Scientometrics Research Journal
Scientific Bi-Quarterly Journal of Shahed University/ Article in Press (Not edited)

Somayeh Ghavidel*[1]

Data Augmentation in Scientometrics and Knowledge Network Analysis: A Systematic Review and Decision-Oriented Conceptual Framework

Ph.D. in Information Science and Knowledge Studies, Kharazmi University; Head of the Scientometrics Committee, Iranian Library and Information Science Association, Tehran, Iran. ORCID ID: 0000-0002-3852-3398; s_ghavidel@isc.co.ir (corresponding author)

Abstract

Purpose: Data augmentation (DA) is increasingly used to overcome limited training samples and improve computational learning, yet its role in scientometrics and knowledge-network analysis remains conceptually unsettled. In this field, apparently incomplete or sparse data may reflect database coverage, inconsistent metadata, choices made during network construction, constraints on labelled training examples, or genuine characteristics of scientific activity. Treating these distinct conditions as interchangeable risks altering the phenomena that scientometric indicators are intended to measure. Moreover, methods for predicting links, recommending citations, or learning network representations are not necessarily augmentation methods. This study critically assesses the available evidence, distinguishes DA from data cleaning, recovery, integration, imputation, and computational inference, and develops an evidence-informed framework for deciding whether intervention is warranted, which approach is defensible, and how resulting data should be interpreted and validated.

Methodology: A systematic literature review was combined with directed qualitative content analysis using a deductive- inductive coding strategy. The Web of Science Core Collection corpus comprised 7,000 bibliographic records archived in seven export files, with publication years examined against the 2018–2026 review window. Temporal screening excluded 393 records, leaving 6,607; title and abstract screening excluded 6,507, and 100 records proceeded to full-text assessment. Historical screening records documented 43 initially included studies and 57 not retained for synthesis. Selection was reported using PRISMA 2020, with PRISMA principles guiding transparent discussion of search documentation. Because the complete original database query and individual screening decisions were not preserved, the historical search cannot be fully replicated. A subsequent recovery audit revisited the fixed corpus without undertaking a new database search. Of the original 43 studies, 31 remained in the core set and 12 were reassigned to supplementary methodological evidence; 12 additional eligible studies were recovered from the archived records. The revised core therefore comprised 43 studies, though its composition differed. These were classified by - evidentiary role into explicit scholarly-context DA applications

(A, six studies), computational mechanisms in scientific data and networks (B, 33), and data-quality or conceptual-boundary studies (C, four). Thirteen supplementary studies (D), including one newly identified, informed interpretation but were excluded from the core denominator; four sources found outside the archived corpus provided contextual background only. A study- evidence matrix recorded bibliographic identifiers, data contexts, mechanisms, analytical tasks, validation approaches, findings, and limitations. Coding distinguished directly reported results, cross-study synthesis, and conceptual inference. Because designs and outcomes were heterogeneous, the synthesis was qualitative rather than a meta-analysis, and no aggregate study-quality score was inferred. No causal effects were quantitatively estimated.

Findings: Only six of the 43 core studies documented DA in scholarly or closely related scientific-data settings. 76.7% concerned computational mechanisms, such as citation or article recommendation, link and collaboration prediction, network embedding, and knowledge-graph representations, which do not by themselves establish that augmentation occurred. Four studies addressed data quality and conceptual boundaries. Direct applications involved heterogeneous tasks, including patent-based technology analysis, synthetic publication time series, paper recommendation, and peer-review text. Their results supported task-specific possibilities rather than a general improvement in scientometric validity. Five recurring problem patterns emerged: uneven database coverage, metadata inconsistency, network-construction effects, class imbalance in particular learning tasks, and restricted training samples. Importantly, sparse collaboration structures or skewed citation distributions may be genuine properties of scientific systems, not defects requiring artificial correction. An inferred citation or predicted collaboration becomes DA only when its output is subsequently used as additional data or representation; predictive accuracy alone does not make that relationship historically observed. Accordingly, four epistemic states were differentiated: observed data, independently verified recovered data, inferred or predicted outputs, and synthetic or modified data. The synthesis identified four broad DA families- statistical or synthetic generation, deep generative methods, graph-based augmentation, and language-based augmentation—while emphasizing that evidence from adjacent domains cannot be transferred uncritically. Validation must extend beyond predictive metrics to statistical, structural, and semantic fidelity, alongside temporal validity where relevant. For scientific networks, changes in degree distributions, centrality, communities, and rankings require particular scrutiny. Provenance records, baseline comparisons, out-of-sample testing, sensitivity analysis, and expert review, where appropriate, support decisions about acceptance, restricted use, refinement, or rejection.

Conclusion: DA is a conditional methodological option, not a universal remedy for missing or imperfect scientometric data. The proposed eight-stage, decision-oriented conceptual framework moves from problem identification and causal diagnosis to specification of data and analytical units, selection of an intervention pathway and suitable mechanism, independent output validation, explicit recording of epistemic status, and sensitivity-informed decisions about use. Three principles govern these decisions: distinguish data operations conceptually, preserve evidence provenance, and maintain the epistemic status of every generated or recovered relationship. The framework does not claim empirical validation or predictive

superiority; rather, it makes the assumptions and evidentiary thresholds governing interventions auditable. Future research should test it against reference datasets and examine whether augmented outputs preserve meaningful citation, collaboration, and thematic structures before their inclusion in substantive scientometric analyses.

Keywords: Data Augmentation (DA); Scientometrics; Knowledge networks; Systematic literature review; Graph learning; Data validation; Epistemic status

زودتر پیش ویرایش نشده

Receive:
../../….
Acceptance:
../../….

پژوهش‌نامه علم‌سنجی
دوفصلنامه علمی ـ پژوهشی دانشگاه شاهد/ زودآیند ویرایش نشده

داده‌افزایی در علم‌سنجی و تحلیل شبکه‌های دانشی: مرور نظام‌مند شواهد و ارائه چارچوب مفهومی تصمیم‌محور

سمیه قویدل [۱*]

۱. دکتری علم اطلاعات و دانش‌شناسی دانشگاه خوارزمی،مسئول کمیته علم‌سنجی انجمن کتابداری و اطلاع‌رسانی ایران؛ شرکت خدمات انفورماتیک
ORCID ID: 0000-0002-3852-3398; s_ghavidel@isc.co.ir (corresponding author)

چکیده

هدف: داده‌افزایی در یادگیری ماشین، پردازش زبان طبیعی و یادگیری گراف گسترش یافته است؛ کاربرد آن در علم‌سنجی نیازمند مرزبندی دقیق با پاک‌سازی، بازیابی، یکپارچه‌سازی و پیش‌بینی روابط است. پژوهش حاضر با هدف تبیین جایگاه داده‌افزایی در علم‌سنجی و شبکه‌های دانشی و ارائه چارچوبی تصمیم‌محور برای استفاده و اعتبارسنجی آن انجام شد.

روش‌شناسی: مرور نظام‌مند با تحلیل محتوای کیفی جهت‌دار و منطق قیاسی- استقرایی تلفیق شد. پیکره ثابت پژوهش ۷۰۰۰ رکورد WoSCC در بازه ۲۰۱۸ تا ۲۰۲۶ بود. پس از حذف ۳۹۳ رکورد خارج از محدوده زمانی، ٦٦٠٧ رکورد غربال شدند؛ ٦٥٠٧ رکورد کنار و ۱۰۰ مدرک به ارزیابی متن کامل رسیدند. انتخاب مطالعات با الگوی PRISMA گزارش شد. ممیزی بازیابی، بدون جست‌وجوی جدید، ترکیب شواهد را اصلاح کرد: از ٤٣ مطالعه اولیه، ۳۱ مطالعه در هسته باقی ماندند، ۱۲ مطالعه به مجموعه مکمل انتقال یافتند و ۱۲ مطالعه دیگر از همان پیکره به هسته افزوده شدند. داده‌ها در ماتریس «مطالعه- شاهد» کدگذاری شدند.

یافته‌ها: بر اساس ماهیت و نقش شواهد، ٤٣ مطالعه هسته‌ای در سه گروه طبقه‌بندی شدند: شش مطالعه کاربرد صریح داده‌افزایی در داده‌ها را گزارش کردند؛ ۳۳ مطالعه به بررسی سازوکارهای محاسباتی مرتبط با شبکه‌های علمی پرداختند که لزوماً مصداق داده‌افزایی نیستند؛ و چهار مطالعه مسائل کیفیت داده و مرزبندی مفهومی داده‌افزایی را بررسی کردند. ۱۳ مطالعه مکمل صرفاً برای شناخت روش‌ها، شرایط انتقال و مخاطرات استفاده شدند. پیش‌بینی پیوند یا توصیه استناد، بدون استفاده بعدی از خروجی به‌عنوان داده افزوده، داده‌افزایی محسوب نمی‌شود. این شواهد بهبود عمومی دقت یا اعتبار علم‌سنجی بر اثر داده‌افزایی را اثبات نکردند. همچنین داده مشاهده‌شده، بازیابی‌شده و مستقلاً تأییدشده، استنباطی و مصنوعی باید متمایز بمانند.

نتیجه‌گیری: داده‌افزایی راهکاری همگانی برای مشکلات داده‌ای علم سنجی نیست. چارچوب تصمیم‌محور هشت‌مرحله‌ای پیشنهادی، تشخیص مسئله، انتخاب مداخله، ردیابی منشأ، اعتبارسنجی آماری، ساختاری، معنایی و درصورت‌لزوم زمانی و تعیین حدود استفاده از خروجی را به یکدیگر پیوند می‌دهد.

واژگان کلیدی: داده‌افزایی؛ علم سنجی؛ شبکه‌های دانشی؛ مرور نظام‌مند ادبیات؛ یادگیری گراف؛ اعتبار سنجی داده؛ وضعیت معرفتی داده

دریافت: ۰۰۰/۰۰/۰۰
پذیرش: ۰۰۰/۰۰/۰۰

زودآیند ویرایش نشده

مقدمه و بیان مسئله

علم‌سنجی[1] و تحلیل شبکه‌های دانشی[2] امکان مطالعه کمّی و ساختاری تولید، انتشار و ارتباط دانش علمی را از طریق تحلیل استنادها، همکاری‌ها و ساختارهای موضوعی فراهم می‌کنند. این تحلیل‌ها عمدتاً بر داده‌های پایگاه‌هایی مانند Web of Science (WoS)، Scopus و Dimensions استوارند؛ بنابراین، نتایج آن‌ها افزون بر شاخص و الگوریتم، به منبع داده، کیفیت فراداده[3] شیوه تشکیل مجموعه و بازنمایی روابط وابسته است (Bornmann & Leydesdorff, 2014; Waltman, 2016). بااین‌حال، «مسئله داده» در علم‌سنجی منشأ واحدی ندارد. تفاوت پوشش پایگاه‌ها ممکن است موجب شود استنادی در یک منبع ثبت شده باشد و در منبع دیگر نباشد؛ چنان‌که رودریگز و همکاران (Rodrigues et al., 2025) در مقایسه OpenAlex، Scopus و WoS نشان داده‌اند. مشکلات دیگر به ناسازگاری فراداده، شیوه‌های متفاوت ثبت نام پژوهشگران، ابهام در شناسایی موجودیت‌ها و نیاز به پیش‌پردازش مربوط‌اند (Nowakowska, 2025). انتخاب شیوه بازنمایی شبکه نیز می‌تواند شاخص‌ها و رتبه‌بندی‌ها را تغییر دهد (Vaccario & Verginer, 2022). در مقابل، کم‌تراکمی شبکه، توزیع نابرابر استنادها یا تفاوت اندازه اجتماعات علمی لزوماً نشانه نقص داده نیست و ممکن است ساختار واقعی علم را بازتاب دهد. ازاین‌رو، پیش از هر مداخله باید روشن شود وضعیت مشاهده‌شده ناشی از خطا، محدودیت پوشش یا فقدان داده است یا ویژگی واقعی پدیده.

گسترش یادگیری ماشین و هوش مصنوعی توجه به داده‌افزایی[4] را افزایش داده است. در ادبیات یادگیری ماشین، داده‌افزایی به عملیات کنترل‌شده‌ای اطلاق می‌شود که با اتکا به داده‌های موجود، نمونه‌ها، ویژگی‌ها یا بازنمایی‌های جدیدی برای اهداف آموزشی یا تحلیلی ایجاد می‌کند. بسته به نوع داده و وظیفه یادگیری، این عملیات از تبدیل‌های ساده تا تولید داده مصنوعی[5] با روش‌های مولد را شامل می‌شود (Feng et al., 2021; Mumuni & Mumuni, 2022; Shorten & Khoshgoftaar, 2019). کاربرد این مفهوم در علم‌سنجی مستلزم مرزبندی آن با پاک‌سازی داده[6]، جایگذاری مقادیر مفقود، غنی‌سازی داده[7] استانداردسازی فراداده و یکپارچه‌سازی منابع است. برای مثال، بازیابی یک استناد واقعی از منبع مستقل یا رفع ابهام نام پژوهشگر و ادغام رکوردهای دو پایگاه، حتی اگر داده‌های موجود را کامل‌تر کنند، به‌خودی‌خود داده‌افزایی نیستند. یکسان‌گرفتن این عملیات، هم دامنه مفهوم داده‌افزایی را گسترش می‌دهد و هم تفسیر نتایج را مخدوش می‌کند. این تمایز در داده‌های شبکه‌ای حساس‌تر است. روش‌هایی مانند «پیش‌بینی پیوند»[8]

[1] Scientometrics
[2] Knowledge Networks Analysis
[3] Metadata
[4] Data Augmentation (DA)
[5] Synthetic Data
[6] Data Cleaning
[7] Data Enrichment
[8] Link Prediction

«تعبیه شبکه»[1]و «شبکه‌های عصبی گرافی»[2]برای پیش‌بینی یا بازنمایی روابط شبکه‌ای توسعه یافته‌اند (Hamilton et al., 2017; Kipf & Welling, 2017; Liben-Nowell & Kleinberg, 2007; Veličković et al., 2018). این روش‌ها می‌توانند احتمال وجود رابطه یا شباهت گره‌ها را برآورد کنند، اما رابطه پیش‌بینی‌شده از نظر وضعیت معرفتی با رابطه تاریخی مشاهده‌شده یکسان نیست. در پژوهش حاضر، پیش‌بینی پیوند، توصیه استناد[3] و تکمیل گراف تنها زمانی در قلمرو داده‌افزایی قرار می‌گیرند که خروجی تولیدشده یا پیش‌بینی‌شده، در گام تحلیلی بعدی به‌عنوان داده یا بازنمایی افزوده استفاده شود؛ اگر پیش‌بینی، خروجی نهایی مدل باشد، شاهد مستقیمی از داده‌افزایی به شمار نمی‌رود. به همین ترتیب، اجرای الگوریتم یادگیری ماشین بر داده‌های متنی یا شبکه‌های علمی، بدون تولید یا به‌کارگیری داده افزوده، مصداق داده‌افزایی نیست. این مرزبندی درباره مدل‌های گرافی و مدل‌های زبانی بزرگ (الال‌ام)[4]نیز اهمیت دارد؛ زیرا موفقیت آن‌ها در یک وظیفه، اعتبار کاربردشان را در شبکه‌های استنادی یا هم‌نویسندگی تضمین نمی‌کند. در پردازش زبان طبیعی (انال‌پی)[5]و متون علمی، برخی کاربردهای داده‌افزایی در طبقه‌بندی متن و شناسایی ساختار معنایی گزارش شده‌اند (Chai et al., 2026; H. Liu et al., 2025)، اما نتایج باید در محدوده داده و وظیفه آزموده‌شده تفسیر شوند. مطالعه لیو و همکاران (Y. Liu et al., 2025) نیز که خارج از پیکره ثابت مرور شناسایی شد، فقط برای تکمیل زمینه نظری استفاده شده و در شمارش مطالعات هسته‌ای وارد نشده است.

تولید متن مصنوعی یا نمونه آموزشی در انال‌پی، از نظر معنا و پیامد تحلیلی با افزودن رابطه به شبکه استنادی یا هم‌نویسندگی یکسان نیست؛ بنابراین، انتقال روش‌های حوزه‌های مجاور به علم‌سنجی، نیازمند آزمون و اعتبارسنجی مستقل در داده و وظیفه مقصد است. فیگیرا و واز (Figueira & Vaz, 2022) بر ارزیابی کیفیت داده مصنوعی تأکید کرده‌اند و ساچدوا و همکاران (Sachdeva et al., 2025) نشان داده‌اند که نمونه‌های مصنوعی کم‌کیفیت می‌توانند وفاداری داده[6]و عملکرد مدل را تحت تأثیر قرار دهند. این موضوع در علم‌سنجی اهمیت ویژه‌ای دارد؛ زیرا تغییر داده ممکن است علاوه بر عملکرد مدل، بر مرکزیت‌ها، رتبه‌بندی‌ها، ساختار اجتماعات و تفسیر روابط علمی اثر بگذارد. ازاین‌رو، ارزیابی داده‌های افزوده باید متناسب با نوع داده، ابعاد وفاداری آماری، ساختاری و معنایی و در مسائل پویا، اعتبار زمانی را نیز دربر گیرد. منشأ داده و اثر مداخله بر نتایج تحلیل نیز باید قابل ردیابی باشد. از نظر وضعیت معرفتی، داده مشاهده‌شده، داده بازیابی‌شده و مستقلاً تأییدشده، رابطه استنباطی یا پیش‌بینی‌شده و داده مصنوعی یا تغییریافته جایگاه یکسانی ندارند. بنابراین، خروجی استنباطی، پیش‌بینی‌شده یا مصنوعی نباید بدون تأیید مستقل به‌عنوان داده مشاهده‌شده به کار رود. مرور ادبیات نشان می‌دهد که شواهد مربوط به داده‌افزایی، «یادگیری گراف»[7]و مدل‌های زبانی

[1] Network Embedding
[2] Graph Neural Networks
[3] Citation Recommendation
[4] LLMs
[5] Natural Language Processing (NLP)
[6] Data Fidelity
[7] Graph Learning

بزرگ در علم‌سنجی، از نظر نوع داده، هدف و نزدیکی به مسئله پژوهش ناهمگون‌اند. برخی مطالعات مستقیماً داده‌ها و فرایندهای علمی را بررسی می‌کنند؛ برخی به سازوکارهای محاسباتی شبکه‌های علمی می‌پردازند و برخی در حوزه‌های مجاور، مانند ان‌ال‌پی و تولید داده مصنوعی، شواهدی روش‌شناختی فراهم می‌آورند. نمی‌توان همه این مطالعات را با وزن تفسیری یکسان، شاهد اثربخشی داده‌افزایی در علم‌سنجی دانست. همچنین مرز داده‌افزایی با غنی‌سازی، یکپارچه‌سازی داده‌ها[1]، پیش‌بینی[2] برآورد و جایگذاری مقادیر مفقود[3] همیشه روشن نیست. در نتیجه، مشخص‌کردن اینکه کدام شواهد از کاربرد مستقیم داده‌افزایی حمایت می‌کنند، کدام قابلیت انتقال روش را نشان می‌دهند و این انتقال تحت چه شرایطی قابل دفاع است، همچنان یک مسئله پژوهشی است.

شکاف پژوهشی موردنظر، نبود تبیینی یکپارچه از منطق تشخیص مسئله داده‌ای، تعیین ضرورت مداخله، انتخاب روش متناسب با نوع داده و وظیفه تحلیلی و اعتبارسنجی خروجی بر اساس سطح شواهد است؛ نه صرفاً یافتن راهی برای تکمیل هر داده ناقص یا اثبات افزایش خودکار دقت علم‌سنجی با داده‌افزایی. در این رویکرد، داده‌افزایی یکی از گزینه‌ها در کنار پاک‌سازی، بازیابی، یکپارچه‌سازی و پیش‌بینی یا استنباط است. پژوهش حاضر با تلفیق مرور نظام‌مند ادبیات و تحلیل محتوای کیفی جهت‌دار، جایگاه و حدود کاربرد داده‌افزایی در علم‌سنجی و شبکه‌های دانشی را تبیین، شواهد مستقیم و انتقالی را تفکیک و الزامات اعتبارسنجی را بررسی می‌کند تا چارچوبی مفهومی و تصمیم‌محور برای استفاده مشروط از داده‌افزایی ارائه دهد. پرسش محوری آن است که در چه شرایطی، برای چه نوع داده و وظیفه‌ای و با چه سطحی از شواهد و اعتبارسنجی می‌توان کاربرد داده‌افزایی را در علم‌سنجی موجه دانست؟

پرسش های پژوهش

۱) بر اساس تحلیل محتوای کیفی جهت‌دار، ساختار شواهد و ابعاد اصلی مطالعات مرتبط با داده‌افزایی در علم‌سنجی و شبکه‌های دانشی چگونه قابل سازمان‌دهی است؟

۲) چه الگوهایی از مسائل داده‌ای، از جمله محدودیت پوشش پایگاه‌ها، ناسازگاری و نویز[4] داده، آثار ناشی از نحوه ساخت و بازنمایی شبکه و عدم‌توازن[5] کلاس‌ها در برخی وظایف طبقه‌بندی، در داده‌ها و تحلیل‌های علم‌سنجی شناسایی شده‌اند و چه پیامدهایی برای تفسیر نتایج در شبکه‌های دانشی دارند؟

۳) چه دسته‌هایی از روش‌های داده‌افزایی در مطالعات بررسی‌شده به‌طور مستقیم به کار رفته‌اند و شواهد روش‌شناختی مکمل چه ظرفیت‌هایی را برای انتقال مشروط این روش‌ها به مسائل علم‌سنجی نشان می‌دهند؟

[1] Data Integration
[2] Prediction
[3] Data Imputation
[4] Noise
[5] Imbalance

٤) چه سازوکارهایی برای پیش‌بینی، توصیه، استنباط و بازنمایی روابط در شبکه‌های استنادی، هم‌نویسندگی، موضوعی و دیگر شبکه‌های دانشی گزارش شده‌اند و خروجی این سازوکارها از نظر وضعیت معرفتی داده یا رابطه چگونه باید تفسیر شود؟

٥) شواهد و یافته‌های روش‌شناختی حوزه‌های مجاور، از جمله پردازش زبان طبیعی، یادگیری گراف و تولید داده مصنوعی، چه شرایط، ملاحظات و الزامات اعتبارسنجی را برای انتقال روش‌های داده‌افزایی به مسائل علم‌سنجی نشان می‌دهند؟

٦) چه مخاطرات، ملاحظات و الزامات اعتبارسنجی برای استفاده از داده‌های افزوده، مصنوعی یا استنباطی در تحلیل‌های علم‌سنجی گزارش شده‌اند؟

٧) بر اساس سنتز شواهد موجود، چه چارچوب مفهومی تصمیم‌محوری برای تشخیص منشأ مسئله داده‌ای، انتخاب نوع مداخله، تعیین وضعیت معرفتی، اعتبارسنجی و استفاده مشروط از داده‌افزایی در علم‌سنجی و شبکه‌های دانشی قابل استخراج است؟

چارچوب نظری

چارچوب نظری پژوهش بر این اصل استوار است که هر مسئله داده‌ای در علم‌سنجی مستلزم داده‌افزایی نیست؛ تشخیص ماهیت و منشأ مسئله بر انتخاب مداخله تقدم دارد. این چارچوب در هفت بُعد «مسئله داده‌ای، نوع داده، عملیات یا سازوکار، وظیفه تحلیلی، اعتبارسنجی، پیامد و محدودیت یا ریسک» سامان می‌یابد و سه اصل «مرزبندی مفهومی مداخلات، قابلیت ردیابی منشأ شواهد و حفظ وضعیت معرفتی داده» بر تفسیر آن حاکم است. در نتیجه، داده‌افزایی یکی از مداخلات ممکن است، نه راه‌حلی ازپیش‌تعیین‌شده.

