. Overall, the methodological profile indicates strong progress on predictive validity, alongside opportunities to strengthen explanatory and behavioral evidence through experiments and qualitative inquiry that connect ML outputs to accounting judgment and accountability. Taken together, the TCM evidence reinforces the central ADO finding: the literature has developed strong evidence on where ML can improve prediction but weaker evidence on how ML should be governed, explained, and incorporated into accounting judgment. This conclusion provides the basis for the implications and future research agenda discussed in the next section.

TABLE 6
Methods

Method Type	Count
Empirical—Regression	102
Empirical—Neural networks	28
Empirical—Other ML (unspecified)	13
Empirical—NLP/text mining	11
Empirical—Random forest	4
Empirical—Unsupervised/clustering	2
Empirical—Time series	1
Empirical—Naïve Bayes	1
Empirical—Survey	1
Empirical—Multicriteria	1
Experimental	7
Qualitative—Case study	7
Qualitative—Interviews	5
Qualitative—Thematic analysis	2
Qualitative—Ethnography	1
Mixed-method	43
Conceptual	32
Systematic literature review	8
Bibliometric analysis	4
Total	273 (100%)

VI. DISCUSSION AND CONCLUSION

Discussion

This discussion is organized around the two research questions that guided the review. RQ1 asked what intellectual structure, application domains, and theory-context-method patterns characterize ML research in accounting. RQ2 asked how accounting antecedents shape ML design and validation decisions and how these decisions condition reported outcomes and future research gaps. We answer RQ1 by showing that the literature is concentrated around two application cores—distress/bankruptcy prediction and reporting quality/fraud/misstatement detection—with audit analytics and textual analysis connecting these domains. We answer RQ2 by showing that the main limitation is not the absence of ML applications, but the incomplete alignment between accounting conditions and ML evaluation, validation, and governance choices.

The ADO synthesis further clarifies *why* ML is adopted (antecedents), *how* it is designed and validated (decisions), and *what* it delivers (outcomes). The most frequently observed antecedents relate to (1) increased assurance and compliance pressure, (2) growth in high-dimensional and unstructured information (e.g., narratives, audit reports, system traces), and (3) the diffusion of digital process infrastructures enabling continuous monitoring. Design choices reflect recurring tradeoffs between predictive power and auditability: many studies benchmark flexible learners (ensembles, neural architectures, NLP pipelines) against interpretable baselines, but validation and transparency practices remain uneven across the corpus (e.g., temporal robustness, leakage controls, and documentation of preprocessing and thresholds). Outcomes converge on incremental predictive performance and broader testing scope (often framed as fuller-population analytics), but reported benefits are conditional on disciplined validation, reproducible pipelines, and governance-aligned deployment assumptions.

Two conceptual points follow. First, our review treats ML as the focal construct (a data-driven subset of AI) and includes studies only when the analytical contribution relies on learning-based models (supervised/unsupervised/deep learning, ensembles, and related pipelines). This boundary is important because broad “AI/digitalization” framing can blur the distinction between what is being reviewed and what is inferred from the evidence base. Second, theoretical lenses are present but not dominant: many papers use theory primarily to motivate the substantive setting (e.g., fraud,

distress, disclosure) rather than to inform ML design, evaluation, or governance decisions. This suggests an opportunity to move from “theory as contextual justification” to “theory as mechanism,” explaining why specific design choices (e.g., asymmetric loss, explainability artifacts, human-algorithm review protocols) should matter in accountability-sensitive contexts. The theoretical implication is therefore direct: ML accounting research can contribute more strongly to accounting theory when it explains how accountability, decision usefulness, professional judgment, and institutional constraints shape the design and evaluation of data-driven models, rather than treating these elements as background conditions.

Implications for Research

Responding to RQ2, the ADO synthesis shows that the largest gaps occur when accounting antecedents and ML design choices are misaligned. Methodologically, future work would benefit from (1) temporally robust validation as a norm in settings with regime shifts, (2) explicit treatment of cost asymmetries and thresholding aligned with audit/planning objectives, and (3) reporting standards that make pipelines reproducible (feature construction, preprocessing, leakage safeguards, stability checks). Conceptually, research can tighten alignment between theory and ML practice by linking adoption and use to accountability, decision usefulness, and institutional constraints, and by examining how ML outputs are incorporated into human judgment and review routines (rather than treating explainability as an add-on) (Table 7). These research implications follow directly from the ADO-TCM evidence: stronger ML accounting research will require not only more sophisticated models but also clearer theorization of why a given model, validation design, or explanation format is appropriate for a specific accounting task.