- **داده‌افزایی و مرز آن با مداخلات داده‌ای مجاور**

داده‌افزایی در این پژوهش به عملیات کنترل‌شده‌ای اطلاق می‌شود که با اتکا به داده موجود، از طریق تولید یا تبدیل کنترل‌شده، نمونه، ویژگی یا بازنمایی جدیدی برای هدف آموزشی یا تحلیلی ایجاد می‌کند و در آن داده مشاهده‌شده مرجع باقی می‌ماند (Mumuni & Mumuni, 2022; Shorten & Khoshgoftaar, 2019). این عملیات با پاک‌سازی، غنی‌سازی، یکپارچه‌سازی، نرمال‌سازی و برآورد و جایگذاری مقادیر مفقود یکسان نیست. برای مثال، بازیابی استناد موجود از منبع دیگر، ادغام رکوردهای WoS و Scopus یا رفع ابهام نام نویسندگان، به‌خودی‌خود داده‌افزایی محسوب نمی‌شود. بنابراین، انتخاب داده‌افزایی به احراز نیاز به تولید یا تبدیل داده وابسته است، نه صرف دسترسی به الگوریتم.

- **مسائل داده‌ای و بازنمایی شبکه‌های علمی**

اعتبار تحلیل‌های علم‌سنجی به کیفیت فراداده، پوشش منبع و شیوه ساخت شبکه وابسته است (Bornmann & Leydesdorff, 2014; Waltman, 2016). باید میان خطا یا ابهام فراداده و تفاوت پوشش استناد (Nowakowska, 2025; Rodrigues et al., 2025)، اثر ساخت و فرافکنی شبکه (Vaccario & Verginer, 2022) و ویژگی‌های

واقعی ساختار علم تمایز گذاشت. کم‌تراکمی شبکه یا توزیع نابرابر استنادها لزوماً نقص داده نیست؛ مداخله زمانی موجه است که محدودیت داده احراز شود یا اثر آن بر یک وظیفه تحلیلی مشخص باشد.

- **تحلیل شبکه‌های دانشی در علم‌سنجی و وابستگی آن به کیفیت داده**

شبکه‌های استنادی، هم‌نویسندگی و موضوعی، روابط متفاوتی میان مقاله‌ها، پژوهشگران و موضوعات بازنمایی می‌کنند. کیفیت داده‌های ورودی، تعریف گره و یال و شیوه ساخت و فرافکنی شبکه می‌تواند بر شاخص‌هایی مانند درجه، مرکزیت، ساختار اجتماعات و رتبه‌بندی‌ها اثر بگذارد (Bornmann & Leydesdorff, 2014; Vaccario & Verginer, 2022; Waltman, 2016). ازاین‌رو، اعتبار تحلیل صرفاً از تراکم یا توازن شبکه به دست نمی‌آید، بلکه به منشأ و معنای روابط بستگی دارد. افزودن داده یا پیوند، بدون تشخیص مستقل مسئله و ضرورت مداخله، ممکن است ساختار واقعی روابط علمی و تفسیر آن‌ها را تحریف کند.

- **سازوکارهای محاسباتی و نسبت آن‌ها با داده‌افزایی**

روش‌های پیش‌بینی پیوند، تعبیه شبکه، شبکه‌های عصبی گراف، تکمیل گراف و توصیه استناد برای مدل‌سازی، استنباط یا بازنمایی روابط به کار می‌روند (Hamilton et al., 2017; Kipf & Welling, 2017; Liben-Nowell & Kleinberg, 2007; Veličković et al., 2018). این روش‌ها تنها زمانی در قلمرو داده‌افزایی قرار می‌گیرند که خروجی آن‌ها در مرحله بعد به‌عنوان داده یا بازنمایی افزوده استفاده شود؛ پیش‌بینی رابطه به‌عنوان خروجی نهایی، به‌خودی‌خود داده‌افزایی نیست. همچنین، پیوند پیش‌بینی‌شده یا استناد توصیه‌شده بدون تأیید مستقل، رابطه تاریخی مشاهده‌شده محسوب نمی‌شود. اجرای الگوریتم بر داده‌های علمی نیز نه وقوع داده‌افزایی را اثبات می‌کند و نه اعتبار علم‌سنجی نتیجه را. موفقیت داده‌افزایی گرافی، مدل‌های مولد یا مدل‌های زبانی در حوزه‌های دیگر صرفاً ظرفیت روش‌شناختی قابل آزمون را نشان می‌دهد و شاهد اثربخشی مستقیم این روش‌ها در شبکه‌های استنادی، هم‌نویسندگی یا موضوعی نیست.

- **اعتبارسنجی و وضعیت معرفتی داده**

افزودن داده تضمین‌کننده کیفیت نیست و نمونه‌های مصنوعی نامناسب می‌توانند وفاداری داده و عملکرد مدل را کاهش دهند (Figueira & Vaz, 2022; Sachdeva et al., 2025). بنابراین، ارزیابی داده‌های افزوده باید افزون بر عملکرد الگوریتم، وفاداری آماری، ساختاری و معنایی، مقایسه با خط مبنا، آزمون خارج از نمونه و تحلیل حساسیت را متناسب با نوع داده دربر گیرد؛ در شبکه‌های علمی، اثر مداخله بر توزیع درجه، مرکزیت‌ها، ساختار اجتماعات و رتبه‌بندی‌ها نیز بررسی شود. همچنین، منشأ و وضعیت معرفتی داده باید قابل ردیابی باشد. چهار وضعیت «مشاهده‌شده»، «بازیابی‌شده و مستقلاً تأییدشده»، «استنباطی/پیش‌بینی‌شده» و «مصنوعی/تغییریافته» از یکدیگر متمایزند. این تمایز باید در ثبت و تفسیر داده‌ها حفظ شود؛ زیرا خروجی‌های استنباطی یا مصنوعی نباید بدون تأیید مستقل به‌عنوان داده مشاهده‌شده گزارش شوند. بنابراین، پذیرش داده افزوده به منشأ، میزان وفاداری و هدف استفاده از آن وابسته است، نه صرفاً عملکرد محاسباتی.

پیشینه پژوهش

ادبیات داده‌افزایی در علم‌سنجی از نظر موضوع، روش و سطح شواهد ناهمگون است: برخی مطالعات کاربرد مستقیم داده‌افزایی در داده‌ها و فرایندهای علمی را بررسی کرده‌اند؛ برخی به کیفیت داده یا سازوکارهای استنباط روابط علمی پرداخته‌اند و گروهی نیز روش‌های قابل انتقال از حوزه‌های مجاور را معرفی کرده‌اند. ازاین‌رو، در این پیشینه، کاربرد مستقیم داده‌افزایی از عملیات پاک‌سازی، بازیابی و یکپارچه‌سازی داده و نیز پیش‌بینی روابط تفکیک می‌شود؛ زیرا موفقیت روش‌های محاسباتی در حوزه‌های دیگر، به‌تنهایی اثربخشی آن‌ها را در علم‌سنجی اثبات نمی‌کند.

- **کیفیت داده و مرز مسئله در علم‌سنجی**

یکی از پیش‌شرط‌های بحث درباره داده‌افزایی، روشن‌شدن ماهیت مسئله داده در علم‌سنجی است. آنجلینی و همکاران (Angelini et al., 2022) با تمرکز بر ارزیابی کیفیت اطلاعات در مدل‌های عملکردی سازمان‌های پژوهشی نشان دادند که قابلیت اتکای تحلیل به کیفیت داده‌ها و اطلاعات ورودی وابسته است. اهمیت این مطالعه برای پژوهش حاضر در آن است که هر ناهنجاری مشاهده‌شده در داده الزاماً نشانه کمبود داده یا ضرورت تولید داده جدید نیست؛ در برخی موارد، مسئله در مرحله شناسایی، پاک‌سازی یا تفسیر داده قرار دارد. در همان سال، واکاریو و ورجینر (Vaccario & Verginer, 2022) نشان دادند که شیوه ساخت و بازنمایی شبکه بر شاخص‌های حاصل اثر می‌گذارد و تغییر آن‌ها الزاماً نشانه فقدان رابطه واقعی نیست. نامبودیری و گویال (Namboodiri & Goyal, 2023) با تلفیق و استانداردسازی داده‌های چند منبع، نمونه‌ای از یکپارچه‌سازی اطلاعات موجود ارائه کردند که با تولید داده افزوده تفاوت دارد. رودریگز و همکاران (Rodrigues et al., 2025) مقایسه داده‌های استنادی OpenAlex، Scopus و WoS نشان دادند که پوشش استنادها در پایگاه‌های کتاب‌سنجی یکسان نیست. در چنین شرایطی، نبود یک استناد در یک پایگاه ممکن است ناشی از تفاوت پوشش منبع باشد، نه فقدان واقعی رابطه استنادی. در همین سال، نوواکوفسکا (Nowakowska, 2025) بر ضرورت پاک‌سازی، استانداردسازی، اتصال رکوردها و رفع ابهام در پیش‌پردازش داده‌های کتاب‌سنجی تأکید کرد. برآیند این مطالعات، تمایز میان خطای فراداده، محدودیت پوشش، ابهام موجودیت، اثر بازنمایی و ویژگی واقعی شبکه را ضروری می‌سازد. ازاین‌رو، تشخیص منشأ مسئله باید بر انتخاب داده‌افزایی تقدم داشته باشد.

- **شکل‌گیری و تحول روش‌های داده‌افزایی**

شورتن و خوشگفتار (Shorten & Khoshgoftaar, 2019)، داده‌افزایی را فراتر از تکثیر نمونه، شامل تبدیل‌های کنترل‌شده وابسته به نوع داده و وظیفه یادگیری معرفی کردند. براون و همکاران (Brown et al., 2020) قابلیت یادگیری چندنمونه‌ای و بدون نمونه در مدل‌های زبانی بزرگ را نشان دادند؛ این پژوهش زمینه روش‌شناختی تحول بعدی مدل‌های مولد بود، نه شاهد مستقیم داده‌افزایی. فنگ و همکاران (Feng et al., 2021) روش‌های جایگزینی واژگان، بازنویسی، ترجمه معکوس و تولید نمونه مصنوعی را در پردازش زبان طبیعی مرور کردند و وابستگی اثربخشی به وظیفه و ساختار داده را نشان دادند. مامونی و مامونی (Mumuni & Mumuni, 2022) نیز تنوع روش‌های داده‌افزایی و تفاوت عملکرد آن‌ها در مسائل کمبود نمونه آموزشی را بررسی کردند. فیگیرا و واز (Figueira & Vaz, 2022)

ضرورت ارزیابی کیفیت و وفاداری داده مصنوعی را در کنار تولید آن برجسته کردند. این تغییر جهت از «افزایش حجم» به «ارزیابی کیفیت» برای انتقال داده‌افزایی به حوزه علم‌سنجی اهمیت اساسی دارد. همچنین لیو و همکاران (Liu et al., 2022) در روش DAGAD از داده‌افزایی برای تشخیص ناهنجاری در داده‌های گرافی استفاده کردند. این پژوهش نشان می‌دهد که داده‌افزایی می‌تواند به ساختارهای گرافی نیز گسترش یابد؛ بااین‌حال، داده و هدف آن علم‌سنجی نبود. بنابراین، اهمیت آن برای پژوهش حاضر در ارائه یک سازوکار روش‌شناختی قابل آزمون است و نه در اثبات امکان اصلاح شبکه‌های استنادی یا هم‌نویسندگی. در پژوهشی دیگر، رعایایی (Roayaei, 2023) داده‌افزایی متنی را در تحلیل احساسات بررسی کرد و نشان داد که راهبردهای مختلف افزایش داده می‌توانند تحت محدودیت داده آموزشی، نتایج متفاوتی ایجاد کنند. در همین سال، زی و همکاران (Xie et al., 2023) از داده‌افزایی گرافی برای هم‌ترازی موجودیت‌ها در گراف دانش استفاده کردند. این مطالعه گسترش داده‌افزایی از داده‌های متنی به ساختارهای گرافی را نشان می‌دهد؛ بااین‌حال، هم‌ترازی موجودیت‌های گراف دانش از نظر معنای رابطه با شبکه‌های استنادی و هم‌نویسندگی یکسان نیست. دو نمونه داخلی نیز کاربرد داده‌افزایی را در حوزه‌هایی خارج از علم‌سنجی بررسی کردند. اسماعیلی و همکاران (Esmaely et al., 2024) داده‌افزایی را در مسئله محاسبه آنی بزرگای زلزله به کار گرفتند و میر و نوفرستی (Mir & Noferesti, 2024) از آن در تحلیل احساسات کاربران استفاده کردند. این مطالعات نشان می‌دهند که داده‌افزایی در داده‌ها و مسائل متنوعی قابل استفاده است، اما در پژوهش حاضر صرفاً شواهد زمینه‌ای حوزه‌های مجاور محسوب می‌شوند و نمی‌توان نتایج آنها را شاهدی برای اثربخشی داده‌افزایی در علم‌سنجی دانست. جوآن و همکاران (Juan et al., 2024) راهبرد تطبیقی داده‌افزایی را متناسب با ویژگی‌های توزیعی گراف بررسی کردند و دودامپگاما و همکاران (Dodampegama et al., 2024) محدودیت داده آموزشی و راهبردهای افزایش داده را در حوزه تفکیک پسماندهای ساختمانی گزارش کردند. ساچدوا و همکاران (Sachdeva et al., 2025) نشان دادند داده مصنوعی کم‌کیفیت می‌تواند وفاداری داده و عملکرد مدل را تضعیف کند. چای و همکاران (Chai et al., 2026) نیز در مرور داده‌افزایی متنی با مدل‌های زبانی بزرگ، روش‌های مبتنی بر پرامپت[1]، بازیابی و ترکیب راهبردها را همراه با لزوم کنترل کیفیت بررسی کردند. در مجموع، تحول از تبدیل‌های ساده به رویکردهای مولد و گرافی، ضرورت اعتبارسنجی وابسته به نوع داده و وظیفه را از میان نمی‌برد.

- **شواهد مستقیم داده‌افزایی در داده‌ها و فرآیندهای علمی**

در مجموعه نهایی پژوهش، شش مطالعه داده‌افزایی را به‌طور صریح در داده‌ها، شبکه‌ها یا فرایندهای علمی و حوزه‌های نزدیک به علم‌سنجی به کار گرفته‌اند. این مطالعات از نظر نوع داده، هدف و شیوه اعتبارسنجی یکسان نیستند و همین ناهمگنی، حدود تعمیم نتایج آنها را مشخص می‌کند. ژو و همکاران (Zhou et al., 2020) با تولید نمونه‌های مصنوعی

[1] Prompt

مبتنی بر شبکه مولد تخاصمی و یادگیری عمیق، پیش‌بینی فناوری‌های نوظهور را بر داده‌های ثبت اختراع بررسی کردند؛ شواهد آنان به همان وظیفه و مسئله محدودیت نمونه مربوط است. کاوثن و همکاران (Cauthen et al., 2022) با تولید منحنی‌های مصنوعی چرخه عمر فناوری از سری‌های زمانی تعداد انتشارات، مجموعه آموزشی تشخیص بلوغ فناوری را گسترش دادند و عملکرد طبقه‌بندی را با داده‌های اصلی و افزوده مقایسه کردند. شو و همکاران (Xu et al., 2022) نیز از داده‌افزایی مبتنی بر شبکه مولد تخاصمی و یادگیری عمیق برای طبقه‌بندی و ترسیم چشم‌انداز فناوری در داده‌های ثبت اختراع، تحت محدودیت و عدم‌توازن نمونه‌ها، استفاده کردند. این سه مطالعه استفاده مستقیم از داده‌افزایی را در تحلیل فناوری یا متغیر کتاب‌سنجی نشان می‌دهند؛ اما نتایجشان به معنای امکان بازسازی معتبر روابط تاریخی در شبکه‌های علمی نیست. در مطالعات جدیدتر، لیو و همکاران (Y. Liu et al., 2025) از داده‌افزایی و یادگیری مقایسه‌ای بر زیرگراف‌های گراف دانش ثبت اختراع برای شناسایی فناوری‌های نوظهور در تولید هوشمند استفاده کردند؛ زیرگراف افزوده، بازنمایی آموزشی است و نباید رابطه تاریخی تأییدشده تلقی شود. ژانگ و همکاران (C. Zhang et al., 2025) در مدل MKCRec داده‌افزایی خودنظارتی را بر گراف ناهمگن نویسنده– مقاله برای توصیه مقاله به کار گرفتند؛ بازنمایی تولیدشده در این مدل نیز با رابطه مشاهده‌شده یکسان نیست. ژانگ و همکاران (H. Zhang et al., 2025) در مطالعه ReviewGuard از داده‌افزایی مبتنی بر مدل زبانی بزرگ برای شناسایی داوری‌های علمی ناکافی استفاده کردند و ارزیابی انسانی را نیز در اعتبارسنجی گنجاندند. داده‌های داوری همتا، از نظر ساختار و وضعیت معرفتی، با روابط استنادی و همکاری علمی متفاوت‌اند. در کنار این شش مطالعه هسته‌ای، لیو و همکاران (H. Liu et al., 2025) داده‌افزایی را برای شناسایی ساختار کارکردی پاراگراف‌های متون علمی بررسی کردند، اما این مطالعه خارج از پیکره ثابت شناسایی شده و صرفاً در تبیین زمینه نظری به کار رفته است؛ بنابراین، در شمارش و درصدهای سنتز نظام‌مند لحاظ نمی‌شود. برآیند شش مطالعه مستقیم، امکان کاربرد داده‌افزایی در وظایف مشخص مرتبط با علم و فناوری را نشان می‌دهد، نه بهبود عام همه تحلیل‌های علم‌سنجی.

- **کیفیت، پالایش و حدود انتقال داده مصنوعی**

هم‌زمان با توسعه روش‌های تولید داده مصنوعی، مسئله کیفیت و اعتبار خروجی نیز برجسته‌تر شده است. فیگیرا و واز (Figueira & Vaz, 2022) ارزیابی کیفیت را مکمل ضروری تولید داده مصنوعی دانستند. ساچدوا و همکاران (Sachdeva et al., 2025) نشان دادند که نمونه‌های کم‌کیفیت می‌توانند عملکرد پیش‌بینی و وفاداری داده را کاهش دهند و بر پالایش پس از تولید تأکید کردند. چای و همکاران (Chai et al., 2026). نیز در داده‌افزایی متنی مبتنی بر مدل‌های زبانی بزرگ، کنترل کیفیت، تناسب زمینه‌ای و ارزیابی خروجی را ضروری شمردند. این شواهد درباره روش‌های مولد مستقیماً اعتبار علم‌سنجی داده افزوده را اثبات نمی‌کنند؛ زیرا تغییر داده در شبکه‌های علمی ممکن است توزیع درجه، مرکزیت، رتبه‌بندی، ساختار اجتماعات و تفسیر روابط را دگرگون کند. نوروزی چاکلی (Noroozi Chakoli, 2026) نیز در بررسی نسبت هوش مصنوعی با تحول علم‌سنجی، توجه به پیامدهای روش‌شناختی ابزارهای هوشمند را مطرح کرده است. بر این مبنا، در سنتز پژوهش حاضر، انتقال روش‌ها مستلزم سنجش وفاداری ساختاری و معنایی و

حفظ تمایز میان داده مشاهده‌شده، بازیابی‌شده و مستقلاً تأییدشده، استنباطی/پیش‌بینی‌شده و مصنوعی/تغییریافته تلقی می‌شود.

- **استنتاج از پیشینه**

پیشینه نشان می‌دهد مسائل کیفیت و پوشش داده با محدودیت نمونه آموزشی یا ویژگی‌های واقعی ساختار شبکه یکسان نیستند؛ همچنین روش‌های پیش‌بینی و بازنمایی روابط، صرفاً هنگامی داده‌افزایی محسوب می‌شوند که خروجی آن‌ها در مرحله بعد به‌عنوان داده یا بازنمایی افزوده استفاده شود. شش مطالعه دارای شواهد مستقیم، کاربردهایی ناهمگون در سری زمانی انتشارات، داده‌های ثبت اختراع، گراف دانش فناوری، توصیه مقاله و داوری همتا را گزارش می‌کنند؛ ازاین‌رو، تعمیم اثربخشی به تمام شبکه‌ها و شاخص‌های علم‌سنجی موجه نیست. مجموعه شواهد، چهار خلأ مرتبط را برجسته می‌کند: مرزبندی عملیاتی داده‌افزایی از مداخلات مجاور؛ محدودیت و ناهمگنی شواهد مستقیم؛ ضرورت حفظ منشأ و وضعیت معرفتی داده؛ و نبود چارچوبی منسجم برای پیوند تشخیص مسئله، انتخاب مداخله، اعتبارسنجی و تعیین حدود استفاده از خروجی. بنابراین، مسئله پژوهش صرفاً انتخاب روش تولید داده نیست، بلکه تعیین شرایطی است که در آن داده‌افزایی، مداخله جایگزین یا عدم مداخله، از نظر شواهد و تفسیر نتایج قابل دفاع باشد.