Practical Implications

Because our evidence is drawn from published studies (not from field deployments), implications for practice should be interpreted as design considerations rather than prescriptive policy guidance. In audit and control contexts, the literature most consistently supports three implementable emphases: (1) validation designs that reflect the time structure of accounting settings, (2) documentation and provenance practices that support reviewability, and (3) explanation interfaces that are evaluated on whether they improve reviewer decisions, not only on whether they can be generated. These points follow directly from recurrent boundary conditions reported across the corpus (performance claims weaken when temporal dependence, leakage controls, or documentation are underspecified). For practitioners, the practical message is that ML systems used in accounting and assurance should be assessed as decision-support infrastructure rather than solely as prediction engines. Their value depends on whether model outputs can be validated over time, traced back to reliable data and preprocessing choices, explained to reviewers, and incorporated into accountable professional judgment.

Limitations

This review has four limitations. First, WoSCC supports transparency and consistent metadata but omits relevant work indexed only in Scopus, Google Scholar, conference proceedings, practitioner outlets, or computer-science databases. Second, topic-field searching can miss papers that use ML without foregrounding ML terms. Third, bibliometric thresholds and ADO-TCM coding involve judgment despite dual screening and reconciliation. Fourth, the review synthesizes reported evidence and cannot establish the field effectiveness of ML tools without deployment evidence.

Conclusion

In conclusion, the literature on ML in accounting is expanding and increasingly clustered around distress prediction and reporting integrity/fraud problems, with growing attention to unstructured disclosures and audit analytics. The ADO-TCM synthesis shows that ML adoption is motivated mainly by scalability and information complexity; design decisions hinge on the tradeoff between predictive lift and accountability-oriented constraints; and reported outcomes emphasize improved screening and expanded coverage but are contingent on validation discipline and transparency. Future work can strengthen the field by aligning theoretical mechanisms with ML design choices, standardizing temporally robust and cost-aware evaluation, and producing evidence on how humans incorporate ML outputs into accounting judgments and assurance processes.

The contribution of this review is to connect these pieces in one cumulative framework. The bibliometric analysis maps the development of ML accounting research; ADO explains why ML is adopted, how it is designed, and the outcomes claimed; and TCM shows how theory, context, and method shape the evidence base. This integrated view reframes future research away from asking only whether ML improves prediction and toward asking when ML

TABLE 7

Synthesized Future Research Agenda (ADO-TCM Aligned)

Theme/Gap Area	Specific Research Directions	Potential Theoretical Lenses	Empirical Strategy/Data and Evaluation
Governance and accountability of ML	(1) When do ML controls (documentation, monitoring, drift management) increase assurance value versus create “procedural compliance”? (2) What minimum documentation makes models reviewable in audit workpapers?	Accountability; institutional legitimacy; socio-technical systems	Field/archival + design-science: compare engagements/units with/without governance artifacts; audits of model documentation quality; longitudinal drift tests
Validation under time, rarity, and asymmetric costs	(1) Cost-sensitive calibration and thresholding consistent with audit/planning objectives. (2) Temporal generalization and leakage-robust evaluation as default reporting standard.	Decision usefulness; judgment and decision making	Out-of-time/blocked splits; sensitivity to base rates; expected-cost metrics; robustness to regime shifts; reporting checklists for pipelines
Human-algorithm teaming and explanation use	(1) Do explanations improve reviewer decisions or just increase confidence? (2) When do interpretable-first versus explainable-after designs change accountability outcomes?	Technology adoption (TAM); socio-technical	Experiments with auditors/analysts: decision quality, effort allocation, calibration; compare explanation formats; measure reliance/overreliance
Data expansion beyond financial ratios (text + process logs + multimodal)	(1) Incremental value of narrative/process traces beyond structured accounting data. (2) Replicability of multimodal pipelines across firms/industries.	Information economics; signaling (for disclosure contexts)	Multimodal benchmarks; ablation studies; cross-context portability; shared feature/pipeline reporting; external validation datasets
Under-covered contexts (managerial/public/SMEs) and outcomes	(1) ML in budgeting/control and public-sector accountability. (2) Outcomes beyond accuracy: sustained effort reallocation, error prevention, fairness effects.	Contingency; institutional; public accountability	Comparative designs across contexts; qualitative case + analytics triangulation; measure operational outcomes (cycle time, exception handling, control failures)

produces accounting value: when evaluation reflects decision costs, validation respects time and data structure, explanations support review, and governance practices make model use reproducible and accountable.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