روش‌شناسی پژوهش

این پژوهش با تلفیق مرور نظام‌مند ادبیات[1] و تحلیل محتوای کیفی جهت‌دار[2] انجام شد. هدف، هدف، صرفاً گردآوری مطالعاتی نبود که از اصطلاح «داده‌افزایی» استفاده کرده‌اند؛ بلکه تلاش شد شواهد مربوط به مسئله‌های داده‌ای، کاربردهای صریح داده‌افزایی، سازوکارهای محاسباتی مرتبط، شیوه‌های اعتبارسنجی، پیامدهای گزارش‌شده و حدود انتقال روش‌های حوزه‌های مجاور به علم‌سنجی، به‌صورت نظام‌مند و قابل ردیابی سنتز شوند. اثربخشی داده‌افزایی از پیش مفروض گرفته نشد و هر گزاره تنها در حد شواهد مستقیم مطالعات اولیه یا سنتز قابل ردیابی چند مطالعه تفسیر شد.

تحلیل محتوای جهت‌دار برای این طرح مناسب بود، زیرا کدگذاری از مفاهیم اولیه برگرفته از مبانی نظری، تعریف عملیاتی داده‌افزایی و پرسش‌های پژوهش آغاز شد و در عین حال امکان افزوده‌شدن استقرایی مفاهیم جدید در جریان مطالعه منابع حفظ گردید (Elo & Kyngäs, 2008; Hsieh & Shannon, 2005). گزارش مراحل شناسایی، غربالگری، ارزیابی صلاحیت و ورود مطالعات با ملاحظات PRISMA 2020 انجام شد و اصول PRISMA-S نیز برای افزایش شفافیت گزارش منبع داده و منطق بازیابی مورد توجه قرار گرفت (Page et al., 2021; Rethlefsen et al., 2021).

[1] Systematic Literature Review - SLR
[2] Directed Qualitative Content Analysis

- **تعریف عملیاتی داده‌افزایی و مرز مفهومی آن**

به‌دلیل هم‌پوشانی مفهومی اصطلاحات در ادبیات، داده‌افزایی پیش از غربالگری و طبقه‌بندی مطالعات به‌صورت عملیاتی تعریف شد. در این پژوهش، داده‌افزایی به مجموعه‌ای از عملیات کنترل‌شده اطلاق می‌شود که با اتکا به داده موجود، نمونه، بازنمایی، ویژگی یا ساختار مشتق‌شده جدیدی را برای یک هدف تحلیلی یا مدل‌سازی ایجاد یا تبدیل می‌کند، بدون آنکه داده مشاهده‌شده اصلی جایگاه خود را به‌عنوان داده مرجع از دست بدهد. بر این اساس، صرف استفاده از «پیش‌بینی پیوند»، «تکمیل گراف»[1]، «توصیه استناد»، «تعبیه شبکه» یا «گراف دانش»[2]به‌طور خودکار داده‌افزایی تلقی نشد. اگر پیش‌بینی یک پیوند صرفاً خروجی نهایی مدل بود، مطالعه در گروه سازوکارهای استنباطی یا بازنمایی قرار گرفت؛ اما اگر خروجی تولیدشده در مرحله بعد به‌عنوان نمونه، ویژگی یا بازنمایی افزوده وارد فرآیند تحلیلی می‌شد، امکان طبقه‌بندی آن به‌عنوان داده‌افزایی بررسی شد. همچنین روابط پیش‌بینی‌شده تا زمانی که با شاهد مستقل تایید نشده باشند، معادل روابط واقعی تاریخی در شبکه‌های استنادی یا همکاری علمی در نظر گرفته نشدند. داده‌افزایی از Data Cleaning، Data Enrichment، Data Integration، Metadata Normalization (نرمال‌سازی متادیتا)، Missing-data Imputation (برآورد و جای‌گذاری داده‌های مفقود)و سایر عملیات بازیابی یا اصلاح داده متمایز شد. مطالعاتی که صرفاً به پاک‌سازی، رفع ناهمگونی، استانداردسازی فراداده یا تکمیل رکورد با اطلاعات بیرونی می‌پرداختند، به‌عنوان شاهد اثربخشی داده‌افزایی محسوب نشدند؛ بااین‌حال، در مواردی که این مطالعات برای روشن‌کردن منشأ مسئله داده‌ای یا تعیین مرز مفهومی داده‌افزایی ضروری بودند، به‌عنوان شواهد مرزبندی حفظ شدند. همچنین عدم‌توازن، تراکم پایین یا ناهمگونی یک شبکه به‌خودی‌خود «نقص داده» تلقی نشد، زیرا این ویژگی‌ها ممکن است بخشی از ساختار واقعی نظام علمی باشند. برای جلوگیری از خلط میان واقعیت مشاهده‌شده و خروجی‌های مدل، چهار وضعیت معرفتی در سراسر تحلیل از یکدیگر جدا نگه داشته شد: داده مشاهده‌شده؛ داده بازیابی‌شده و مستقلاً تاییدشده؛ داده استنباطی یا پیش‌بینی‌شده؛ و داده مصنوعی یا تغییریافته. این تفکیک مانع از آن شد که داده یا رابطه تولیدشده توسط مدل صرفاً به دلیل عملکرد مناسب، به‌عنوان واقعیت مشاهده‌شده گزارش شود.

- **منبع داده و راهبرد جست‌وجو**

منبع اصلی شناسایی مطالعات WoSCC[3]یا مجموعه هسته‌ای وب‌آوساینس، یکی از منابع شناخته‌شده نمایه‌سازی و تحلیل استنادی بین‌المللی بود (Birkle et al., 2020). جست‌وجوی اصلی بر اساس تاریخ ثبت‌شده در مستندات پژوهش، در ۱۸ ژوئن ۲۰۲۶ (۲۸ خرداد ۱۴۰۵) انجام شد. مجموعه خام پژوهش شامل ۷۰۰۰ رکورد کتاب‌شناختی از WoSCC بود که در هفت فایل متوالی، هر یک شامل ۱۰۰۰ رکورد، ذخیره شدند. بررسی مجدد فایل‌های اولیه نشان داد که تاریخ خروجی ثبت‌شده در هر هفت فایل، ٤ ژوئیه ۲۰۲۶ است. بنابراین، تاریخ جست‌وجوی گزارش‌شده و

[1] Graph Completion
[2] Knowledge Graph
[3] Web of Science Core Collection (WoSCC)

تاریخ ذخیره خروجی‌ها به‌عنوان دو رویداد متمایز در نظر گرفته شدند. رکوردهای موجود شامل اطلاعات کتابشناختی نظیر عنوان، چکیده، کلیدواژه‌های نویسندگان، کلیدواژه‌های افزوده[1] سال انتشار، DOI، اطلاعات موضوعی و شناسه یکتای WoS بودند. از آنجا که «تاریخچه جست‌وجو»[2] اولیه WoSCC، شامل عبارت دقیق اجراشده و تمامی تنظیمات تاریخی جست‌وجو، به‌صورت کامل در پرونده پژوهش باقی نمانده بود، در مرحله بازنگری از نسبت‌دادن یک عبارت بازسازی‌شده به‌عنوان «پرس‌وجو»[3] تاریخی قطعیِ مولد همه ۷۰۰۰ رکورد پرهیز شد. در مقابل، برای پاسخ به مسئله جامعیت و اختصاصیت انتخاب مطالعات، منطق جست‌وجو بر پایه تعریف عملیاتی پژوهش بازسازی و به‌صورت یک «جست‌وجوی بازیابی مجدد درون مجموعه ثابت»[4] بازتولیدپذیر بر همان مجموعه ثابت ۷۰۰۰ رکورد اعمال شد. در این مرحله هیچ رکورد جدیدی از WoSCC استخراج یا به مجموعه خام افزوده نشد. «جست‌وجوی بازیابی مجدد درون مجموعه ثابت» در چهار مسیر مفهومی صورت‌بندی شد. در هر مسیر، اصطلاحات مترادف یا هم‌خانواده داخل یک بلوک با OR و بلوک‌های مفهومی مستقل با AND ترکیب شدند. استفاده از چند مسیر به این دلیل بود که یک «پرس‌وجو» منفرد از نوع Data Augmentation AND Scientometrics نمی‌توانست هم‌زمان شواهد مستقیم داده‌افزایی، سازوکارهای محاسباتی مرتبط با شبکه‌های علمی، شواهد کیفیت داده و مطالعات روش‌شناختی انتقالی را با حساسیت و اختصاصیت کافی پوشش دهد..

جدول ۱. عبارات Topic (TS) مورد استفاده در «جست‌وجوی بازیابی مجدد درون مجموعه ثابت» و ممیزی نهایی مجموعه ثابت

مسیرها	عنوان مسیر	راهبرد جست‌وجو
مسیر A	داده‌افزایی صریح در داده‌ها و فرآیندهای علمی	TS=(("data augment*" OR "data-augment*" OR "synthetic data" OR "synthetic sample*" OR "synthetic data generation" OR "data synthesis" OR "text augment*" OR "graph augment*" OR "network augment*" OR "generative augment*") AND (scientometric* OR bibliometric* OR "science of science" OR "science mapping" OR "research evaluation" OR "scholarly data" OR "scientific literature" OR "scientific text*" OR "scientific paper*" OR "peer review" OR "citation function*" OR "citation intent*" OR "citation context*" OR "citation network*" OR "co-citation network*" OR "co-authorship network*" OR coauthorship OR "bibliographic network*" OR "scholarly network*" OR "academic network*" OR "scientific network*" OR patent* OR "technology intelligence"))
مسیر B	سازوکارهای محاسباتی مستقیم در داده‌ها و شبکه‌های علمی	TS=((scientometric* OR bibliometric* OR "science of science" OR "science mapping" OR "research evaluation" OR "scholarly data" OR "citation network*" OR "co-citation network*" OR "co-authorship network*" OR coauthorship OR "bibliographic network*" OR "scholarly network*" OR "academic network*" OR "scientific network*" OR patent*) AND ("link prediction" OR "network embedding" OR "graph embedding" OR "network representation learning" OR "citation recommendation" OR "paper recommendation" OR "coauthorship prediction" OR "co-authorship prediction" OR "collaboration prediction" OR "knowledge graph*" OR "citation function classif*" OR "citation intent classif*" OR "citation classification" OR "citation time series"))

[1] Keywords Plus
[2] Search History
[3] Query
[4] Recovery Search

مسیر C	کیفیت داده و مرزبندی مفهومی	TS=((scientometric* OR bibliometric* OR "citation data" OR "bibliographic data" OR "Web of Science" OR Scopus OR OpenAlex) AND ("data quality" OR "metadata quality" OR preprocess* OR clean* OR normaliz* OR disambiguat* OR "data integration" OR "database integration" OR "missing citation*" OR "incoherent citation*" OR "citation inconsisten*" OR "database merging"))
مسیر D	شواهد روش‌شناختی مجاور	TS=(("data augment*" OR "data-augment*" OR "synthetic data" OR "synthetic data generation" OR "data synthesis" OR "text augment*" OR "graph augment*" OR "generative augment*") AND ("natural language processing" OR NLP OR "graph neural network*" OR GNN OR "large language model*" OR LLM OR "knowledge graph*" OR "class imbalance" OR "mixed-type data"))

در مرحله بازنگری، برای روشن‌شدن دامنه موضوعی پژوهش، چهار محور مفهومی مرتبط با داده‌افزایی صریح، سازوکارهای محاسباتی شبکه‌های علمی، مسائل کیفیت داده و شواهد روش‌شناختی حوزه‌های مجاور از یکدیگر تفکیک شدند. این محورها برای تبیین منطق انتخاب و طبقه‌بندی مطالعات به کار رفتند و نباید معادل عبارت دقیق جست‌وجوی اولیه در پایگاه WoSCC تلقی شوند. ازآنجاکه تاریخچه کامل جست‌وجوی اولیه در مستندات پژوهش محفوظ نمانده بود، مجموعه ۷۰۰۰ رکوردی به‌عنوان پیکره بایگانی‌شده و محدوده شناسایی اولیه مطالعات معرفی شد. بر این اساس، دامنه استنتاج پژوهش به شواهد انتخاب‌شده از این مجموعه محدود است و ادعای بازتولید کامل راهبرد تاریخی جست‌وجو یا پوشش تمامی آثار مرتبط مطرح نمی‌شود.

- **ممیزی زمانی و تثبیت مجموعه قابل بررسی**

دامنه زمانی پژوهش سال‌های ۲۰۱۸ تا ۲۰۲۶ بود. بازبینی هفت فایل اولیه استخراج‌شده از WoSCC شامل ۷۰۰۰ رکورد کتاب‌شناختی، نتایج ممیزی زمانی گزارش‌شده را تأیید کرد. از مجموع رکوردها، ۳۹۲ رکورد به سال‌های پیش از ۲۰۱۸ تعلق داشتند و یک رکورد دارای سال انتشار رسمی ۲۰۲۷ بود. بر این اساس، پس از کنارگذاری ۳۹۳ رکورد خارج از محدوده زمانی، ٦٦٠٧ رکورد برای مرحله غربالگری موضوعی باقی ماندند. بررسی شناسه‌های یکتای وب‌آوساینس (UT) نشان داد که تمامی ۷۰۰۰ رکورد دارای شناسه معتبر و متمایز هستند و هیچ شناسه تکراری در مجموعه هفت فایل مشاهده نشد. همچنین، تطبیق شناسه‌های مطالعات نهایی با مجموعه اولیه، حضور هر ٤٣ مطالعه هسته‌ای و ۱۳ مطالعه روش‌شناختی مکمل را در پیکره اولیه تأیید کرد. این بررسی، تعداد رکوردهای اولیه، یکتایی شناسه‌ها و نتایج ممیزی زمانی را مستند می‌کند؛ بااین‌حال، به معنای بازسازی کامل تاریخچه جست‌وجو یا تصمیم‌های غربالگری اولیه نیست. در مرحله بازنگری، ممیزی بازیابی درون مجموعه ثابت نشان داد برخی مطالعات مرتبط، با وجود حضور در مجموعه اولیه، پیش‌تر به‌درستی شناسایی یا طبقه‌بندی نشده بودند. این مطالعات بدون اجرای جست‌وجوی جدید در پایگاه یا افزودن رکورد به مجموعه اولیه، از فایل‌های خام بازیابی و با معیارهای ثابت ورود و خروج دوباره ارزیابی شدند.

- **معیارهای ورود و خروج**

انتخاب مطالعات برای ورود به مجموعه هسته‌ای پژوهش بر اساس سه دسته معیار کتاب‌شناختی، ارتباط موضوعی و کفایت اطلاعات روش‌شناختی انجام شد. از نظر کتاب‌شناختی، مطالعات باید در بازه زمانی ۲۰۱۸ تا تاریخ اجرای

جست‌وجوی اولیه در سال ۲۰۲۶ منتشر شده و در قالب مقاله پژوهشی[1] مقاله مروری[2] یا مقاله همایشی (مندرج در مجموعه مقالات همایش)[3] قرار گرفته باشند. از نظر محتوایی، ورود به مجموعه هسته‌ای مستلزم احراز دست‌کم یکی از سه شرط بود: نخست، مطالعه داده‌افزایی را به‌طور صریح در داده‌ها یا فرایندهای علمی، علم‌سنجی یا تحلیل علم و فناوری به کار گرفته باشد (گروه A)؛ دوم، سازوکارهای محاسباتی مرتبط را مستقیماً بر داده‌ها یا شبکه‌های علمی بررسی کرده باشد (گروه B)؛ و سوم، شواهدی مرتبط با مسائل کیفیت داده، محدودیت پوشش منابع، ویژگی‌های بازنمایی شبکه یا مرزبندی مفهومی داده‌افزایی ارائه کرده باشد (گروه C). افزون بر این، هر مطالعه باید متناسب با ماهیت پژوهشی، مروری یا مفهومی خود، اطلاعات کافی برای تشخیص مسئله مورد بررسی، نوع داده یا موضوع تحلیل، روش یا سازوکار و نقش یافته‌های آن در پاسخ به پرسش‌های پژوهش ارائه می‌کرد. در مطالعات تجربی نیز اطلاعات مربوط به وظیفه تحلیلی و شیوه اعتبارسنجی، در صورت گزارش، استخراج و بررسی شد. مطالعاتی که صرفاً به کاربرد عمومی هوش مصنوعی، یادگیری ماشین یا تحلیل شبکه پرداخته بودند و ارتباط مشخصی میان مسئله بررسی‌شده، سازوکار مورد استفاده و پرسش‌های پژوهش برقرار نمی‌کردند، واجد شرایط ورود به مجموعه هسته‌ای محسوب نشدند. همچنین عملیاتی مانند پاک‌سازی داده‌ها، غنی‌سازی داده‌ها[4]، نرمال‌سازی یا یکسان‌سازی داده‌ها[5]، یکپارچه‌سازی داده‌ها[6] و برآورد و جایگذاری مقادیر مفقود[7] به‌خودی‌خود مصداق کاربرد مستقیم داده‌افزایی تلقی نشدند. بااین‌حال، مطالعات مرتبط با این عملیات، در صورتی که به شناخت منشأ مسائل کیفیت داده یا روشن‌شدن مرز مفهومی داده‌افزایی در علم‌سنجی کمک می‌کردند، می‌توانستند در گروه C قرار گیرند. در طبقه‌بندی مطالعات، علاوه بر ارتباط موضوعی، نقش هر مطالعه در ساختار استنتاج پژوهش نیز مورد توجه قرار گرفت. بر این اساس، مطالعاتی که داده‌افزایی را در حوزه‌های خارج از داده‌ها و فرایندهای علمی به کار گرفته بودند، تنها در صورت برخورداری از ارتباط روشن‌شناختی مشخص با مسائل پژوهش و امکان بررسی شرایط انتقال روش، در مجموعه روش‌شناختی مکمل (گروه D) قرار گرفتند. شواهد این گروه برای شناسایی روش‌ها، بررسی الزامات اعتبارسنجی و شناخت مخاطرات احتمالی انتقال داده‌افزایی از حوزه‌هایی مانند پردازش زبان طبیعی، یادگیری گراف و تولید داده مصنوعی مورد استفاده قرار گرفتند، اما در محاسبه فراوانی و درصد مطالعات هسته‌ای منظور نشدند. مطالعات مروری نیز برای شناسایی مفاهیم، طبقه‌بندی روش‌ها، بررسی شرایط انتقال و تبیین محدودیت‌ها مورد استفاده قرار گرفتند. بااین‌حال، در مواردی که یک مطالعه مروری یافته‌های پژوهش‌های اولیه موجود در مجموعه را بازگو می‌کرد، محتوای آن به‌عنوان شاهد تجربی مستقل و مضاعف در تفسیر نتایج منظور نشد.

[1] Article
[2] Review
[3] Proceedings Paper
[4] Data Enrichment
[5] Normalization
[6] Integration
[7] Imputation

این تمایز برای جلوگیری از چندباره‌شماری شواهد و حفظ تناسب میان نوع مطالعه و جایگاه آن در سنتز ضروری بود. در مجموع، در فرایند انتخاب مطالعات، میان ارتباط واژگانی با موضوع پژوهش و صلاحیت محتوایی برای ورود به سنتز تمایز برقرار شد. صرف حضور اصطلاحاتی مانند داده‌افزایی، یادگیری ماشین، پیش‌بینی پیوند یا تحلیل شبکه در عنوان و چکیده، برای احراز صلاحیت مطالعه کافی نبود؛ بلکه تناسب مسئله داده‌ای، نوع داده، سازوکار مورد بررسی و نقش شواهد با پرسش‌های پژوهش، مبنای اصلی تصمیم‌گیری بود. این رویکرد امکان تفکیک مطالعات مرتبط با دامنه کلی موضوع از مطالعات واجد شرایط سنتز را فراهم کرد و مانع از آن شد که یافته‌های حوزه‌های مجاور، بدون توجه به شرایط انتقال و اعتبارسنجی، به‌عنوان شواهد مستقیم کاربرد داده‌افزایی در علم‌سنجی تفسیر شوند.

- **فرآیند انتخاب مطالعات و ممیزی بازیابی**

برای گزارش جریان انتخاب مطالعات، از ساختار پیشنهادی PRISMA 2020 در تفکیک مراحل شناسایی اولیه، غربالگری عنوان و چکیده، ارزیابی صلاحیت متن کامل و ورود مطالعات به سنتز استفاده شد (Page et al., 2021). مجموعه اولیه پژوهش شامل ۷۰۰۰ رکورد بازیابی‌شده از WoSCC بود. پس از بررسی سال انتشار، ۳۹۳ رکورد خارج از محدوده زمانی ۲۰۱۸ تا ۲۰۲۶ کنار گذاشته شدند و ٦٦٠٧ رکورد برای غربالگری موضوعی باقی ماندند. در مرحله غربالگری عنوان و چکیده، ٦٥٠٧ رکورد، معادل حدود ۹۸٫۵ درصد مجموعه باقی‌مانده، به ارزیابی متن کامل راه نیافتند و ۱۰۰ مدرک برای بررسی تفصیلی انتخاب شدند. مطابق آمار ثبت‌شده فرایند اولیه، از میان این مدارک، ٤٣ مطالعه وارد ترکیب تحلیلی شدند و ۵۷ مدرک در مجموعه اولیه سنتز قرار نگرفتند.

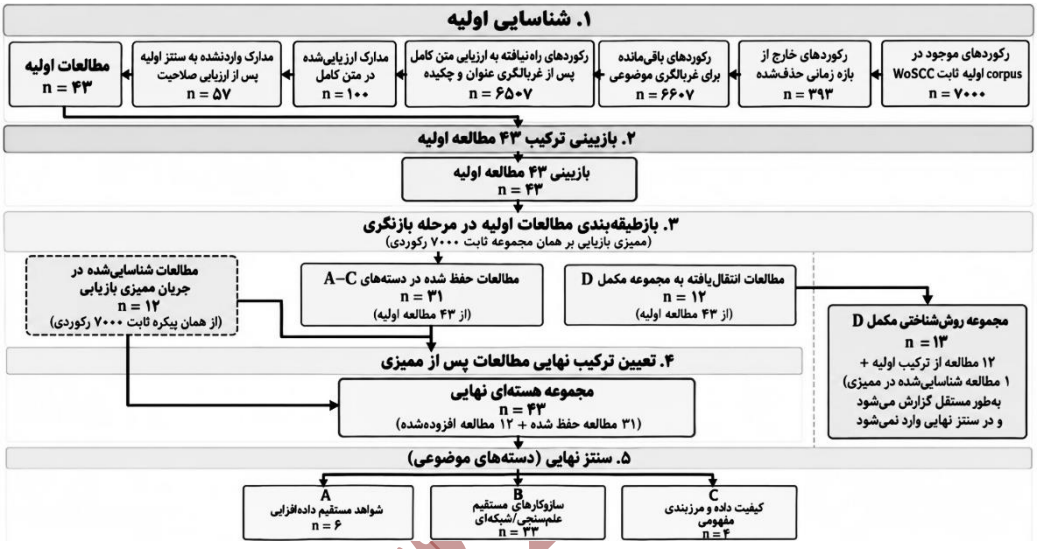
در تفسیر این فرایند، میان شناسایی اولیه رکوردها، غربالگری ارتباط موضوعی و احراز صلاحیت برای ورود به سنتز تمایز برقرار شد. تعداد رکوردهای بازیابی‌شده لزوماً بیانگر تعداد مطالعات واجد شرایط ورود به مجموعه نهایی نیست؛ زیرا ارتباط واژگانی یک مدرک با موضوع پژوهش، به‌تنهایی برای تأیید صلاحیت محتوایی آن کفایت نمی‌کند. در پژوهش حاضر، داده‌افزایی با حوزه‌هایی مانند یادگیری ماشین، پردازش زبان طبیعی و تحلیل شبکه هم‌پوشانی مفهومی داشت؛ درحالی‌که ورود به سنتز هسته‌ای مستلزم برخورداری از شواهد مرتبط با داده‌ها و فرایندهای علمی، سازوکارهای محاسباتی قابل بررسی در شبکه‌های علمی یا مسائل کیفیت داده و مرزبندی مفهومی بود. بر این اساس، نسبت بالای رکوردهای کنارگذاشته‌شده در غربالگری عنوان و چکیده را نمی‌توان به‌تنهایی نشانه نادرستی راهبرد جست‌وجو دانست. در راهنمای کاکرین نیز توضیح داده شده است که افزایش حساسیت جست‌وجو در مرورهای نظام‌مند می‌تواند با کاهش دقت بازیابی و افزایش تعداد مدارک نامرتبط همراه باشد (Lefebvre et al., 2025). بااین‌حال، این رابطه عمومی به معنای تأیید حساسیت بالای راهبرد جست‌وجوی پژوهش حاضر نیست. به‌دلیل در دسترس نبودن تاریخچه کامل جست‌وجوی اولیه و اطلاعات لازم درباره مطالعات مرتبط شناسایی‌نشده، امکان سنجش مستقل حساسیت و دقت بازیابی فراهم نبود. ازاین‌رو، نسبت ۹۸٫۵ درصد صرفاً به‌عنوان یکی از ویژگی‌های توصیفی فرایند غربالگری گزارش شد و مبنای داوری قطعی درباره عملکرد راهبرد جست‌وجو قرار نگرفت. در مرحله ارزیابی متن کامل، انتخاب ۱۰۰ مدرک به معنای تأیید نهایی صلاحیت آنها برای ورود به سنتز نبود. در این مرحله، تناسب محتوای مطالعات با معیارهای

ورود، نوع داده، ماهیت عملیات یا سازوکار و نقش شواهد در پاسخ به پرسش‌های پژوهش اهمیت داشت. بااین‌حال، ازآنجاکه سوابق تفصیلی دلایل موردی عدم ورود ۵۷ مدرک به‌طور کامل در مستندات باقی‌مانده پژوهش در دسترس نبود، امکان ارائه توزیع عددی مستند از دلایل خروج آنها فراهم نشد. با توجه به تأکید PRISMA 2020 بر شفافیت فرایند انتخاب مطالعات و گزارش دلایل کنارگذاشتن مدارک در مرحله ارزیابی صلاحیت (Page et al., 2021)، از انتساب پسینی دلایل فرضی به این مدارک خودداری شد و محدودیت ناشی از فقدان سوابق یادشده در روش‌شناسی پژوهش تصریح گردید. در مرحله بازنگری مقاله، به‌منظور بررسی سازگاری ترکیب مطالعات با تعریف عملیاتی داده‌افزایی و معیارهای ورود به مجموعه هسته‌ای، «ممیزی بازیابی مطالعات»[1] بر همان پیکره ثابت ۷۰۰۰ رکوردی انجام شد. این مرحله جست‌وجوی جدیدی در WoSCC نبود و رکورد تازه‌ای از پایگاه به مجموعه اولیه افزوده نشد. هدف آن، بازبینی ترکیب شواهد و شناسایی مطالعاتی بود که در پیکره اولیه حضور داشتند، اما در ترکیب تحلیلی پیشین وارد نشده یا متناسب با نقش روش‌شناختی خود طبقه‌بندی نشده بودند.

بر اساس نتایج ممیزی گزارش‌شده، از ۴۳ مطالعه مجموعه اولیه، ۳۱ مطالعه با معیارهای ورود به گروه‌های هسته‌ای A تا C انطباق داشتند و ۱۲ مطالعه در زمره شواهد روش‌شناختی مجاور قرار گرفتند. در این پژوهش، «مطالعات روش‌شناختی مجاور» به پژوهش‌هایی اطلاق می‌شود که روش‌های داده‌افزایی را در حوزه‌های مرتبط بررسی کرده‌اند (Feng et al., 2021; Mumuni & Mumuni, 2022)، اما شواهد مستقیمی از کاربرد آنها در علم‌سنجی ارائه نمی‌کنند و یافته‌هایشان صرفاً قابلیت انتقال مشروط به مسائل این حوزه را دارند. اهمیت این مطالعات در فراهم‌کردن مبانی روش‌شناختی برای شناخت سازوکارهای داده‌افزایی، بررسی شرایط انتقال روش‌ها به داده‌های علمی، شناسایی الزامات اعتبارسنجی و تبیین مخاطرات احتمالی آنهاست. بنابراین، این ۱۲ مطالعه به‌دلیل تفاوت در زمینه کاربرد و نقش شواهد، از مجموعه هسته‌ای به گروه روش‌شناختی مکمل D منتقل شدند، نه به‌دلیل نامرتبط‌بودن یا ضعف اعتبار علمی. یافته‌های آنها در تدوین چارچوب مفهومی و بررسی شرایط انتقال روش‌ها مورد استفاده قرار گرفت، اما به‌عنوان شواهد مستقیم کاربرد یا اثربخشی داده‌افزایی در علم‌سنجی و در محاسبه فراوانی مطالعات هسته‌ای منظور نشد. این ۱۲ مطالعه از پژوهش حذف نشدند، بلکه به مجموعه مکمل D منتقل شدند تا در محاسبه فراوانی و درصد مطالعات هسته‌ای منظور نشوند. همچنین ۱۲ مطالعه واجد شرایط دیگر که در ترکیب اولیه حضور نداشتند، از همان پیکره ثابت شناسایی شدند و مطابق گزارش ممیزی، پس از بررسی محتوایی، کنترل شناسه‌های DOI و UT و تطبیق با اطلاعات مبدأ، به مجموعه هسته‌ای افزوده شدند. در نتیجه، مجموعه هسته‌ای اصلاح‌شده همچنان شامل ۴۳ مطالعه بود، اما ترکیب آن تغییر یافت؛ به‌گونه‌ای که شش مطالعه در گروه A، تعداد ۳۳ مطالعه در گروه B و چهار مطالعه در گروه C قرار گرفتند. افزون بر این، ۱۳ مطالعه در مجموعه روش‌شناختی مکمل D حفظ شدند که شامل ۱۲ مطالعه بازطبقه‌بندی‌شده از مجموعه اولیه

[1] Recovery Audit

و یک مطالعه مکمل شناسایی‌شده در مرحله ممیزی بود. شواهد این گروه برای طبقه‌بندی روش‌ها، بررسی شرایط انتقال، تبیین الزامات اعتبارسنجی و شناسایی مخاطرات استفاده شدند، اما در مخرج محاسبات مربوط به ۴۳ مطالعه هسته‌ای قرار نگرفتند. ثابت‌ماندن تعداد مطالعات هسته‌ای در عدد ۴۳، حاصل تغییر ترکیب مطالعات در مرحله بازنگری بود و نباید به معنای ثابت‌بودن هویت یا نقش مطالعات اولیه و نهایی تفسیر شود. برای جلوگیری از خلط میان فرایند تاریخی انتخاب و اصلاحات مرحله بازنگری، این دو جریان به‌صورت متمایز در شکل ۱ نمایش داده شدند. همچنین، با توجه به محفوظ‌نماندن سوابق کامل غربالگری اولیه، از انتساب قطعی مطالعات بازیابی‌شده به یکی از مراحل خروج در جریان تاریخی انتخاب خودداری شد.



شکل ۱. جریان انتخاب اولیه مطالعات و ممیزی بازیابی در مرحله بازنگری، با اقتباس از الگوی PRISMA 2020

بخش نخست شکل، جریان ثبت‌شده شناسایی و انتخاب اولیه مطالعات از مجموعه ۷۰۰۰ رکوردی WoSCC را نشان می‌دهد. بخش دوم به ممیزی و بازطبقه‌بندی مطالعات در مرحله بازنگری مقاله اختصاص دارد که بدون استخراج رکورد جدید از پایگاه انجام شد. مجموعه هسته‌ای نهایی از ۳۱ مطالعه حفظ‌شده از ترکیب اولیه و ۱۲ مطالعه افزوده‌شده در مرحله ممیزی تشکیل شد. همچنین ۱۳ مطالعه روش‌شناختی مکمل (گروه D)، شامل ۱۲ مطالعه بازطبقه‌بندی‌شده و یک مطالعه مکمل شناسایی‌شده در مرحله ممیزی، به‌عنوان مجموعه‌ای مستقل نگهداری شدند. این مطالعات و چهار منبع زمینه‌ای شناسایی‌شده خارج از پیکره ثابت، در محاسبه فراوانی و درصد مطالعات هسته‌ای منظور نشدند. تعداد رکوردها و مدارک در مراحل انتخاب اولیه، بر اساس گزارش تاریخی غربالگری ارائه شده است و نباید با جریان بازطبقه‌بندی مرحله بازنگری یکسان تلقی شود.

- **مطالعات شناسایی‌شده خارج از مجموعه خام**

در بازنگری منابع، مطالعات مرتبطی شناسایی شدند که در هفت فایل خامِ حاوی مجموعاً ۷۰۰۰ رکورد حضور نداشتند. برای جلوگیری از ورود انتخابی و پسینی مطالعات به پیکره ثابت مرور نظام‌مند، این منابع در فایل مکمل ثبت شدند و به مجموعه هسته‌ای راه نیافتند. استناد به آن‌ها صرفاً برای تکمیل زمینه نظری و معرفی تحولات جدید حوزه

بود؛ ازاین‌رو، در تعداد مطالعات، فراوانی‌ها، درصدها و استنتاج‌های مبتنی بر مجموعه ۴۳ مطالعه هسته‌ای منظور نشدند.

- **استخراج، سازماندهی و تحلیل داده‌ها**

برای هر مطالعه هسته‌ای، رکوردی ساختاریافته شامل مشخصات کتاب‌شناختی (نویسندگان، سال انتشار، عنوان، منبع و نوع مدرک)، شناسه‌ها و اطلاعات ردیابی (شناسه مطالعه، DOI، شناسه UT، فایل خام مبدأ، شماره ردیف و وضعیت بازیابی) و متغیرهای تحلیلی (لایه شواهد، صراحت کاربرد داده‌افزایی، زمینه داده، روش یا سازوکار، وظیفه تحلیلی، شیوه ارزیابی، نقش مطالعه، دلیل ورود و محدودیت تفسیری) ایجاد شد. اطلاعات در ماتریس «مطالعه–شاهد» در سه سطح «یافته مستقیم مطالعه اولیه»، «سنتز حاصل از چند مطالعه» و «استنتاج مفهومی نویسندگان» سازمان‌دهی شدند تا یافته‌های مستقیم با استنتاج‌های نویسندگان خلط نشوند. ویژگی‌های ۴۳ مطالعه هسته‌ای از نظر سال انتشار، نوع داده، زمینه کاربرد، خانواده روش، وظیفه تحلیلی و لایه شواهد با فراوانی و درصد توصیف شدند. مخرج محاسبه درصدهای لایه‌های A، B و C مجموعه ۴۳ مطالعه هسته‌ای بود و ۱۳ مطالعه مکمل لایه D و مطالعات خارج از پیکره ثابت در این محاسبات وارد نشدند. این شاخص‌ها صرفاً برای توصیف ساختار شواهد به کار رفتند و مبنای استنتاج اثربخشی، برتری روش یا رابطه علّی نبودند. توزیع لایه‌ها و نقش آن‌ها در سنتز شواهد در جدول ۲ ارائه شده است.

جدول ۲. ساختار نهایی مجموعه مطالعات و نقش لایه‌های شواهد پس از ممیزی بازیابی

لایه	تعریف عملیاتی	تعداد	سهم از هسته ۴۳ مطالعه	نقش در سنتز
A	داده‌افزایی صریح در داده‌ها یا فرآیندهای علمی	۶	٪ ۱۴٫۰	شواهد مستقیم داده‌افزایی
B	سازوکارهای مستقیم در داده‌ها و شبکه‌های علمی	۳۳	٪ ۷۶٫۷	شواهد سازوکاری؛ نه لزوماً داده‌افزایی
C	کیفیت داده و مرزبندی مفهومی	۴	٪ ۹٫۳	تعیین منشأ مسئله و حدود داده‌افزایی
هسته اصلی	A + B + C	۴۳	٪ ۱۰۰	مبنای سنتز اصلی
D	شواهد روش‌شناختی مکمل	۱۳	—	انتقال روش، اعتبارسنجی و ریسک
Post-search	مطالعات مرتبط خارج از پیکره ثابت	۴	—	فقط زمینه و تحولات جدید؛ خارج از محاسبات نظام‌مند

- **تحلیل محتوای کیفی جهت‌دار**

تحلیل محتوای کیفی جهت‌دار با رویکرد قیاسی– استقرایی انجام شد. دفترچه کد اولیه بر اساس تعریف عملیاتی داده‌افزایی، پرسش‌های پژوهش و مفاهیم تکرارشونده در پیشینه تدوین شد و ده محور را دربر گرفت: نوع مسئله داده‌ای، نوع داده، عملیات یا سازوکار، خانواده روش، وظیفه تحلیلی، شیوه اعتبارسنجی، پیامد گزارش‌شده، محدودیت یا ریسک، منشأ شواهد و وضعیت معرفتی خروجی. در جریان خوانش مطالعات، مفاهیم پیش‌بینی‌نشده پس از بررسی مصداق و مرز مفهومی، به‌صورت استقرایی به دفترچه کد افزوده شدند. کدهای هم‌پوشان در مقوله‌های سطح بالاتر سازمان‌دهی و شواهد ناسازگار یا محدودکننده نیز حفظ شدند. ازآنجاکه روش پژوهش تحلیل محتوای جهت‌دار بود و نه نظریه داده‌بنیاد، از اصطلاحات «کدگذاری باز، محوری و گزینشی» استفاده نشد. در مرحله بازنگری، تصمیم‌های ورود و

طبقه‌بندی مطالعات بازیابی‌شده بر اساس همان تعاریف عملیاتی و دفترچه کد کنترل و برای هر مطالعه، دلیل ورود، نقش در استنتاج و محدودیت انتقال ثبت شد. بدین‌ترتیب، تصمیم‌ها صرفاً بر شباهت واژگانی استوار نبودند و قابلیت ردیابی مفهومی و رکوردی داشتند.

- **ارزیابی کفایت و کیفیت شواهد**

با تغییر ترکیب مطالعات پس از ممیزی بازیابی، امتیازهای ارزیابی کیفیت مجموعه پیشین به‌عنوان نتایج ارزیابی مجموعه نهایی گزارش نشدند. همچنین برای مطالعات بازیابی‌شده، هیچ امتیاز QuADS[1] (Harrison et al., 2021) یا امتیاز فنی از روی عنوان، چکیده یا فراداده برآورد نشد. کفایت شواهد هر مطالعه بر اساس اطلاعات قابل مشاهده و ردیابی شامل شفافیت منبع و نوع داده، ماهیت عملیات یا سازوکار، وجود روش پایه یا مقایسه در صورت گزارش، طراحی اعتبارسنجی، شاخص‌های ارزیابی، تفکیک داده مرجع از داده مصنوعی یا استنباطی و محدودیت‌های انتقال نتایج بررسی شد. این اطلاعات در ماتریس «مطالعه–شاهد» و فایل داده مکمل ثبت شدند. با توجه به ناهمگنی طرح‌ها و ماهیت مطالعات، از امتیاز کیفیت کلی یا آستانه عددی ثابت برای حذف، وزن‌دهی یا طبقه‌بندی مطالعات به «باکیفیت» و «کم‌کیفیت» استفاده نشد؛ بلکه قوت و محدودیت شواهد هر مطالعه متناسب با ادعایی که از آن پشتیبانی می‌کرد، در تفسیر نتایج لحاظ شد.

- **ساخت چارچوب مفهومی**

چارچوب مفهومی با تلفیق نظام‌مند روابط استخراج‌شده از مطالعات تدوین شد. در هر مطالعه، ابتدا مسئله و نوع داده و سپس عملیات یا سازوکار، وظیفه تحلیلی، شیوه اعتبارسنجی، پیامد و محدودیت یا ریسک بررسی شد. این مؤلفه‌ها مبنای مسیر تصمیم‌گیری از تشخیص مسئله و ضرورت مداخله تا انتخاب راهکار متناسب، اعتبارسنجی و ارزیابی پیامدها قرار گرفتند؛ به‌گونه‌ای که پاک‌سازی، بازیابی و یکپارچه‌سازی از داده‌افزایی و روش‌های استنباطی تفکیک شوند. در تبیین روابط چارچوب، شواهد کاربرد مستقیم داده‌افزایی (لایه A)، سازوکارهای محاسباتی در داده‌های علمی (لایه B) و کیفیت داده و مرزبندی مفهومی (لایه C) از شواهد روش‌شناختی مکمل (لایه D) متمایز شدند. روابط عمدتاً متکی بر لایه D صرفاً ظرفیت روش‌شناختی قابل انتقال و نیازمند آزمون مستقل تلقی شدند، نه یافته‌های تثبیت‌شده در علم‌سنجی. چارچوب حاصل سنتزی و تصمیم‌محور است و ادعایی درباره اعتبارسنجی تجربی یا عملکرد پیش‌بینانه آن مطرح نمی‌شود.

- **حدود استنتاج و کنترل بیش‌تعمیم**

برای کنترل بیش‌تعمیم، سه قاعده رعایت شد: نخست، عدم‌توازن، کم‌تراکمی یا ناهمگونی شبکه صرفاً در صورت وجود شواهدی از نقص پوشش، خطای فراداده، فقدان رکورد یا سوگیری اندازه‌گیری، نشانه مسئله داده‌ای تلقی شد؛ زیرا این ویژگی‌ها ممکن است بازتاب ساختار واقعی نظام علمی باشند. دوم، داده‌ها و پیوندهای مصنوعی یا پیش‌بینی‌شده

[1] Quality Assessment with Diverse Studies (QuADS)

از داده‌های مشاهده‌شده متمایز نگه داشته شدند؛ رابطه پیش‌بینی‌شده، هرچند برای توصیه، تحلیل سناریو یا مدل‌سازی سودمند است، بدون تأیید مستقل نباید به‌عنوان رابطه تاریخی مشاهده‌شده ثبت یا گزارش شود. سوم، شواهد حوزه‌های روش‌شناختی مجاور صرفاً برای شناسایی سازوکارها، معیارهای ارزیابی، محدودیت‌ها و ریسک‌های قابل انتقال استفاده شدند، نه اثبات اثربخشی در علم‌سنجی. دامنه مرور به مجموعه ثابت ۷۰۰۰ رکوردی و معیارهای عملیاتی اعلام‌شده محدود بود؛ بنابراین، نه پوشش تمام ادبیات جهانی ادعا می‌شود و نه چارچوب مفهومی نهایی تضمینی برای اثربخشی داده‌افزایی در همه مسائل علم‌سنجی ارائه می‌دهد.