While preparing this work, the authors used the Bibliometrix “Biblio AI” module (powered by Gemini 2.5 Flash) to assist with topic labeling and contextual summarization of keyword clusters within the bibliometric analyses. After using this tool, the authors reviewed and edited all outputs and took full responsibility for the publication’s content. The AI software Grammarly was then used to proofread the entire paper.

REFERENCES

- Amani, F. A., and A. M. F. Fadlalla. 2017. Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems* 24: 32–58. <https://doi.org/10.1016/j.accinf.2016.12.004>
- Aria, M., and C. Cuccurullo. 2017. Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics* 11 (4): 959–975. <https://doi.org/10.1016/j.joi.2017.08.007>
- Ashraf, M. 2025. Does automation improve financial reporting? Evidence from internal controls. *Review of Accounting Studies* 30 (1): 436–479. <https://doi.org/10.1007/s11142-024-09822-y>
- Atman, Z., and E. Lamarque. 2025. Which financial indicators do stakeholders have to focus on to manage risk of bank failure? *Review of Accounting and Finance* 24 (4): 503–550. <https://doi.org/10.1108/RAF-04-2024-0155>
- Bao, Y., B. Ke, B. Li, Y. J. Yu, and J. Zhang. 2020. Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research* 58 (1): 199–235. <https://doi.org/10.1111/1475-679X.12292>
- Bertomeu, J. 2020. Machine learning improves accounting: Discussion, implementation, and research opportunities. *Review of Accounting Studies* 25 (3): 1135–1155. <https://doi.org/10.1007/s11142-020-09554-9>
- Bertomeu, J., E. Cheynel, E. Floyd, and W. Pan. 2021. Using machine learning to detect misstatements. *Review of Accounting Studies* 26 (2): 468–519. <https://doi.org/10.1007/s11142-020-09563-8>
- Bhimani, A., and L. Willcocks. 2014. Digitisation, “big data” and the transformation of accounting information. *Accounting and Business Research* 44 (4): 469–490. <https://doi.org/10.1080/00014788.2014.910051>
- Brown, N. C., R. M. Crowley, and W. B. Elliott. 2020. What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research* 58 (1): 237–291. <https://doi.org/10.1111/1475-679X.12294>
- Calandra, D., T. Wang, M. Cane, and S. Alfiero. 2023. Management of hydrogen mobility challenges: A systematic literature review. *Journal of Cleaner Production* 410: 137305. <https://doi.org/10.1016/j.jclepro.2023.137305>
- Coats, P. K., and L. F. Fant. 1993. Recognizing financial distress patterns using a neural network tool. *Financial Management* 22 (3): 142. <https://doi.org/10.2307/3665934>
- De Filippo, D., and J. Gorraiz. 2020. Is the emerging source citation index an aid to assess the citation impact in social science and humanities? *Journal of Informetrics* 14 (4): 101088. <https://doi.org/10.1016/j.joi.2020.101088>
- Dong, W., S. Liao, and Z. Zhang. 2018. Leveraging financial social media data for corporate fraud detection. *Journal of Management Information Systems* 35 (2): 461–487. <https://doi.org/10.1080/07421222.2018.1451954>
- Duan, H. K., M. A. Vasarhelyi, and M. Codesso. 2025. Integrating process mining and machine learning for advanced internal control evaluation in auditing. *Journal of Information Systems* 39 (1): 55–75. <https://doi.org/10.2308/ISYS-2022-028>
- Fulcer, K., H. Gu, H. Hu, Q. Huang, A. Kogan, M. A. Vasarhelyi, D. Wei, and J. Young. 2025. Application of outlier detection methods in audit analytics. *Accounting Horizons* 39 (3): 143–157. <https://doi.org/10.2308/HORIZONS-2023-071>
- Green, B. P., and J. H. Choi. 1997. Assessing the risk of management fraud through neural network technology. *Auditing: A Journal of Practice & Theory* 16 (1): 14–28. <https://doi.org/10.2308/AJPT-9708303198>
- Grønsund, T., and M. Aanestad. 2020. Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems* 29 (2): 101614. <https://doi.org/10.1016/j.jsis.2020.101614>
- Hardies, K., F. Ohlrogge, J. Mentens, and J. Vandennieuwenhuysen. 2024. A guide for accounting researchers to conduct and report systematic literature reviews. *Behavioral Research in Accounting* 36 (1): 21–43. <https://doi.org/10.2308/BR1A-2022-042>
- Hayes, L., and J. E. Boritz. 2021. Classifying restatements: An application of machine learning and textual analytics. *Journal of Information Systems* 35 (3): 107–131. <https://doi.org/10.2308/ISYS-19-003>
- Huang, F., and Y. Wang. 2023. Introducing machine learning in auditing courses. *Journal of Emerging Technologies in Accounting* 20 (1): 195–211. <https://doi.org/10.2308/JETA-2022-017>
- Janvrin, D. J., R. L. Raschke, G. P. Schneider, and S.-F. P. Hsieh. 2025. Using semantic similarity to proxy for knowledge structures in accounting research. *Journal of Information Systems* 39 (2): 41–70. <https://doi.org/10.2308/ISYS-2023-010>
- Jordan, M. I., and T. M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349 (6245): 255–260. <https://doi.org/10.1126/science.aaa8415>

- Krupa, J., and M. Minutti-Meza. 2022. Regression and machine learning methods to predict discrete outcomes in accounting research. *Journal of Financial Reporting* 7 (2): 131–178. <https://doi.org/10.2308/JFR-2021-010>
- Kureljusic, M., and E. Karger. 2024. Forecasting in financial accounting with artificial intelligence: A systematic literature review and future research agenda. *Journal of Applied Accounting Research* 25 (1): 81–104. <https://doi.org/10.1108/JAAR-06-2022-0146>
- Lee, E. J., and D. Tahmoush. 2025. Auditors' decision-making aid for going concern audit opinions through machine learning analysis. *International Journal of Accounting Information Systems* 56: 100732. <https://doi.org/10.1016/j.accinf.2025.100732>
- Lin, Y.-C., R. Padliansyah, Y.-H. Lu, and W.-R. Liu. 2025. Bankruptcy prediction: Integration of convolutional neural networks and explainable artificial intelligence techniques. *International Journal of Accounting Information Systems* 56: 100744. <https://doi.org/10.1016/j.accinf.2025.100744>
- Liu, W.-P., M.-F. Yen, and T.-Y. Wu. 2022. Report users' perceived sentiments of key audit matters and firm performance: Evidence from a deep learning-based natural language processing approach. *Journal of Information Systems* 36 (3): 191–209. <https://doi.org/10.2308/ISYS-2020-061>
- Martín-Martín, A., E. Orduna-Malea, M. Thelwall, and E. Delgado López-Cózar. 2018. Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics* 12 (4): 1160–1177. <https://doi.org/10.1016/j.joi.2018.09.002>
- Massaro, M., J. Dumay, and J. Guthrie. 2016. On the shoulders of giants: Undertaking a structured literature review in accounting. *Accounting, Auditing & Accountability Journal* 29 (5): 767–801. <https://doi.org/10.1108/AAAJ-01-2015-1939>
- Munoko, I., H. L. Brown-Liburd, and M. A. Vasarhelyi. 2020. The ethical implications of using artificial intelligence in auditing. *Journal of Business Ethics* 167 (2): 209–234. <https://doi.org/10.1007/s10551-019-04407-1>
- Otero, A. R., and M. Agu. 2025. Benefits and drawbacks of incorporating ChatGPT in financial audits. *Current Issues in Auditing* 19 (2): 1–8. <https://doi.org/10.2308/CIIA-2024-016>
- Page, M. J., J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 372: n71. <https://doi.org/10.1136/bmj.n71>
- Parker, C., L. Jiang, S. Cho, and M. A. Vasarhelyi. 2025. Predicting material misstatements using machine learning. *The Accounting Review* 100 (6): 225–262. <https://doi.org/10.2308/TAR-2024-0035>
- Paul, J., and G. R. G. Benito. 2018. A review of research on outward foreign direct investment from emerging countries, including China: What do we know, how do we know and where should we be heading? *Asia Pacific Business Review* 24 (1): 90–115. <https://doi.org/10.1080/13602381.2017.1357316>
- Paul, J., S. Parthasarathy, and P. Gupta. 2017. Exporting challenges of SMEs: A review and future research agenda. *Journal of World Business* 52 (3): 327–342. <https://doi.org/10.1016/j.jwb.2017.01.003>
- Perols, J. 2011. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory* 30 (2): 19–50. <https://doi.org/10.2308/AJPT-50009>
- Pranckutė, R. 2021. Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. *Publications* 9 (1): 12. <https://doi.org/10.3390/publications9010012>