- **محدودیت‌های پژوهش**

نخست، شناسایی مطالعات به WoSCC و بازه زمانی ۲۰۱۸ تا ۲۰۲۶ محدود بود؛ بنابراین، یافته‌ها به تمام ادبیات داده‌افزایی تعمیم‌پذیر نیستند. چهار منبع مرتبط خارج از پیکره ثابت نیز صرفاً برای تکمیل زمینه نظری و بررسی تحولات اخیر استفاده شدند و در محاسبات مطالعات هسته‌ای وارد نشدند. اگرچه دسترسی به هفت فایل اولیه امکان بازبینی تعداد رکوردها، کنترل یکتایی شناسه‌ها، تأیید ممیزی زمانی و ردیابی مطالعات منتخب در پیکره اولیه را فراهم کرد، تاریخچه کامل جست‌وجوی اصلی، شامل عبارت دقیق اجراشده و تمامی تنظیمات آن، در مستندات موجود محفوظ نمانده بود. همچنین، سوابق تفصیلی تمام تصمیم‌های غربالگری اولیه در دسترس نبود. ازاین‌رو، تأیید ساختار و ترکیب پیکره اولیه نباید معادل بازتولید کامل راهبرد تاریخی جست‌وجو یا تمام مراحل انتخاب مطالعات تلقی شود. دوم، ناهمگنی مطالعات از نظر نوع داده، وظیفه تحلیلی، روش محاسباتی و اعتبارسنجی، تجمیع کمّی نتایج در قالب فراتحلیل و مقایسه مستقیم اثربخشی روش‌ها را ناموجه می‌ساخت. ازاین‌رو، شواهد با تحلیل محتوای کیفی جهت‌دار و ماتریس «مطالعه–شاهد» بررسی شدند و نتایج صرفاً برای تبیین دامنه کاربرد، شرایط استفاده و حدود اعتبار شواهد به کار رفتند، نه تعیین برتری عمومی روش‌ها. سوم، تحول سریع روش‌های مولد، مدل‌های زبانی بزرگ و یادگیری گراف، تعمیم یافته‌ها به کاربردهای آینده را محدود می‌کند. اگرچه چارچوب پیشنهادی به الگوریتم مشخصی وابسته نیست و بر تشخیص منشأ مسئله، انتخاب مداخله متناسب، ردیابی منشأ داده، حفظ وضعیت معرفتی و اعتبارسنجی مستقل استوار است، کارایی و تعمیم‌پذیری آن در انواع داده‌ها و شبکه‌های علم‌سنجی هنوز به‌صورت تجربی ارزیابی نشده است. بنابراین، چارچوب حاضر باید صورت‌بندی مفهومی مبتنی بر شواهد موجود تلقی شود، نه مدلی با عملکرد تجربی تأییدشده.

یافته‌های پژوهش

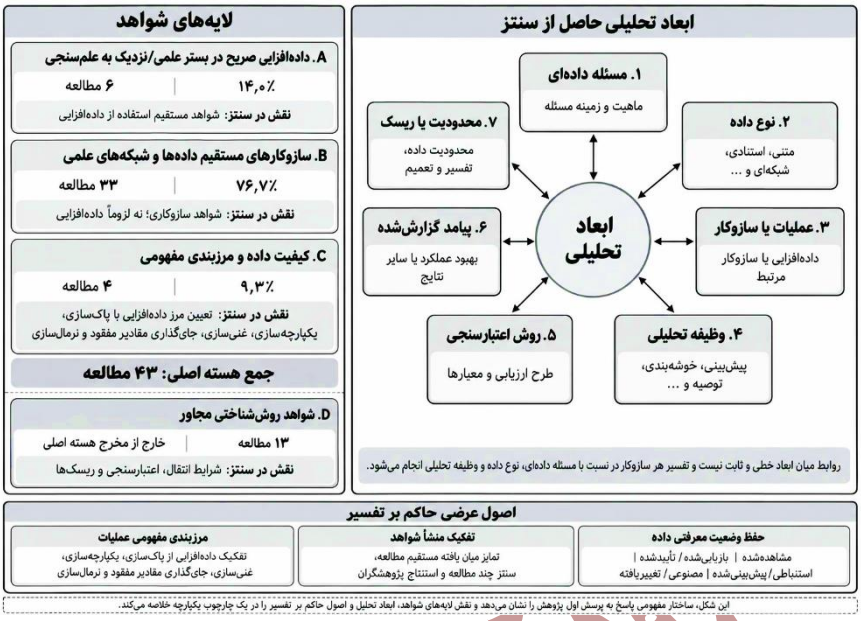
یافته‌ها از سنتز ۴۳ مطالعه هسته‌ای، با رعایت مرزبندی‌های مفهومی و سطوح شواهد تعریف‌شده در روش‌شناسی، استخراج شدند. در تحلیل نتایج، شواهد کاربرد مستقیم داده‌افزایی، سازوکارهای محاسباتی مرتبط با داده‌ها و شبکه‌های علمی، شواهد کیفیت و مرزبندی داده و مطالعات روش‌شناختی مکمل از یکدیگر تفکیک شدند. همچنین، یافته‌های

زودآیند ویرایش نشده

مستقیم مطالعات از سنتز میان‌مطالعه‌ای و استنتاج مفهومی پژوهشگران متمایز شدند تا هر نتیجه متناسب با پشتوانه شواهدی آن تفسیر شود. یافته‌ها به‌ترتیب هفت پرسش پژوهش ارائه می‌شوند و شواهد مکمل صرفاً برای تبیین شرایط انتقال روش‌ها، محدودیت‌ها و الزامات اعتبارسنجی به کار می‌روند.

پا سخ به پر سش اول پژوهش. بر ا ساس تحلیل محتوای کیفی جهت‌دار، ساختار شواهد و ابعاد ا صلی مطالعات مرتبط با داده‌افزایی در علم‌سنجی و شبکه‌های دانشی چگونه قابل سازمان‌دهی است؟

تحلیل محتوای کیفی جهت‌دار ۴۳ مطالعه هسته‌ای نشان داد که شواهد مرتبط با داده‌افزایی در علم‌سنجی و شبکه‌های دانشی، ساختاری یکدست ندارند، بلکه در سه لایه هسته‌ای و یک لایه روش‌شناختی مکمل سازمان می‌یابند. در هسته اصلی، ٦ مطالعه شواهد مستقیم استفاده صریح از داده‌افزایی را در بسترهای علمی مرتبط با علم‌سنجی فراهم کردند؛ ۳۳ مطالعه به سازوکارهای محاسباتی مرتبط با داده‌ها و شبکه‌های علمی، مانند پیش‌بینی پیوند، تعبیه شبکه، توصیه استناد و بازنمایی گراف اختصاص داشتند؛ و ٤ مطالعه به کیفیت داده و مرزبندی مفهومی داده‌افزایی با عملیات مجاور، مانند پاک‌سازی، یکپارچه‌سازی، غنی‌سازی، جای‌گذاری مقادیر مفقود و نرمال‌سازی، پرداختند. همچنین ۱۳ مطالعه روش‌شناختی مکمل برای تبیین شرایط انتقال، معیارهای اعتبارسنجی و ریسک‌های احتمالی به کار رفتند، اما در مخرج تحلیل اصلی وارد نشدند. سنتز نهایی نشان داد که این مجموعه شواهد حول هفت بُعد تحلیلی سازمان می‌یابد: مسئله داده‌ای، نوع داده، عملیات یا سازوکار، وظیفه تحلیلی، روش اعتبارسنجی، پیامد گزارش‌شده و محدودیت یا ریسک. این ابعاد به‌صورت منفک تفسیر نمی‌شوند، بلکه معنای هر سازوکار در پیوند با نوع مسئله داده‌ای، ماهیت داده و هدف تحلیلی آن شکل می‌گیرد. ازاین‌رو، موفقیت یک روش در یک وظیفه خاص، لزوماً به معنای قابلیت تعمیم آن به همه مسائل علم‌سنجی یا بازسازی روابط واقعی شبکه‌های علمی نیست. در تفسیر این سنتز، سه اصل رعایت شد: مرزبندی مفهومی داده‌افزایی از عملیات مجاور، تفکیک شواهد مستقیم از شواهد روش‌شناختی مکمل، و حفظ وضعیت معرفتی داده‌ها و روابط در سطوح مشاهده‌شده، بازیابی‌شده/تأییدشده، استنباطی/پیش‌بینی‌شده و مصنوعی/ تغییریافته. بر این اساس، پاسخ به پرسش نخست نشان می‌دهد که داده‌افزایی در این حوزه، بیش از آنکه یک تکنیک منفرد باشد، قلمرویی چندلایه از شواهد، سازوکارها و ملاحظات تفسیری است که باید با توجه به نوع داده، مسئله و سطح شواهد تحلیل شود. شکل ۲ این ساختار چندلایه را همراه با ابعاد اصلی تحلیل و اصول حاکم بر تفسیر، به‌صورت یکپارچه نشان می‌دهد.



شکل ۲. ساختار چندلایه شواهد و ابعاد تحلیلی مطالعات داده‌افزایی در علم‌سنجی و شبکه‌های دانشی

این شکل، سازمان‌یابی شواهد مرتبط با داده‌افزایی را در چهار لایه شامل داده‌افزایی صریح، سازوکارهای محاسباتی مرتبط با داده‌ها و شبکه‌های علمی، کیفیت داده و مرزبندی مفهومی، و شواهد روش‌شناختی مکمل نشان می‌دهد. شکل همچنین هفت بُعد تحلیلی و سه اصل تفسیری حاکم بر سنتز یافته‌ها را نمایش می‌دهد.

پاسخ به پرسش دوم پژوهش. چه الگوهایی از مسائل داده‌ای، از جمله محدودیت پوشش پایگاه‌ها، ناسازگاری و نویز فراداده، آثار ناشی از نحوه ساخت و بازنمایی شبکه و عدم‌توازن کلاس‌ها در برخی وظایف طبقه‌بندی، در داده‌ها و تحلیل‌های علم‌سنجی شناسایی شده‌اند و چه پیامدهایی برای تفسیر نتایج در شبکه‌های دانشی دارند؟

سنتز شواهد مرتبط با این پرسش، پنج الگوی مسائل داده‌ای را مشخص کرد: ناهمسانی پوشش پایگاه‌ها، مسائل کیفیت و سازگاری فراداده، آثار ساخت و بازنمایی شبکه، عدم‌توازن کلاس‌ها در برخی وظایف تحلیلی و کمبود نمونه یا محدودیت داده‌های در دسترس برای پیش‌بینی و طبقه‌بندی. این الگوها منشأ متفاوتی دارند و همگی مصداق «نقص داده» نیستند. تفاوت پوشش پایگاه‌هایی مانند WoS، Scopus و OpenAlex می‌تواند به بازنمایی مجموعه‌های متفاوتی از استنادها و روابط علمی منجر شود؛ بااین‌حال، بازیابی رکورد از منبع مستقل یا ادغام داده‌های پایگاه‌ها، در قلمرو بازیابی و یکپارچه‌سازی قرار می‌گیرد، نه داده‌افزایی. همچنین، خطاهای فراداده، ناهمگونی ثبت نام‌ها و ابهام در تطبیق موجودیت‌ها، پیش از هر مداخله به پاک‌سازی، استانداردسازی یا رفع ابهام نیاز دارند. در الگوی سوم، شیوه ساخت، تجمیع یا بازنمایی شبکه ممکن است روابطی مشتق‌شده ایجاد کند که به همان صورت در داده اولیه ثبت نشده‌اند و بر شاخص‌ها و رتبه‌بندی‌ها اثر بگذارد؛ بنابراین، تفسیر روابط باید با توجه به شیوه ساخت شبکه انجام شود. در مقابل، عدم‌توازن کلاس‌ها و کمبود نمونه آموزشی می‌توانند در برخی وظایف مشخص، عملکرد مدل را محدود کنند. این

مسائل در مطالعات مرتبط با طبقه‌بندی زمینه استناد، توصیه مقاله با تعاملات ثبت‌شده محدود، طبقه‌بندی اسناد ثبت اختراع با نمونه‌های کم یا نامتوازن و پیش‌بینی فناوری با داده‌های برچسب‌خورده محدود مشاهده شدند. بااین‌حال، محدودیت نمونه آموزشی را نباید با توزیع نامتقارن استنادها، همکاری‌ها یا اندازه اجتماعات علمی یکسان دانست؛ زیرا این ویژگی‌ها ممکن است ساختار واقعی نظام علم را بازتاب دهند. برآیند یافته‌ها نشان می‌دهد که تشخیص منشأ مسئله باید بر انتخاب مداخله تقدم داشته باشد. داده‌افزایی تنها زمانی می‌تواند گزینه‌ای روش‌شناختی محسوب شود که محدودیت داده در ارتباط با وظیفه تحلیلی مشخص احراز شده باشد و کیفیت و اعتبار داده‌های افزوده نیز مستقلاً ارزیابی شود.

جدول ۳. الگوهای مسائل داده‌ای، پیامدهای تحلیلی و جایگاه مداخله در علم‌سنجی

تفسیر و مداخله مناسب	پیامد اصلی برای تحلیل	الگوی مسئله
بازیابی، اعتبارسنجی یا یکپارچه‌سازی بین‌پایگاهی؛ نه داده‌افزایی	تفاوت در رکوردها و روابط استنادی و در نتیجه تفاوت در ساختار شبکه	ناهمسانی پوشش میان پایگاه‌ها
پاک‌سازی، استانداردسازی، رفع ابهام و یکپارچه‌سازی	خطا در تطبیق، شناسایی یا شکل‌گیری روابط شبکه	ناسازگاری فراداده و ابهام موجودیت‌ها
کنترل اعتبار بازنمایی و تفکیک رابطه مشاهده‌شده از رابطه مشتق‌شده	ایجاد یا تغییر روابط مشتق‌شده و اثر بر شاخص‌ها و رتبه‌بندی‌ها	نحوه ساخت و بازنمایی شبکه
استفاده از روش‌های مدیریت عدم‌توازن؛ داده‌افزایی فقط در صورت تولید یا تبدیل واقعی نمونه و پس از اعتبارسنجی	افت یا سوگیری عملکرد مدل طبقه‌بندی	عدم‌توازن کلاس‌ها در وظایف مشخص
بررسی مشروط داده‌افزایی؛ نتایج به همان نوع داده و وظیفه تحلیلی محدود می‌ماند	محدودیت در آموزش یا تعمیم مدل	کمبود نمونه یا محدودبودن داده‌های در دسترس

جدول ۳ نشان می‌دهد که مسائل داده‌ای در علم‌سنجی منشأ واحدی ندارند. تفاوت پوشش پایگاه‌ها، مشکلات فراداده، آثار ساخت و بازنمایی شبکه، عدم‌توازن کلاس‌ها و کمبود نمونه آموزشی، ماهیت متفاوتی دارند و ضرورت مداخله و نوع آن باید برای هر مورد جداگانه ارزیابی شود. همچنین، کم‌تراکمی یا نابرابری ساختاری شبکه، بدون شواهد مستقل از نقص یا خطا، نشانه کمبود داده نیست؛ زیرا مداخله ناموجه می‌تواند ساختار واقعی پدیده و تفسیر نتایج را تحریف کند.

پاسخ به پرسش سوم پژوهش. چه دسته‌هایی از روش‌های داده‌افزایی در مطالعات بررسی‌شده به‌طور مستقیم به کار رفته‌اند و شواهد روش‌شناختی مکمل چه ظرفیت‌هایی را برای انتقال مشروط این روش‌ها به مسائل علم‌سنجی نشان می‌دهند؟

سنتز مطالعات، چهار خانواده اصلی روش‌های داده‌افزایی را بر اساس سازوکار غالب مشخص کرد: روش‌های آماری و تولید داده مصنوعی، مدل‌های مولد مبتنی بر یادگیری عمیق، داده‌افزایی ساختاری و گراف‌محور، و داده‌افزایی متنی مبتنی بر مدل‌های زبانی. پشتوانه تجربی این خانواده‌ها یکسان نیست؛ شش مطالعه هسته‌ای کاربرد مستقیم داده‌افزایی را در داده‌ها یا فرایندهای علمی گزارش کرده‌اند، درحالی‌که مطالعات مکمل، ظرفیت‌های روش‌شناختی قابل انتقال و نیازمند اعتبارسنجی مستقل را نشان می‌دهند.

گروه نخست، روش‌های آماری و تولید داده مصنوعی: این گروه شامل تولید سری‌های زمانی مصنوعی، بازنمونه‌گیری،

بیش‌نمونه‌گیری مصنوعی اقلیت[1]، تولید مبتنی بر توزیع‌های آماری و روش‌های مبتنی بر کوپولا[2]است. در شواهد مستقیم، کاوثن و همکاران (Cauthen et al., 2022) با تولید سری‌های زمانی مصنوعی انتشارات، داده‌های آموزشی طبقه‌بندی بلوغ فناوری را گسترش دادند. سایر روش‌های این گروه عمدتاً ظرفیت‌هایی برای مواجهه با محدودیت نمونه یا عدم‌توازن در وظایف مشخص فراهم می‌کنند؛ بااین‌حال، وفاداری آماری به‌تنهایی اعتبار داده افزوده را در شبکه‌های علمی تضمین نمی‌کند. گروه دوم، مدل‌های مولد و یادگیری عمیق: شبکه‌های مولد تخاصمی (جی‌ای‌ان)[3]در مطالعات مستقیم برای افزایش داده‌های ثبت اختراع در طبقه‌بندی و پیش‌بینی فناوری‌های نوظهور به کار رفته‌اند (Xu et al., 2022; Zhou et al., 2020). این شواهد، کاربرد داده‌افزایی مولد را در وظایف مشخص پشتیبانی می‌کنند، اما به معنای اعتبار بازسازی روابط تاریخی شبکه‌های استنادی یا هم‌نویسندگی نیستند. همچنین، استفاده از مدل مولد، بدون تولید و به‌کارگیری داده افزوده در فرایند تحلیل، به‌خودی‌خود داده‌افزایی محسوب نمی‌شود. گروه سوم، داده‌افزایی ساختاری و گراف‌محور: شواهد مستقیم شامل داده‌افزایی در توصیه مقاله و تحلیل زیرگراف‌های دانش ثبت اختراع برای شناسایی فناوری‌های نوظهور است (Y. Liu et al., 2025; C. Zhang et al., 2025). مطالعات مکمل نیز روش‌های داده‌افزایی گرافی، داده‌افزایی تطبیقی[4]و تغییر کنترل‌شده ساختار گراف را معرفی کرده‌اند. انتقال این روش‌ها به شبکه‌های علمی مستلزم ارزیابی ساختاری و معنایی است؛ زیرا افزودن یا تغییر گره‌ها و یال‌ها می‌تواند شاخص‌ها و معنای روابط را تغییر دهد. بنابراین، روابط تولیدشده باید از روابط مشاهده‌شده متمایز باقی بمانند. گروه چهارم، داده‌افزایی متنی و مبتنی بر مدل‌های زبانی: در شواهد مستقیم، داده‌افزایی مبتنی بر مدل زبانی بزرگ برای تولید نمونه‌های مصنوعی داوری همتا و آموزش مدل شناسایی داوری‌های ناکافی استفاده شده است (H. Zhang et al., 2025). شواهد مکمل نیز روش‌های بازنویسی، نمونه‌گیری، تولید مبتنی بر پرامپت و تولید متن مصنوعی را پوشش می‌دهند. این رویکردها برای وظایف مبتنی بر متن علمی قابل بررسی‌اند، اما انتقال نتایج آن‌ها به شبکه‌های استنادی و همکاری علمی نیازمند آزمون مستقل است. در مجموع، شواهد مستقیم از کاربرد داده‌افزایی در وظایف معین طبقه‌بندی، پیش‌بینی، توصیه و پردازش متن علمی پشتیبانی می‌کنند؛ اما هیچ خانواده روشی را نمی‌توان راه‌حلی عمومی برای مسائل علم‌سنجی دانست. انتخاب و انتقال روش باید به منشأ مسئله، نوع داده، وظیفه تحلیلی و اعتبارسنجی متناسب با آن وابسته باشد. جدول ٤ خانواده‌های روش، شواهد مستقیم و حدود انتقال آن‌ها را نشان می‌دهد.

جدول ۴. رویکردهای داده‌افزایی و دامنه شواهد موجود برای کاربرد در علم‌سنجی

گروه روش	نمونه روش‌ها و سازوکارها	شواهد مستقیم در بستر علمی	دامنه شواهد و محدودیت انتقال

[1] Synthetic Minority Over-sampling Technique (SMOTE)
[2] Copula
[3] Generative Adversarial Network (GAN)
[4] Adaptive Augmentation

قابل بررسی در شرایط محدودیت نمونه در وظایف مشخص؛ وفاداری آماری به‌تنهایی برای داده‌های شبکه‌ای کافی نیست	داده‌افزایی سری زمانی کتاب‌سنجی برای طبقه‌بندی بلوغ فناوری	سری زمانی مصنوعی، بازنمونه‌گیری، بیش‌نمونه‌گیری مصنوعی اقلیت، کوپولا، نویز کنترل‌شده	آماری و تولید داده مصنوعی
شواهد مستقیم وظیفه‌محور است؛ استفاده از مدل مولد به‌خودی‌خود داده‌افزایی محسوب نمی‌شود	جی‌ای‌ان برای افزایش داده در طبقه‌بندی پتنت و پیش‌بینی فناوری	جی‌ای‌ان، تولید نمونه مصنوعی، پالایش نمونه‌های تولیدشده	مدل‌های مولد و یادگیری عمیق
کاربرد در شبکه‌های علمی نیازمند کنترل ساختار، معنا و تفکیک روابط مصنوعی از روابط مشاهده‌شده است	داده‌افزایی در توصیه مقاله و طبقه‌بندی زیرگراف‌های پتنت	Self-supervised Graph Augmentation ،Graph Augmentation ،Adaptive Augmentation	ساختاری و گراف‌محور
با مسائل مبتنی بر متن علمی تناسب نزدیک‌تری دارد؛ انتقال به شبکه‌های استنادی یا همکاری نیازمند آزمون مستقل است	داده‌افزایی مبتنی بر ال‌ال‌ام در داده‌های داوری علمی	بازنویسی، نمونه‌گیری، تولید مبتنی بر پرامپت، متن مصنوعی	متنی و مبتنی بر مدل‌های زبانی

جدول ٤ دامنه شواهد را از کاربرد مستقیم داده‌افزایی در داده‌های علمی تا ظرفیت روش‌های حوزه‌های مجاور نشان می‌دهد. هدف این طبقه‌بندی، رتبه‌بندی روش‌ها نیست، بلکه تفکیک میزان آزمودگی و ضرورت اعتبارسنجی مستقل آن‌ها در مسائل علم‌سنجی است. بازنمونه‌گیری، بیش‌نمونه‌گیری مصنوعی اقلیت، تولید داده مبتنی بر توزیع‌های آماری و روش‌های مبتنی بر کوپولا از جمله رویکردهای مطرح‌شده در شواهد روش‌شناختی مکمل‌اند.

پاسخ به پرسش چهارم پژوهش. چه سازوکارهایی برای پیش‌بینی، توصیه، استنباط و بازنمایی روابط در شبکه‌های استنادی، هم‌نویسندگی، موضوعی و دیگر شبکه‌های دانشی گزارش شده‌اند و خروجی این سازوکارها از نظر وضعیت معرفتی داده یا رابطه چگونه باید تفسیر شود؟

سنتز مطالعات، پنج گروه از سازوکارهای محاسباتی در شبکه‌های دانشی را مشخص کرد: توصیه استناد و مقاله، پیش‌بینی پیوند، پیش‌بینی یا توصیه همکاری علمی، تعبیه و بازنمایی شبکه، و ساخت یا غنی‌سازی بازنمایی‌های مبتنی بر گراف دانش. این گروه‌ها از نظر کارکرد و ماهیت خروجی یکسان نیستند و کاربرد آن‌ها لزوماً به معنای داده‌افزایی نیست. در شبکه‌های استنادی، سامانه‌های توصیه با بهره‌گیری از اطلاعات متنی، ساختار استناد، تعبیه شبکه یا گراف دانش، منابع مرتبط را پیشنهاد می‌کنند؛ درحالی‌که پیش‌بینی پیوند، احتمال وجود یا شکل‌گیری رابطه را برآورد می‌کند. خروجی این سازوکارها ماهیت پیشنهادی یا استنباطی دارد و حتی با عملکرد مناسب مدل، نباید استناد تاریخی مفقود یا رابطه واقعی بازیابی‌شده تلقی شود. در شبکه‌های هم‌نویسندگی نیز اطلاعات ساختاری، موضوعی و علایق پژوهشی برای پیشنهاد یا پیش‌بینی همکاری به کار می‌روند؛ اما رابطه پیشنهادی تا زمان وقوع یا تأیید مستقل، در شمار روابط مشاهده‌شده نیست. روش‌های تعبیه شبکه و بازنمایی در فضای نهفته، اطلاعات ساختاری و معنایی را به ویژگی‌ها یا بردارهای مشتق‌شده تبدیل می‌کنند؛ بنابراین، شباهت محاسبه‌شده الزاماً بیانگر وجود رابطه واقعی نیست. در شبکه‌های موضوعی و گراف‌های دانش نیز ترکیب منابع، استخراج روابط و بازنمایی معنایی برای نقشه‌برداری علم، پیش‌بینی موضوعات و تحلیل مسیرهای تحول فناوری به کار می‌رود. در این ساختارها، منشأ روابط مشاهده‌شده، استخراج‌شده و استنباطی باید قابل ردیابی باشد؛ زیرا حضور یک یال در گراف یکپارچه، مشاهده مستقیم آن در منبع اولیه را اثبات نمی‌کند. در بخشی از مطالعات مستقیم، داده‌افزایی گرافی در توصیه مقاله و طبقه‌بندی زیرگراف‌های ثبت اختراع به کار

رفته است. شواهد روش‌شناختی مکمل نیز تغییر کنترل‌شده ساختار گراف، افزایش نمونه و راهبردهای تطبیقی را بررسی کرده‌اند. خروجی‌های مصنوعی یا تغییریافته باید از داده مرجع متمایز و از نظر ساختاری و معنایی اعتبارسنجی شوند. سایر سازوکارهای محاسباتی نیز تنها زمانی مصداق داده‌افزایی‌اند که خروجی آن‌ها در مرحله بعد به‌عنوان داده یا بازنمایی افزوده استفاده شود. در سنتز پژوهش، چهار وضعیت معرفتی از یکدیگر تفکیک شد: «مشاهده‌شده» (ثبت مستقیم در منبع مرجع)، «بازیابی‌شده و مستقلاً تأییدشده» (بازیابی داده غایب از مجموعه اولیه و تأیید آن با شاهد مستقل)، «استنباطی/پیش‌بینی‌شده» (حاصل برآورد مدل) و «مصنوعی/تغییریافته» (حاصل تولید یا تبدیل الگوریتمی). بر این اساس، دقت پیش‌بینی یا بهبود عملکرد مدل، به‌تنهایی وضعیت معرفتی یک رابطه را از «استنباطی» به «مشاهده‌شده» یا «بازیابی‌شده» تغییر نمی‌دهد. حفظ این تمایز برای تفسیر معتبر روابط و شاخص‌های شبکه‌های علمی ضروری است.

جدول ۵. سازوکارهای شبکه‌ای، ماهیت خروجی و حدود تفسیر معرفتی

سازوکار	نمونه مطالعات هسته‌ای/مکمل	خروجی اصلی	وضعیت معرفتی غالب	حدود تفسیر
توصیه استناد و مقاله	C01, C09, C12, C18, C20, C26, C28, C33, C34	استناد یا مقاله پیشنهادی	استنباطی/ پیش‌بینی‌شده	پیشنهاد مدل است و بدون تایید مستقل، رابطه تاریخی یا استناد مفقود محسوب نمی‌شود
پیش‌بینی پیوند	C11, C16, C31, C39	پیوند احتمالی یا آینده	استنباطی/ پیش‌بینی‌شده	احتمال وجود یا شکل‌گیری رابطه را برآورد می‌کند؛ معادل رابطه مشاهده‌شده نیست
پیش‌بینی و توصیه همکاری	C13, C15, C27, C35, C37, C38	همکاری بالقوه یا پیشنهادی	استنباطی/ پیش‌بینی‌شده	بیانگر امکان یا احتمال همکاری است، نه همکاری تاریخی تاییدشده
تعبیه و بازنمایی شبکه	C11, C19, C22, C23, C25, C29, C36	بردار، فاصله، شباهت یا بازنمایی ساختاری	مصنوعی/ تغییریافته در سطح بازنمایی	بازنمایی محاسباتیِ مشتق‌شده از داده است و نباید خودبه‌خود به‌عنوان رابطه واقعی تفسیر شود
گراف دانش و بازنمایی ساختاری–معنایی	C07, C14, C17, C21, C26, C30	ساختار یکپارچه، ویژگی یا رابطه استخراج‌شده/ استنباطی	وابسته به منشأ: مشاهده‌شده یا استنباطی/ پیش‌بینی‌شده	وضعیت هر رابطه باید بر اساس منشأ آن حفظ شود؛ یکپارچگی گراف به معنای مشاهده‌شده بودن همه روابط نیست
داده‌افزایی گرافی	C01, C03, M07, M09, M10	نمونه، زیرگراف یا ساختار تغییریافته	مصنوعی/ تغییریافته	خروجی باید از داده مشاهده‌شده متمایز و پیش از استفاده تحلیلی از نظر ساختاری و معنایی اعتبارسنجی شود

شواهد، کاربرد روش‌های محاسباتی را در پیش‌بینی، توصیه و بازنمایی روابط علمی نشان می‌دهند، اما بازسازی قطعی روابط پنهان را تأیید نمی‌کنند. خروجی این روش‌ها عمدتاً روابط بالقوه یا بازنمایی‌ها و ساختارهای متناسب با وظایف تحلیلی‌اند؛ بنابراین، خروجی‌های استنباطی، پیش‌بینی‌شده و مصنوعی باید متناسب با منشأ و ماهیت خود اعتبارسنجی شوند و بدون تأیید مستقل، معادل روابط مشاهده‌شده تلقی نشوند.

پاسخ به پرسش پنجم پژوهش. شواهد و یافته‌های روش‌شناختی حوزه‌های مجاور، از جمله پردازش زبان طبیعی، یادگیری گراف و تولید داده مصنوعی، چه شرایط، ملاحظات و الزامات اعتبارسنجی را برای انتقال

روش‌های داده‌افزایی به مسائل علم‌سنجی نشان می‌دهند؟

سنتز شواهد این پرسش بر ۱۵ مطالعه شامل دو مطالعه هسته‌ای مرتبط با روش‌های محاسباتی در داده‌ها یا فرایندهای علمی و ۱۳ مطالعه روش‌شناختی مکمل استوار بود. مطالعات مکمل برای شناسایی ظرفیت روش‌ها، شرایط انتقال، الزامات اعتبارسنجی و مخاطرات احتمالی استفاده شدند؛ بنابراین، بخش عمده شواهد ماهیت انتقالی دارد و اثربخشی مستقیم داده‌افزایی در علم‌سنجی را اثبات نمی‌کند. برآیند مطالعات، شش ملاحظه اصلی را مشخص کرد.

نخست، تناسب روش با نوع داده و وظیفه تحلیلی: موفقیت روش‌های داده‌افزایی در طبقه‌بندی متن یا گره، تشخیص ناهنجاری و تولید داده جدولی، به‌تنهایی اعتبار آن‌ها را در شبکه‌های استنادی و همکاری علمی تأیید نمی‌کند. تفاوت ماهیت داده و معنای روابط، آزمون مستقل روش را در داده و وظیفه هدف ضروری می‌سازد. دوم، کنترل کیفیت داده مصنوعی: شواهد نشان دادند که نمونه‌های مصنوعی نامناسب می‌توانند وفاداری داده و عملکرد مدل را کاهش دهند؛ بنابراین، تولید داده باید با ارزیابی کیفیت و در صورت لزوم، پالایش نمونه‌ها همراه باشد. در داده‌های شبکه‌ای، حفظ ساختار روابط نیز افزون بر شباهت توزیعی اهمیت دارد. سوم، اعتبارسنجی فراتر از عملکرد پیش‌بینی: شاخص‌هایی مانند Accuracy، Precision، Recall و F1 عملکرد مدل را در وظیفه مشخص ارزیابی می‌کنند، اما به‌تنهایی اعتبار داده افزوده را اثبات نمی‌کنند. اعتبارسنجی در علم‌سنجی باید متناسب با نوع داده، وفاداری آماری، ساختاری و معنایی را نیز بررسی کند؛ زیرا تغییر داده می‌تواند مرکزیت‌ها، ساختار اجتماعات و الگوهای ارتباطی را دگرگون سازد. بنابراین، بهبود دقت پیش‌بینی الزاماً به معنای افزایش اعتبار تحلیل علم‌سنجی نیست. چهارم، تشخیص ماهیت عدم‌توازن: عدم‌توازن کلاس‌ها زمانی می‌تواند مبنای داده‌افزایی باشد که عملکرد وظیفه‌ای مشخص، مانند طبقه‌بندی، را محدود کند. این وضعیت با نابرابری توزیع استنادها، همکاری‌ها یا اندازه اجتماعات علمی یکسان نیست؛ زیرا چنین الگوهایی ممکن است ویژگی واقعی نظام علم باشند. ازاین‌رو، ضرورت مداخله باید پیش از متعادل‌سازی یا افزایش مصنوعی داده احراز شود. پنجم، وابستگی انتقال روش‌های گراف‌محور به ساختار شبکه هدف: شواهد مکمل، کاربرد داده‌افزایی گرافی را در هم‌ترازی موجودیت‌ها، طبقه‌بندی گره و تشخیص ناهنجاری نشان می‌دهند؛ اما این نتایج، اعتبار مداخله در شبکه‌های استنادی یا هم‌نویسندگی را اثبات نمی‌کنند. با توجه به معنای مشخص یال‌ها در شبکه‌های علمی، انتقال این روش‌ها نیازمند آزمون مستقل و ارزیابی حساسیت شاخص‌ها و تفسیر روابط نسبت به تغییر ساختار گراف است. ششم، کنترل مخاطرات داده‌افزایی مبتنی بر مدل‌های زبانی بزرگ: مطالعات مکمل، روش‌هایی مانند تولید مبتنی بر پرامپت، بازنویسی، استفاده از اطلاعات بازیابی‌شده و تولید متن مصنوعی را گزارش کرده‌اند؛ اما مخاطراتی مانند اطلاعات نادرست، تفاوت توزیعی با داده واقعی، تقویت سوگیری و ابهام در منشأ داده همچنان مطرح‌اند. بنابراین، ثبت منشأ، مقایسه با داده مرجع، ارزیابی واقع‌نمایی و در موارد نیازمند قضاوت معنایی، بازبینی انسانی از الزامات انتقال این روش‌ها به داده‌های علمی است.

جدول ۶. ملاحظات روش‌شناختی و شرایط انتقال داده‌افزایی به مسائل علم‌سنجی

شرط انتقال به مسائل علم‌سنجی	شواهد پشتیبان نهایی	ملاحظه حاصل از شواهد

آزمون مستقل روش روی داده و وظیفه علم‌سنجی هدف	مطالعات مکمل متنی، گرافی و تولید داده مصنوعی	وابستگی روش به نوع داده و وظیفه تحلیلی
ارزیابی و در صورت لزوم پالایش داده مصنوعی پیش از ورود به تحلیل	M01, M02, M13	متغیر بودن کیفیت داده مصنوعی
ارزیابی هم‌زمان عملکرد و وفاداری آماری، ساختاری و معنایی متناسب با نوع داده	M02, M07, M09, M10, M13	ناکافی بودن معیارهای صرفاً پیش‌بینی
تشخیص منشأ و اثر عدم‌توازن بر وظیفه تحلیلی پیش از مداخله	M05, M10	عدم‌توازن لزوماً نشان‌دهنده نقص داده نیست
آزمون روی شبکه علمی هدف و ارزیابی حساسیت نسبت به تغییر ساختار	M07, M09, M10	وابستگی داده‌افزایی گرافی به ساختار شبکه
ثبت منشأ، مقایسه با داده مرجع، کنترل اعتبار محتوایی و بازبینی متناسب با حساسیت مسئله	M03, M04, M06	قابلیت تولیدی الالام همراه با خطر خطا و سوگیری

شواهد حوزه‌های مجاور در این پژوهش، مبنایی برای تعیین شرایط آزمون و ارزیابی انتقال روش‌ها به علم‌سنجی فراهم می‌کنند، نه اثبات اعتبار یا اثربخشی مستقیم آن‌ها. پیش از انتقال، تناسب نوع داده و وظیفه تحلیلی، کیفیت داده تولیدشده، حفظ ساختار و معنای روابط، منشأ عدم‌توازن و الزامات اعتبارسنجی باید بررسی شوند. بنابراین، موفقیت روش‌ها در پردازش زبان طبیعی، یادگیری ماشین یا یادگیری گراف، صرفاً توجیه‌کننده بررسی و آزمون مستقل آن‌ها در مسائل علم‌سنجی است و به‌تنهایی اعتبار یا اثربخشی این روش‌ها را در حوزه مذکور اثبات نمی‌کند.

پاسخ به پرسش ششم پژوهش. چه مخاطرات، ملاحظات و الزامات اعتبارسنجی برای استفاده از داده‌های افزوده، مصنوعی یا استنباطی در تحلیل‌های علم‌سنجی گزارش شده‌اند؟

سنتز شواهد نشان داد که اعتبارسنجی داده‌های افزوده، مصنوعی یا استنباطی در علم‌سنجی فراتر از عملکرد مدل یا شباهت آماری داده‌های تولیدشده است و به منشأ و وضعیت معرفتی داده، وفاداری به داده مرجع، حفظ ساختار و معنای روابط و حدود تعمیم نتایج وابسته است. بر این اساس، هفت محور اصلی و یک محدودیت تکمیلی شناسایی شد.

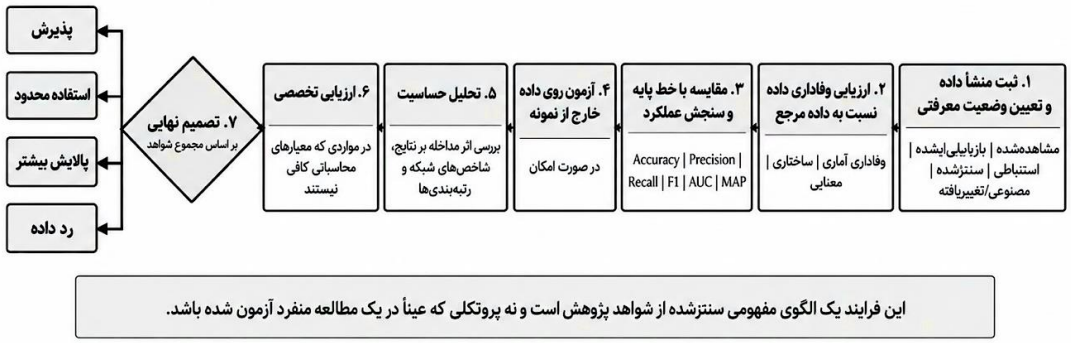
نخست، حفظ منشأ و وضعیت معرفتی داده: داده مشاهده‌شده، داده بازیابی‌شده و مستقلاً تأییدشده، خروجی استنباطی یا پیش‌بینی‌شده و داده مصنوعی یا تغییریافته، از نظر منشأ و پشتوانه شواهدی یکسان نیستند. ادغام نامتمایز آن‌ها، تشخیص سهم داده واقعی و مداخله محاسباتی را دشوار می‌کند؛ بنابراین، منشأ و وضعیت هر نمونه یا رابطه باید در سراسر تحلیل قابل ردیابی باشد.دوم، جلوگیری از تفسیر روابط پیش‌بینی‌شده به‌عنوان روابط واقعی: عملکرد مناسب الگوریتم‌های پیش‌بینی پیوند، توصیه استناد یا همکاری، وجود تاریخی روابط پیشنهادی را اثبات نمی‌کند. این روابط تا زمان تأیید با شواهد مستقل یا ثبت وقوع واقعی، ماهیت استنباطی یا پیش‌بینی‌شده دارند و نباید به‌عنوان روابط مشاهده‌شده یا بازیابی‌شده گزارش شوند. سوم، کنترل تغییرات ساختاری شبکه: افزودن، حذف یا تغییر روابط می‌تواند توزیع درجه، مرکزیت، ساختار اجتماعات و رتبه‌بندی گره‌ها را دگرگون کند. همچنین، برخی روابط حاصل نحوه تجمیع یا بازنمایی شبکه‌اند و به همان صورت در داده اولیه مشاهده نشده‌اند. ازاین‌رو، مداخلات شبکه‌ای مستلزم مقایسه ساختار پیش و پس از مداخله و تحلیل حساسیت نتایج هستند. چهارم، تفکیک عملکرد محاسباتی از اعتبار علم‌سنجی: شاخص‌هایی

مانند Accuracy، Precision، Recall، F1، AUCو MAP عملکرد مدل را در وظایف مشخص ارزیابی می‌کنند، اما به‌تنهایی اعتبار داده افزوده را اثبات نمی‌کنند. اعتبارسنجی باید متناسب با نوع داده، وفاداری آماری، ساختاری و معنایی و در مسائل پویا، اعتبار زمانی را نیز بررسی کند؛ زیرا بهبود عملکرد مدل الزاماً به معنای اعتبار بیشتر نتایج علم‌سنجی نیست. پنجم، ارزیابی وفاداری و کیفیت داده مصنوعی: نمونه‌های مصنوعی از کیفیت یکسانی برخوردار نیستند و نمونه‌های نامناسب می‌توانند وفاداری به داده مرجع و عملکرد مدل را کاهش دهند. در شبکه‌های علمی، شباهت آماری به‌تنهایی کافی نیست؛ زیرا حفظ توزیع متغیرها لزوماً به معنای حفظ توپولوژی و معنای روابط نیست. بنابراین، ارزیابی کیفیت و در صورت لزوم، پالایش نمونه‌های مصنوعی باید پیش از ورود آن‌ها به تحلیل اصلی انجام شود. ششم، تشخیص ماهیت عدم‌توازن و کم‌تراکمی: نابرابری توزیع استنادها، همکاری‌ها یا اندازه اجتماعات ممکن است بازتاب ساختار واقعی علم باشد. متعادل‌سازی مصنوعی چنین داده‌هایی، بدون احراز نقص یا محدودیت مؤثر بر وظیفه تحلیلی، می‌تواند پدیده مورد مطالعه را تحریف کند. ازاین‌رو، باید میان عدم‌توازن مسئله‌ساز برای مدل و ویژگی واقعی ساختار شبکه تمایز گذاشت. هفتم، کنترل مخاطرات داده‌های تولیدشده با مدل‌های زبانی بزرگ: تولید اطلاعات نادرست اما ظاهراً معتبر، تفاوت توزیعی با داده واقعی، تقویت سوگیری و ابهام در منشأ داده از مخاطرات اصلی این روش‌هاست. ورود چنین خروجی‌هایی به متون علمی، فرادادهٔ یا روابط شبکه‌ای می‌تواند نتایج تحلیل را مخدوش کند؛ بنابراین، ثبت منشأ، مقایسه با داده مرجع، ارزیابی اعتبار محتوایی و در مواردی که سنجه‌های محاسباتی کافی نیستند، بازبینی انسانی ضرورت دارد. در نهایت، محدودیت انتقال روش‌ها از حوزه‌های مجاور: موفقیت داده‌افزایی در متون، داده‌های جدولی یا گراف‌های عمومی، اعتبار آن را در علم‌سنجی تضمین نمی‌کند. هر روش انتقال‌یافته باید متناسب با ساختار داده، معنای روابط و وظیفه تحلیلی مقصد، به‌طور مستقل آزمون و اعتبارسنجی شود.