- Schneider, M., and R. Brühl. 2023. Disentangling the black box around CEO and financial information-based accounting fraud detection: Machine learning-based evidence from publicly listed U.S. firms. *Journal of Business Economics* 93 (9): 1591–1628. <https://doi.org/10.1007/s11573-023-01136-w>
- Serrano-Cinca, C., B. Gutiérrez-Nieto, and M. Bernate-Valbuena. 2019. The use of accounting anomalies indicators to predict business failure. *European Management Journal* 37 (3): 353–375. <https://doi.org/10.1016/j.emj.2018.10.006>
- Siano, F., and P. Wysocki. 2021. Transfer learning and textual analysis of accounting disclosures: Applying big data methods to small(er) datasets. *Accounting Horizons* 35 (3): 217–244. <https://doi.org/10.2308/HORIZONS-19-161>
- Smith, L., and C. Lamprecht. 2024. Identifying the limitations associated with machine learning techniques in performing accounting tasks. *Journal of Financial Reporting and Accounting* 22 (2): 227–253. <https://doi.org/10.1108/JFRA-05-2023-0280>
- Tragouda, M., M. Doumpos, and C. Zopounidis. 2024. Identification of fraudulent financial statements through a multi-label classification approach. *Intelligent Systems in Accounting, Finance and Management* 31 (2): e1564. <https://doi.org/10.1002/isaf.1564>
- Trinkle, B. S., and A. A. Baldwin. 2016. Research opportunities for neural networks: The case for credit. *Intelligent Systems in Accounting, Finance and Management* 23 (3): 240–254. <https://doi.org/10.1002/isaf.1394>
- Wang, Y., T. Chiu, V. Chiu, and M. A. Vasarhelyi. 2025. Detecting suspicious transactions using deep learning-based process mining. *Journal of Information Systems*. <https://doi.org/10.2308/ISYS-2024-006>
- Wei, D., S. Cho, M. A. Vasarhelyi, and L. Te-Wierik. 2024. Outlier detection in auditing: Integrating unsupervised learning within a multilevel framework for general ledger analysis. *Journal of Information Systems* 38 (2): 123–142. <https://doi.org/10.2308/ISYS-2022-026>
- Wu, C., and R. B. Dull. 2025. Using supervised learning algorithms to predict discontinued operations in nonprofit organizations. *Journal of Information Systems* 39 (3): 67–84. <https://doi.org/10.2308/ISYS-2021-035>

- Xie, X., and S. Li. 2026. Does reporting key audit matters change corporate disclosure strategy: Evidence from Chinese audit report reform. *Asia-Pacific Journal of Accounting & Economics* 33 (2): 364–379. <https://doi.org/10.1080/16081625.2025.2485985>
- Xiong, L., L. Peng, L. Xiao, and L. Yang. 2025. The importance of data security risk in setting audit fees: Evidence from a machine learning-based approach. *Emerging Markets Finance and Trade* 61 (15): 4756–4777. <https://doi.org/10.1080/1540496X.2025.2520882>
- Yang, Z. R., M. B. Platt, and H. D. Platt. 1999. Probabilistic neural networks in bankruptcy prediction. *Journal of Business Research* 44 (2): 67–74. [https://doi.org/10.1016/S0148-2963\(97\)00242-7](https://doi.org/10.1016/S0148-2963(97)00242-7)
- Zhang, C., S. Cho, and M. A. Vasarhelyi. 2022. Explainable artificial intelligence (XAI) in auditing. *International Journal of Accounting Information Systems* 46: 100572. <https://doi.org/10.1016/j.accinf.2022.100572>