جدول ۷. ریسک‌ها و الزامات اعتبارسنجی داده‌های افزوده، مصنوعی و استنباطی در علم‌سنجی

الزام اعتبارسنجی	پیامد تحلیلی محتمل	ریسک یا محدودیت
ثبت منشأ و برچسب‌گذاری داده به‌صورت مشاهده‌شده، بازیابی‌شده/ تاییدشده، استنباطی/ پیش‌بینی‌شده و مصنوعی/ تغییریافته	از بین رفتن قابلیت ردیابی و ادغام نامتمایز داده واقعی و تولیدشده	اختلاط وضعیت‌های معرفتی داده
تایید با منبع مستقل، داده مرجع یا رخداد واقعی	ورود روابط احتمالی یا پیشنهادی به شبکه مشاهده‌شده	پذیرش رابطه استنباطی به‌عنوان رابطه واقعی
مقایسه پیش و پس از مداخله و تحلیل حساسیت ساختاری	تغییر توزیع درجه، مرکزیت، اجتماعات یا رتبه‌بندی‌ها	تغییر ساختار شبکه
ارزیابی هم‌زمان عملکرد و وفاداری آماری، ساختاری و معنایی	بهبود محاسباتی بدون اطمینان از اعتبار تحلیلی	اتکای صرف به عملکرد مدل
ارزیابی کیفیت، مقایسه با داده مرجع و در صورت نیاز پالایش نمونه‌ها	کاهش وفاداری داده یا افت عملکرد	کیفیت ناکافی داده مصنوعی
تعیین منشأ و نقش عدم‌توازن پیش از مداخله	تحریف ویژگی واقعی ساختار علم	اصلاح ناموجه عدم‌توازن طبیعی
ثبت منشأ، کنترل اعتبار محتوایی و بازبینی متناسب با حساسیت مسئله	ورود اطلاعات یا الگوهای غیرقابل اتکا به تحلیل	خطا یا سوگیری در داده تولیدشده با الالام
آزمون مستقل روش روی داده و وظیفه علم‌سنجی هدف	اعتبار بیرونی ناکافی برای داده و شبکه علمی	انتقال مستقیم روش از حوزه‌های مجاور

جدول ۷ نشان می‌دهد که اعتبارسنجی فراتر از بهبود عملکرد مدل است و باید منشأ، نوع و وضعیت معرفتی داده، وفاداری به داده مرجع و اثر مداخله بر ساختار و تفسیر نتایج را دربرگیرد. الگوی پیشنهادی با ثبت این مشخصات آغاز می‌شود و ارزیابی وفاداری آماری، ساختاری و معنایی، مقایسه با خط پایه، آزمون خارج از نمونه، تحلیل حساسیت و در صورت لزوم ارزیابی تخصصی را شامل می‌شود. برآیند این مراحل، مبنای پذیرش، استفاده محدود، پالایش یا رد داده است. این الگو حاصل سنتز شواهد پژوهش حاضر است و به‌صورت یک پروتکل یکپارچه در مطالعه‌ای منفرد به‌طور کامل آزمون نشده است.



شکل ۳. فرآیند چندمرحله‌ای اعتبارسنجی داده‌های افزوده، مصنوعی یا استنباطی در تحلیل‌های علم‌سنجی

شکل ۳ مسیر اعتبارسنجی داده‌های افزوده، مصنوعی یا استنباطی را از ثبت منشأ و وضعیت معرفتی تا ارزیابی وفاداری، مقایسه با خط پایه، آزمون خارج از نمونه و تحلیل حساسیت نشان می‌دهد. نتایج این مراحل، مبنای پذیرش، استفاده محدود، پالایش یا رد داده قرار می‌گیرند. این الگو حاصل سنتز شواهد پژوهش حاضر است و به‌صورت یک پروتکل یکپارچه در مطالعه‌ای منفرد به‌طور کامل آزمون نشده است.

پاسخ به پرسش هفتم پژوهش. بر اساس سنتز شواهد موجود، چه چارچوب مفهومی تصمیم‌محوری برای تشخیص منشأ مسئله داده‌ای، انتخاب نوع مداخله، تعیین وضعیت معرفتی، اعتبارسنجی و استفاده مشروط از داده‌افزایی در علم‌سنجی و شبکه‌های دانشی قابل استخراج است؟

سنتز شواهد هسته‌ای و مطالعات روش‌شناختی مکمل نشان داد که نمی‌توان رابطه‌ای عمومی و ازپیش‌اثبات‌شده میان داده‌افزایی و بهبود کیفیت داده یا دقت تحلیل در علم‌سنجی فرض کرد. بر این اساس، چارچوبی مفهومی، سنتزی و تصمیم‌محور در هشت گام صورت‌بندی شد که داده‌افزایی را یکی از گزینه‌های مواجهه با مسائل داده‌ای می‌داند. این چارچوب، مدلی تجربی یا اعتبارسنجی‌شده نیست، بلکه مسیر تصمیم‌گیری روش‌شناختی را بر پایه شواهد سازمان می‌دهد.

گام نخست، تشخیص مسئله داده‌ای: کمبود، کم‌تراکمی، عدم‌توازن یا ناهمگنی مشاهده‌شده لزوماً نقص داده نیست. نابرابری توزیع استنادها، همکاری‌ها یا اندازه اجتماعات ممکن است ویژگی واقعی نظام علم باشد؛ بنابراین، ضرورت مداخله باید با شواهد مستقل احراز شود. گام دوم، تعیین منشأ مسئله: ناهمسانی پوشش پایگاه‌ها، خطاهای فراداده، آثار

بازنمایی شبکه، کمبود نمونه آموزشی و عدم‌توازن کلاس‌ها منشأ و پیامد یکسانی ندارند. تشخیص اینکه محدودیت از کیفیت داده، پوشش منبع، شیوه بازنمایی یا وظیفه تحلیلی ناشی شده است، مبنای انتخاب مداخله قرار می‌گیرد. گام سوم، تعیین نوع داده و واحد تحلیل: متن علمی، سری زمانی کتاب‌سنجی، شبکه استنادی، شبکه هم‌نویسندگی، داده‌های ثبت اختراع و گراف دانش از نظر ساختار، معنا و الزامات اعتبارسنجی متفاوت‌اند؛ ازاین‌رو، انتقال روش میان آن‌ها نیازمند آزمون مستقل است. گام چهارم، انتخاب مسیر مداخله: پاک‌سازی و استانداردسازی، یکپارچه‌سازی، بازیابی و تأیید داده، داده‌افزایی و پیش‌بینی یا استنباط باید از یکدیگر متمایز باشند. اگر مسئله با اصلاح فراداده یا بازیابی از منبع مستقل برطرف شود، داده‌افزایی توجیه روش‌شناختی ندارد. همچنین، روابط پیش‌بینی‌شده بدون تأیید مستقل جایگزین روابط مشاهده‌شده نمی‌شوند. گام پنجم، انتخاب سازوکار متناسب: روش باید بر اساس منشأ مسئله، نوع داده و وظیفه تحلیلی انتخاب شود، نه صرفاً دسترس‌پذیری الگوریتم. تولید متن مصنوعی، داده‌افزایی سری زمانی، داده‌افزایی گرافی، پیش‌بینی پیوند و توصیه رابطه کارکردهای متفاوتی دارند؛ همچنین، پیش‌بینی یا توصیه تنها در صورت استفاده از خروجی به‌عنوان داده یا بازنمایی افزوده، مصداق داده‌افزایی است. گام ششم، اعتبارسنجی مستقل خروجی: عملکرد مدل به‌تنهایی برای پذیرش داده افزوده کافی نیست. اعتبارسنجی باید متناسب با نوع داده، وفاداری آماری، ساختاری و معنایی و در صورت لزوم اعتبار زمانی را بررسی کند. در شبکه‌های علمی، آثار مداخله بر توزیع درجه، مرکزیت، ساختار اجتماعات و رتبه‌بندی گره‌ها نیز باید ارزیابی شود. گام هفتم، تعیین و ثبت وضعیت معرفتی: خروجی باید در یکی از چهار وضعیت «مشاهده‌شده»، «بازیابی‌شده و مستقلاً تأییدشده»، «استنباطی یا پیش‌بینی‌شده» و «مصنوعی یا تغییریافته» مشخص شود. منشأ و وضعیت معرفتی داده باید از آغاز فرایند قابل ردیابی و در خروجی نهایی ثبت شود؛ زیرا عملکرد مناسب مدل، ماهیت استنباطی یا مصنوعی رابطه را تغییر نمی‌دهد. گام هشتم، تحلیل حساسیت و تصمیم نهایی: مناسب‌بودن داده برای آموزش مدل لزوماً به معنای اعتبار آن برای تحلیل اصلی شبکه نیست. روابط پیش‌بینی‌شده ممکن است در توصیه، پیش‌بینی یا سناریوسازی کاربرد داشته باشند، اما بدون تأیید مستقل نباید وارد شبکه تاریخی شوند. تصمیم نهایی، بر پایه نتایج اعتبارسنجی و تحلیل حساسیت، به پذیرش، استفاده محدود، پالایش بیشتر یا رد خروجی منجر می‌شود. سه اصل «تفکیک مفهومی عملیات»، «ردیابی منشأ شواهد» و «حفظ وضعیت معرفتی داده» بر تمامی گام‌ها حاکم‌اند. بدین‌ترتیب، ارزش چارچوب در فراهم‌کردن مسیری شفاف و قابل ردیابی از تشخیص مسئله تا استفاده مشروط از خروجی است؛ مسیری که داده‌افزایی را مداخله‌ای وابسته به مسئله، سطح شواهد و اعتبارسنجی می‌داند، نه راه‌حلی عمومی و خودکار.

زودآیند ویرایش نشده



شکل ۴. چارچوب مفهومی تصمیم‌محور برای استفاده و اعتبارسنجی داده‌افزایی در علم‌سنجی و شبکه‌های دانشی

شکل ٤ مسیر تصمیم‌گیری را از تشخیص مسئله و منشأ آن تا انتخاب مداخله، اعتبارسنجی و تعیین حدود استفاده از خروجی نشان می‌دهد. در این چارچوب، داده‌افزایی تنها در صورت تناسب با نوع مسئله، داده و هدف تحلیل انتخاب می‌شود و نحوه استفاده از خروجی به اعتبارسنجی متناسب با منشأ، نوع و وضعیت معرفتی آن وابسته است.

بحث و نتیجه‌گیری

پژوهش حاضر با هدف تبیین جایگاه داده‌افزایی در علم‌سنجی و تحلیل شبکه‌های دانشی انجام شد؛ حوزه‌ای که در آن مسئله داده لزوماً به معنای فقدان داده نیست و کاربرد الگوریتم‌های هوش مصنوعی نیز همواره مصداق داده‌افزایی محسوب نمی‌شود. سنتز ٤٣ مطالعه هسته‌ای نشان داد که شواهد موجود ناهمگون‌اند: ٦ مطالعه (١٤ درصد) کاربرد صریح داده‌افزایی را در داده‌ها یا فرایندهای علمی گزارش کرده‌اند؛ ٣٣ مطالعه (٧٦٫٧ درصد) به سازوکارهای محاسباتی مرتبط با داده‌ها و شبکه‌های علمی، مانند پیش‌بینی پیوند، توصیه استناد، تعبیه شبکه و گراف دانش پرداخته‌اند و ٤ مطالعه (٩٫٣ درصد) برای تبیین مسائل کیفیت داده و مرزبندی مفهومی به کار رفته‌اند. همچنین، ١٣ مطالعه روش‌شناختی مکمل برای شناخت ظرفیت روش‌ها، شرایط انتقال، اعتبارسنجی و مخاطرات استفاده شدند، اما در مخرج تحلیل هسته اصلی وارد نشدند. این ترکیب نشان می‌دهد که شواهد کاربرد مستقیم داده‌افزایی در حوزه مورد بررسی، محدودتر از ادبیات گسترده پیش‌بینی، توصیه و بازنمایی داده‌های علمی است.

مرورهای عمومی داده‌افزایی، از جمله شورتن و خوشگفتار (Shorten & Khoshgoftaar, 2019)، فنگ و همکاران (Feng et al., 2021) و مامونی و مامونی (Mumuni & Mumuni, 2022)، کاربرد افزایش کنترل‌شده نمونه‌های آموزشی را در برخی مسائل کمبود داده و عدم‌توازن نشان داده‌اند. یافته‌های پژوهش حاضر با این ظرفیت همسوست، اما تعمیم آن را در علم‌سنجی مشروط می‌داند. کمبود نمونه آموزشی در طبقه‌بندی، با کم‌تراکمی شبکه استنادی یا توزیع نامتوازن همکاری‌های علمی یکسان نیست؛ زیرا ویژگی‌های اخیر ممکن است ساختار واقعی نظام علم را بازتاب دهند.

بنابراین، بدون شواهد مستقل از نقص داده یا محدودیت مؤثر بر وظیفه تحلیلی، مداخله برای اصلاح این ویژگی‌ها توجیه نمی‌شود. بر اهمیت کیفیت اطلاعات ورودی، واکاریو و ورجینر (Angelini et al., 2022) اهمیت کیفیت اطلاعات ورودی را برای قابلیت اتکای تحلیل برجسته کرده‌اند و واکاریو و ورجینر (Vaccario & Verginer, 2022) بر آثار شیوه ساخت و بازنمایی شبکه بر شاخص‌ها و رتبه‌بندی‌ها، و رودریگز و همکاران (Rodrigues et al., 2025) بر تفاوت پوشش استنادی پایگاه‌های WoS، Scopus و OpenAlex تأکید کرده‌اند. همچنین، نوواکوفسکا (Nowakowska, 2025) نقش پیش‌پردازش و استانداردسازی داده‌های کتاب‌سنجی را بررسی کرده و کاپوتو و کارجینا (Caputo & Kargina, 2022) ظرفیت ادغام منابع را برای گسترش پوشش داده نشان داده‌اند. وجه اشتراک این مطالعات با یافته‌های حاضر، وابستگی اعتبار تحلیل به کیفیت منبع، فراداده و نحوه ساخت مجموعه داده است. بااین‌حال، پژوهش حاضر میان این مسائل و داده‌افزایی مرز روشنی برقرار می‌کند: بازیابی رکورد از منبع مستقل، یکپارچه‌سازی، پاک‌سازی، غنی‌سازی، جای‌گذاری مقادیر مفقود و رفع ابهام موجودیت‌ها، با تولید کنترل‌شده نمونه، ویژگی، بازنمایی یا ساختار افزوده یکسان نیستند. به همین ترتیب، پیش‌بینی صرف یک رابطه، بدون استفاده از خروجی به‌عنوان داده یا بازنمایی افزوده، داده‌افزایی محسوب نمی‌شود. این مرزبندی، انتخاب مداخله را به منشأ مسئله وابسته می‌سازد و مانع از تلقی داده‌افزایی به‌عنوان راه‌حل عمومی داده‌های ناقص می‌شود.

محدودبودن شواهد مستقیم نیز ضرورت این احتیاط را تأیید می‌کند. کائوثن و همکاران (Cauthen et al., 2022) با تولید سری‌های زمانی مصنوعی انتشارات، داده‌های آموزشی طبقه‌بندی بلوغ فناوری را گسترش دادند. ژانگ و همکاران (H. Zhang et al., 2025) در ReviewGuard از داده‌افزایی مبتنی بر مدل زبانی بزرگ برای تولید نمونه‌های مرتبط با داوری همتا و شناسایی داوری‌های ناکافی استفاده کردند. مطالعات دیگر، داده‌افزایی مولد را در طبقه‌بندی داده‌های ثبت اختراع و پیش‌بینی فناوری‌های نوظهور (Zhou et al., 2020; Xu et al., 2022)و داده‌افزایی گرافی را در توصیه مقاله و تحلیل زیرگراف‌های دانش ثبت اختراع به کار بردند (Y. Liu et al., 2025; C. Zhang et al., 2025). وجه مشترک این شش مطالعه، وجود داده هدف، وظیفه تحلیلی و مداخله مشخص است؛ اما ناهمگونی آن‌ها از نظر نوع داده، سازوکار و اعتبارسنجی، استنتاج اثربخشی عمومی داده‌افزایی در علم‌سنجی را ناموجه می‌سازد. در همین زمینه، لیو و همکاران (H. Liu et al., 2025) داده‌افزایی متنی را برای شناسایی ساختار کارکردی پاراگراف‌های متون علمی به کار برده‌اند. این مطالعه، به دلیل شناسایی خارج از مجموعه ثابت مرور نظام‌مند، صرفاً شاهدی زمینه‌ای برای توجه به تحولات حوزه است و در شمارش‌ها و استنتاج‌های نظام‌مند پژوهش حاضر وارد نشده است. این تفکیک ضمن حفظ ثبات پیکره بررسی‌شده، از افزودن انتخابی مطالعات جدید به نتایج مرور جلوگیری می‌کند.

یافته دیگر به تفاوت میان سازوکارهای محاسباتی شبکه و داده‌افزایی مربوط است. ادبیات پیش‌بینی پیوند (Liben-Nowell & Kleinberg, 2007) و یادگیری بازنمایی گراف (Perozzi et al., 2014; Hamilton et al., 2017; Kipf & Welling, 2017; Veličković et al., 2018) ظرفیت روش‌های محاسباتی را برای برآورد روابط، استخراج شباهت‌ها و بازنمایی ساختار شبکه نشان داده است. در مجموعه هسته‌ای پژوهش حاضر نیز ۳۳ مطالعه به سازوکارهای

مرتبط با توصیه مقاله و استناد، پیش‌بینی همکاری، تعبیه شبکه و گراف‌های دانش اختصاص داشتند؛ بااین‌حال، این کاربردها لزوماً داده‌افزایی مستقیم نبودند. مهم‌تر آنکه، توان پیش‌بینی یک رابطه، وجود تاریخی آن را اثبات نمی‌کند. استناد توصیه‌شده، پیوند پیش‌بینی‌شده یا همکاری پیشنهادی تا زمان تأیید مستقل، خروجی استنباطی باقی می‌ماند. این نتیجه با یافته واکاریو و ورجینر (Vaccario & Verginer, 2022) درباره آثار فرافکنی شبکه نیز سازگار است؛ زیرا روابط مشتق‌شده ممکن است به همان صورت در داده اولیه مشاهده نشده باشند. بر این اساس، پژوهش حاضر چهار وضعیت معرفتی را تفکیک می‌کند: «مشاهده‌شده»، «بازیابی‌شده و مستقلاً تأییدشده»، «استنباطی یا پیش‌بینی‌شده» و «مصنوعی یا تغییریافته». اهمیت این تمایز در آن است که دقت الگوریتم یا حضور یک یال در گراف یکپارچه، به‌تنهایی منشأ و وضعیت معرفتی رابطه را تغییر نمی‌دهد. در نتیجه، خروجی‌های پیش‌بینی‌شده یا مصنوعی باید از روابط تاریخی مشاهده‌شده متمایز باقی بمانند و تنها متناسب با هدف و اعتبارسنجی خود استفاده شوند.

مقایسه با شواهد حوزه‌های مجاور نیز ضرورت انتقال مشروط روش‌ها را نشان می‌دهد. زی و همکاران (Xie et al., 2023)، جوآن و همکاران (Juan et al., 2024) و لیو و همکاران (Liu et al., 2022) کاربرد داده‌افزایی گرافی را در وظایفی مانند هم‌ترازی موجودیت‌ها، طبقه‌بندی گره و تشخیص ناهنجاری بررسی کرده‌اند. همچنین، فنگ و همکاران (Feng et al., 2021) و چای و همکاران (Chai et al., 2026) روش‌های متنی، از بازنویسی و نمونه‌گیری تا تولید مبتنی بر مدل‌های زبانی، را گزارش کرده‌اند. این مطالعات ظرفیت روش‌ها را نشان می‌دهند، اما تفاوت معنای روابط و اهداف تحلیلی، انتقال بی‌واسطه آن‌ها به شبکه‌های استنادی، هم‌نویسندگی یا متون علمی را محدود می‌کند. برخلاف بسیاری از وظایف عمومی یادگیری ماشین، در علم‌سنجی خود ساختار و معنای داده نیز موضوع تحلیل است؛ بنابراین، افزودن یک یال ممکن است مرکزیت، ساختار اجتماعات، رتبه‌بندی گره‌ها و تفسیر روابط علمی را تغییر دهد. این ملاحظه با ادبیات تولید و اعتبارسنجی داده مصنوعی نیز همسوست. فیگیرا و واز (Figueira & Vaz, 2022) بر ضرورت ارزیابی کیفیت داده مصنوعی تأکید کرده‌اند و ساچدوا و همکاران (Sachdeva et al., 2025) نشان داده‌اند که نمونه‌های نامناسب می‌توانند وفاداری داده و عملکرد مدل را کاهش دهند. پژوهش حاضر این منطق را در علم‌سنجی گسترش می‌دهد: معیارهایی مانند Accuracy، Precision، Recall، F1، AUC و MAP برای ارزیابی عملکرد مدل مفیدند، اما اعتبار داده افزوده را به‌تنهایی اثبات نمی‌کنند. اعتبارسنجی باید متناسب با مسئله، وفاداری آماری، ساختاری و معنایی و در صورت لزوم اعتبار زمانی را نیز دربرگیرد. بر همین مبنا، الگوی پیشنهادی شامل ثبت منشأ و وضعیت معرفتی، مقایسه با داده مرجع و خط پایه، آزمون خارج از نمونه در صورت امکان، تحلیل حساسیت و ارزیابی تخصصی در موارد لازم است. نتیجه این ارزیابی می‌تواند پذیرش، استفاده محدود، پالایش یا رد داده باشد؛ زیرا داده‌ای که برای آموزش مدل مناسب است، الزاماً برای ورود به شبکه تاریخی یا محاسبه شاخص‌های علم‌سنجی اعتبار ندارد.

از منظر تحول علم‌سنجی در مواجهه با هوش مصنوعی، یافته‌ها با دیدگاه بورنر (Börner, 2010) درباره ظرفیت

روش‌های محاسباتی برای بازنمایی ساختار علم و بحث نوروزی چاکلی (Noroozi Chakoli, 2026) درباره مواجهه مسئولانه با هوش مصنوعی در علم‌سنجی هم‌راستاست. بااین‌حال، پژوهش حاضر بر این تمایز تأکید می‌کند که مدل‌های زبانی، شبکه‌های عصبی گراف و روش‌های مولد، ابزارهای محاسباتی‌اند و اعتبار خروجی آن‌ها همچنان به منطق اندازه‌گیری، معنای روابط علمی و هدف تحلیل وابسته است.

برآیند مقایسه با پیشینه، سه محور همسویی را مشخص می‌کند: اهمیت کیفیت و پوشش داده، تناسب روش با نوع داده و وظیفه تحلیلی، و ضرورت اعتبارسنجی داده مصنوعی. نوآوری مفهومی پژوهش حاضر، پیوند این محورها در یک منطق تصمیم‌گیری واحد است. مرورهای عمومی داده‌افزایی عمدتاً روش‌های افزایش داده را بررسی می‌کنند؛ مطالعات کیفیت داده بر اعتبار و کامل‌بودن داده تمرکز دارند و مطالعات شبکه‌ای به پیش‌بینی روابط و بازنمایی ساختار می‌پردازند. چارچوب حاضر با اتصال این رویکردها، پرسشی مقدماتی مطرح می‌کند: آیا در مسئله مورد بررسی، تولید یا افزودن داده اساساً موجه است؟ چارچوب پیشنهادی برای پاسخ به این پرسش، مسیری هشت‌گامی از تشخیص مسئله و منشأ آن، تعیین نوع داده و واحد تحلیل، انتخاب مسیر مداخله و سازوکار متناسب، اعتبارسنجی مستقل، ثبت نهایی وضعیت معرفتی و تحلیل حساسیت تا تصمیم درباره نحوه استفاده از خروجی فراهم می‌کند. سه اصل «مرزبندی مفهومی عملیات»، «ردیابی منشأ شواهد» و «حفظ وضعیت معرفتی داده» نیز بر تمام مراحل حاکم‌اند. این چارچوب حاصل سنتز شواهد است و نباید مدل پیش‌بینانه یا الگویی با اعتبارسنجی تجربی کامل تلقی شود. ارزش آن در شفاف‌سازی ترتیب تصمیم‌ها و فراهم‌کردن مبنایی قابل ردیابی برای انتخاب، ارزیابی و استفاده مشروط از مداخلات داده‌ای است. در نتیجه، داده‌افزایی در علم‌سنجی نه راه‌حلی عمومی برای کمبود داده است و نه روشی با اثر تضمین‌شده بر کیفیت و دقت تحلیل. شواهد مستقیم، امکان کاربرد آن را در وظایفی مشخص نشان می‌دهند، اما برای تعمیم اثربخشی به همه داده‌ها و شبکه‌های علمی کافی نیستند. گاهی اصلاح فراداده، بازیابی رکورد معتبر، یکپارچه‌سازی منابع یا حتی عدم مداخله، انتخاب روش‌شناختی مناسب‌تری است. زمانی نیز که داده‌افزایی توجیه می‌شود، خروجی باید مستقل از عملکرد صرف مدل و متناسب با منشأ، ماهیت و هدف استفاده اعتبارسنجی شود. بنابراین، دستاورد اصلی پژوهش حاضر، ارائه چارچوبی تصمیم‌محور برای انتخاب مداخله متناسب با مسئله، حفظ وضعیت معرفتی داده و تعیین حدود استفاده معتبر از خروجی در علم‌سنجی و شبکه‌های دانشی است.

پیشنهادهای اجرایی پژوهش

بر پایه یافته‌ها، انتخاب مداخله در علم‌سنجی باید با تشخیص منشأ مسئله آغاز شود؛ زیرا بسیاری از مشکلات با پاک‌سازی، استانداردسازی، بازیابی یا یکپارچه‌سازی داده برطرف می‌شوند. همچنین، کم‌تراکمی یا عدم‌توازن شبکه لزوماً نقص داده نیست. ازاین‌رو، داده‌افزایی تنها در صورت احراز ضرورت تولید یا تبدیل کنترل‌شده داده برای وظیفه‌ای مشخص توصیه می‌شود.

– در سامانه‌های علم‌سنجی، منشأ و وضعیت معرفتی داده‌ها و روابط ثبت و قابل ردیابی باشد. داده‌های مشاهده‌شده، بازیابی‌شده و مستقلاً تأییدشده، استنباطی یا پیش‌بینی‌شده و مصنوعی یا تغییریافته باید از یکدیگر تفکیک شوند و روابط تولیدشده یا پیشنهادی بدون تأیید مستقل، معادل روابط مشاهده‌شده تلقی نشوند.

– اعتبارسنجی داده‌های افزوده، افزون بر عملکرد مدل، وفاداری آماری، ساختاری، معنایی و در صورت لزوم زمانی را دربرگیرد. آزمون خارج از نمونه، تحلیل حساسیت و ارزیابی آثار مداخله بر شاخص‌های شبکه نیز متناسب با مسئله انجام شود تا مبنای پذیرش، استفاده محدود، پالایش یا رد خروجی فراهم آید.

در مراکز علم‌سنجی و ارزیابی پژوهش، داده‌های مصنوعی و پیش‌بینی‌شده از داده‌های مبنای گزارش‌های رسمی تفکیک شوند. همچنین، منشأ داده، روش تولید یا تبدیل، نتایج اعتبارسنجی و حدود استفاده از خروجی به‌صورت شفاف مستند شود تا امکان بازبینی و ارزیابی مستقل فراهم باشد.

پیشنهاد برای پژوهش‌های آتی

– آزمون تجربی روش‌های داده‌افزایی: پژوهش‌های آینده می‌توانند با ایجاد مجموعه‌داده‌های مرجع از داده‌های واقعی علم‌سنجی و شبیه‌سازی کنترل‌شده حذف یا کمبود داده، عملکرد داده‌افزایی را با بازیابی از منابع مستقل، برآورد و جای‌گذاری مقادیر مفقود و سایر مداخلات مقایسه کنند.

– ارزیابی داده‌افزایی گرافی: آزمون مستقل این روش‌ها در شبکه‌های استنادی، هم‌نویسندگی و موضوعی، با توجه به تفاوت معنای روابط، پیشنهاد می‌شود. ارزیابی باید افزون بر عملکرد مدل، آثار مداخله بر ساختار شبکه، مرکزیت، رتبه‌بندی و اجتماعات را بررسی کند. مطالعه طولی روابط پیش‌بینی‌شده نیز می‌تواند میزان تحقق آن‌ها را در دوره‌های بعد مشخص سازد.

– بررسی داده‌افزایی مبتنی بر مدل‌های زبانی بزرگ: کاربرد متن مصنوعی در طبقه‌بندی کارکرد استناد، تحلیل زمینه استناد و شناسایی موضوع، از مسیرهای پژوهشی آینده است. مقایسه داده‌های انسانی و مصنوعی از نظر واقع‌نمایی، سوگیری، تنوع، شباهت توزیعی و عملکرد خارج از نمونه، همراه با کنترل تولید استنادها یا روابط غیرواقعی، ضرورت دارد.

– توسعه سنجه‌های وفاداری علم‌سنجی و اعتبارسنجی چارچوب: پیشنهاد می‌شود مفهوم «وفاداری علم‌سنجی» با تأکید بر حفظ ساختار استنادی، الگوهای همکاری، توزیع موضوعی و شاخص‌های علم‌سنجی، فراتر از شباهت آماری، عملیاتی و آزمون شود. همچنین، ارزیابی تجربی چارچوب تصمیم‌محور و مقایسه داده‌های WoS، Scopus، Dimensions و OpenAlex می‌تواند به تفکیک محدودیت‌های واقعی داده از تفاوت‌های پوشش و فراداده پایگاه‌ها کمک کند.

بیانیه‌ها و شفافیت پژوهش

زودآیند ویرایش نشده

نویسنده این اثر اعلام می‌دارد که هیچ‌گونه تضاد منافع مالی یا غیرمالی در فرایند پژوهش و انتشار این دست‌نوشته وجود ندارد. همچنین، نویسندگان متعهد می‌شوند که تمامی اصول اخلاق نشر- شامل اصالت اثر و پرهیز از سرقت ادبی، اخذ رضایت آگاهانه، پرهیز از ساخت و دستکاری داده‌ها (جعل و تحریف)، خودداری از ارسال موازی و انتشار چندباره، و پایبندی به دستورالعمل مجله پیرامون کاربرد فناوری‌های هوش مصنوعی- را به‌دقت رعایت کرده‌اند.

■ بیانیه‌ی تعارض منافع و اخلاق پژوهش

این مطالعه حاصل پژوهشی مستقل است و هیچ‌گونه حمایت مالی یا گرنت پژوهشی از سازمان‌های دولتی، تجاری (خصوصی) یا بخش غیرانتفاعی دریافت نکرده است.

■ گرنت‌ها و حمایت مالی

این مطالعه حاصل پژوهشی مستقل است و هیچ‌گونه حمایت مالی یا گرنت پژوهشی از سازمان‌های دولتی، تجاری (خصوصی) یا بخش غیرانتفاعی دریافت نکرده است.

■ سهم/ مشارکت نویسندگان

نویسنده مقاله، به‌عنوان تنها نویسنده مقاله، مسئولیت مفهوم‌سازی، طراحی روش‌شناسی پژوهش، گردآوری و مدیریت داده‌ها، تحلیل رسمی، اعتبارسنجی یافته‌ها، تصویرسازی، نگارش پیش‌نویس اولیه، بازبینی و ویرایش علمی متن را بر عهده داشته و نسخه نهایی مقاله را بررسی و تأیید کرده است.

■ دسترسی‌پذیری داده‌ها

داده‌های پشتیبان یافته‌های این پژوهش هم‌زمان با ارسال نسخه اولیه دست‌نوشته در اختیار مجله قرار گرفته‌اند و اکنون در مخزن عمومی Mendeley Data با عنوان «Evidence Dataset for a Systematic Review of Data Augmentation in Scientometrics and Scholarly Knowledge Networks» منتشر شده‌اند. این مجموعه‌داده در نسخه نخست (V1) با شناسه دیجیتال دائمی DOI: 10.17632/cc8p8nhfnv.1 به‌صورت آزاد در دسترس پژوهشگران قرار دارد و امکان بررسی، ارزیابی مستقل و ردیابی شواهد و داده‌های تحلیل‌شده پژوهش را فراهم می‌کند (GhaviDel, 2026).

■ تقدیر و تشکر

نویسنده مراتب قدردانی صمیمانه‌ی خود را از سردبیر و داوران محترم و ناشناس «پژوهشنامه علم‌سنجی» به پاس ارائه‌ی نظرات، رهنمودها و بازخوردهای ارزنده‌ای که موجب ارتقای کیفیت علمی و ساختاری این مقاله شد، ابراز می‌دارد.

فهرست منابع

اسماعیلی، ر.، کیمیایی‌فر، ر.، حاجیان، ع.، سلیمانی، خ.، و هدهدی، م. (۱۴۰۳). بررسی تاثیر داده‌افزایی در هوشمند سازی مطالعات مخاطرات محیطی- مطالعه موردی: محاسبه آنی بزرگا در سیستمهای هشدار سریع زلزله. *جغرافیا و مطالعات محیطی*، ۱۳(۵۲)، ۷-۱۷. https://doi.org/10.71740/ges.2024.1130183

میر، م.، و نوفرستی، س. (۱۴۰۳). به‌کارگیری روشهای داده‌افزایی برای تحلیل احساسات کاربران درباره بازگشایی مدارس در دوران همه‌گیری کووید-۱۹. *پردازش علائم و داده‌ها*، ۲۱(۲)، ۱۴-۳. https://doi.org/10.61186/jsdp.21.2.3

نوروزی چاکلی، ع. (۱۴۰۵). سخن سردبیر: علم‌سنجی در آستانه دگرگونی: از سنجش و ارزیابی انسان‌محور تا همزیستی مسئولانه با هوش مصنوعی. *پژوهش نامه علم سنجی*، ۱۲(۱)، ۱-۶. https://doi.org/10.22070/rsci.2026.5033

Angelini, M., Daraio, C., & Urban, L. (2022). A visual analytics approach for the assessment

of information quality of performance models—A software review. *Scientometrics*, *127*, 6827–6853. https://doi.org/10.1007/s11192-022-04399-2

Börner, K. (2010). *Atlas of science: Visualizing what we know*. MIT Press.

Bornmann, L., & Leydesdorff, L. (2014). Scientometrics in a changing research landscape: Bibliometrics has become an integral part of research quality evaluation and has been changing the practice of research. *EMBO Reports*, *15*(12), 1228–1232. https://doi.org/10.15252/embr.201439608.

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a Data Source for Research on Scientific and Scholarly Activity. *Quantitative Science Studies*, *1*(1), 363-376. https://doi.org/10.1162/qss_a_00018

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Caputo, A., & Kargina, M. (2022). A user-friendly method to merge Scopus and Web of Science data during bibliometric analysis. *Journal of Marketing Analytics*, *10*(1), 82–88. https://doi.org/10.1057/s41270-021-00142-7

Cauthen, K., Rai, P., Hale, N., Freeman, L. J., & Ray, J. (2022). Detecting technological maturity from bibliometric patterns. *Expert Systems with Applications*, *201*, 117177. https://doi.org/10.1016/j.eswa.2022.117177

Chai, Y., Xie, H., & Qin, J. S. (2026). Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities. *Artificial Intelligence Review*, *59*(35). https://doi.org/10.1007/s10462-025-11405-5

Dodampegama, S., Hou, L., Asadi, E., Zhang, G. M., & Setunge, S. (2024). Revolutionizing construction and demolition waste sorting: Insights from artificial intelligence and robotic applications. *Resources, Conservation and Recycling*, *202*, 107375. https://doi.org/10.1016/j.resconrec.2023.107375.

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, *62*(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Esmaely, R., Kimiaeifar, R., Hajian, A., Soleimani, K., Hodhodi, M. (2024). Investigating the Effect of Data Augmentation on the Intelligentization of Environmental Hazard Studies - Case Study: Real-Time Calculation of Earthquake Magnitudes in Early Warning Systems. *Journal of Geography and Environmental Studies*, *13* (52), 1- 17. https://doi.org/10.71740/ges.2024.1130183 **[In Persian]**

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. *In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 968–988). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.84

Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, *10*(15), 2733. https://doi.org/10.3390/math10152733

Hamilton, W. L., Ying, R., Leskovec, J. (2017). Representation learning on graphs. Morgan & Claypool. Representation learning on graphs: Methods and applications. *IEEE Data*

Engineering Bulletin, 40(3), 52–74. https://dblp.org/rec/journals/debu/HamiltonYL17.html

Harrison, R., Jones, B., Gardner, P., & Lawton, R. (2021). Quality assessment with diverse studies (QuADS): An appraisal tool for methodological and reporting quality in systematic reviews of mixed- or multi-method studies. *BMC Health Services Research*, *21*, 144. https://doi.org/10.1186/s12913-021-06122-y

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Juan, X., Liang, X., Xue, H., & Wang, X. (2024). Multi-strategy adaptive data augmentation for graph neural networks. *Expert Systems with Applications*, *258*, 125076. https://doi.org/10.1016/j.eswa.2024.125076

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR),* 1-14. https://doi.org/10.48550/arXiv.1609.02907

Lefebvre, C., Glanville, J., Briscoe, S., Featherstone, R., Littlewood, A., Metzendorf, M.-I., Noel-Storr, A., Paynter, R., Rader, T., Thomas, J., & Wieland, L. S. (2025). Searching for and selecting studies. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* (Version 6.5.1, Chapter 4). Cochrane.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019–1031. https://doi.org/10.1002/asi.20591

Liu, F., Ma, X., Wu, J., Yang, J., Xue, S., Beheshti, A., Zhou, C., Peng, H., Sheng, Q. Z., & Aggarwal, C. C. (2022). DAGAD: Data augmentation for graph anomaly detection. In X. Zhu, S. Ranka, M. T. Thai, T. Washio, & X. Wu (Eds.), *2022 IEEE International Conference on Data Mining (ICDM)* (pp. 259–268). IEEE. https://doi.org/10.1109/ICDM54844.2022.00036

Liu, H., Chen, Z., Ding, Q., Zhang, J., Liu, J., Hu, J., Lu, W. (2025). Research on paragraph-level functional structure recognition in scientific literature: A data augmentation method based on LLMs and lexical function. *Scientometrics, 130*, 5473–5502. https://doi.org/10.1007/s11192-025-05355-6

Liu, Y., Zhou, Y., He, C., Liu, Y., & Dong, F. (2025). Uncovering emerging technologies in intelligent manufacturing via graph classification of community characteristics. *Journal of Engineering Design, 36*(7–9), 1502–1526. https://doi.org/10.1080/09544828.2024.2411486

Mir, M., Noferesti, S. (2024). Using Data Augmentation Techniques for Sentiment Analysis of Users' Opinions on Reopening of Schools During the Covid-19 Epidemic. *Signal and Data Processing*, *21*(2), 3-14. https://doi.org/10.61186/jsdp.21.2.3 **[In Persian]**

Mumuni, A., & Mumuni, M. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, *16*, 100258. https://doi.org/10.1016/j.array.2022.100258

Namboodiri, W., & Goyal, R. (2023). Benchmarking the Darrieus wind turbine configurations through review and data envelopment analysis. *Clean Technologies and Environmental Policy*, *25*(7), 2123–2155. https://doi.org/10.1007/s10098-023-02554-8

Noroozi Chakoli, A. (2026). Scientometrics at a turning point: From human-centered assessment and evaluation to responsible coexistence with artificial intelligence. *Scientometrics Research Journal*, *12*(1), 1–6. https://doi.org/10.22070/rsci.2026.5033 **[In Persian]**

Nowakowska, M. (2025). A comprehensive approach to preprocessing data for bibliometric analysis. *Scientometrics 130*, 5191–5225 .https://doi.org/10.1007/s11192-025-05415-x

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. *Proceedings of KDD*, 701–710. https://doi.org/10.1145/2623330.2623732

Rethlefsen, M.L., Kirtley, S., Waffenschmid, S., Ayala, A.P., Moher, D., Page, M.J., Koffel, J.B.; PRISMA-S Group (2021). PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Systematic Reviews, 10*(1), Article 39. https://doi.org/10.1186/s13643-020-01542-z

Roayaei, M. (2023). Learning an efficient text augmentation strategy: A case study in sentiment analysis. *International Journal of Web Research*, *6*(2), 67–75. https://doi.org/10.22133/ijwr.2024.441414.1202

Rodrigues, D., Lopes, A., & Batista, F. (2025). Detecting incoherent citation data among three bibliometric platforms: OpenAlex, Scopus and Web of Science. *Journal of Information Science*. Advance online publication. https://doi.org/10.1177/01655515251330579

Sachdeva, P., Malhotra, A., & Gupta, K. (2025). Quality Data Extractor (QDE): Elevating synthetic data augmentation through post-generation filtration. *IEEE Access*, *13*, 150901–150932. https://doi.org/10.1109/ACCESS.2025.3603435

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(60), 1–48. https://doi.org/10.1186/s40537-019-0197-0

Vaccario, G., Verginer, L. (2022). When standard network measures fail to rank journals: A theoretical and empirical analysis. *Quantitative Science Studies*; *3* (4), 1040–1053. https://doi.org/10.1162/qss_a_00225

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lió, P., & Bengio, Y. (2018). Graph attention networks. *International Conference on Learning Representations (ICLR)*, 1-12. https://doi.org/10.48550/arXiv.1710.10903

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007

Xie, F., Zeng, X., Zhou, B., Tan, Y. (2023). Improving Knowledge Graph Entity Alignment with Graph Augmentation. In: Kashima, H., Ide, T., Peng, WC. (eds) *Advances in Knowledge Discovery and Data Mining*. PAKDD 2023. Lecture Notes in Computer Science, 13936. Springer, Cham. https://doi.org/10.1007/978-3-031-33377-4_1

Xu, G., Dong, F., & Feng, J. (2022). Mapping the technological landscape of emerging

industry value chain through a patent lens: An integrated framework with deep learning. IEEE Transactions on Engineering Management, *69*(6), 3367–3378. https://doi.org/10.1109/TEM.2020.3041933

Zhang, C., Zhang, Z., Huang, J., Liu, Y., Jin, D., Xiao, X., & Shen, Z. (2025a). MKCRec: Meta-relation guided knowledge coupling for paper recommendation. *ACM Transactions on Information Systems*, *43*(3), Article 77, 1–29. https://doi.org/10.1145/3715101

Zhang, H., Li, R., Shrestha, S., Mamidala, S. H., Putta, R., Aggarwal, A. K., Xiao, T., Ding, J., & Chen, H. (2025b). ReviewGuard: Enhancing deficient peer review detection via LLM-driven data augmentation. In *2025 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE. 207-216. https://doi.org/10.1109/JCDL67857.2025.00032

Zhou, Y., Dong, F., Liu, Y., Li, Z., Du, J., & Zhang, L. (2020). Forecasting emerging technologies using data augmentation and deep learning. *Scientometrics*, *123*, 1–29. https://doi.org/10.1007/s11192-020-03351-6

– Research Declarations and Transparency

▪ Declaration of Competing Interests and Research Ethics

The author declares that no financial or non-financial conflicts of interest exist in relation to the conduct of this research or the publication of this manuscript. Furthermore, the author confirms full adherence to the principles of publication ethics, including the originality of the work and avoidance of plagiarism, obtaining informed consent, prevention of data fabrication and falsification, avoidance of simultaneous submission and duplicate publication, and compliance with the journal's guidelines concerning the use of artificial intelligence technologies.

▪ Funding and Grants

This study was conducted as independent research and received no financial support or research grants from governmental, commercial (private-sector), or non-profit organizations.

▪ Author Contributions (CRediT Statement)

The sole author of this manuscript was responsible for conceptualization, research methodology, data collection and curation, formal analysis, validation of findings, visualization, writing the original draft, and critically reviewing and editing the manuscript. The author has reviewed and approved the final version of the manuscript.

▪ Data Availability Statement

The data supporting the findings of this study were provided to the journal at the time of the initial manuscript submission and are now publicly available in the Mendeley Data repository under the title "Evidence Dataset for a Systematic Review of Data Augmentation in Scientometrics and Scholarly Knowledge Networks."

The dataset is openly accessible as Version 1 (V1) under the permanent Digital Object Identifier (DOI): 10.17632/cc8p8nhfnv.1. It enables researchers to examine, independently evaluate, and trace the supporting evidence and analytical data underlying the findings of this study (GhaviDel, 2026).

▪ Acknowledgments

The author wishes to express her sincere gratitude to the Editor-in-Chief and the esteemed anonymous reviewers of the Scientometrics Research Journal for their valuable comments, guidance, and constructive feedback, which contributed to improving the scientific quality and structure of this manuscript.