Scientometric Analysis of the Relationship Between Artificial Intelligence and Data Engineering: Trends, Collaboration, and Evolution*

Análisis cienciométrico de la relación entre inteligencia artificial e ingeniería de datos: tendencias, colaboración y evolución.

Angelo José Britto-Berrocal **
Delcides Miguel Cordoba-Rizo ***

Recibido: abril 23 de 2024 - Evaluado: septiembre 28 de 2024 - Aceptado: diciembre 21 de 2024

Para citar este artículo / To cite this Article
A. J. Britto-Berrocal, D. M.Cordoba-Rizo, "Scientometric Analysis of the Relationship Between Artificial Intelligence and Data Engineering: Trends, Collaboration, and Evolution", Revista de Ingenierías Interfaces, vol. 7, no.2, pp.1-17, 2024.

Abstract

Data engineering has become a fundamental step for artificial intelligence (AI) because the quality of training depends on the quality of the data. This thematic connection is relatively new, and the literature on the topic is quite scattered. Therefore, the purpose of this review is to identify the main contributions in this area by applying the Tree of Science algorithm. This algorithm processes query results from Scopus and Web of Science to classify articles into root, trunk, and branches. The main result of the study was the identification of three emerging areas. The first focuses on the use of AI to analyze and enhance complex scientific data, emphasizing solving specific optimization challenges. The second addresses the practical implementation of AI, tackling issues such as data cleaning and improving operational efficiency. The third highlights the development of AI from its creation to its maintenance once implemented, leveraging data engineering as a key tool to enhance AI training and performance. These findings are noteworthy because they shed light on the current use of AI applications to optimize processes in various sectors. The ability to process large volumes of data quickly improves efficiency and accelerates decision-making in sectors such as healthcare, industry, and cybersecurity. This enables personalized diagnostics, process optimization, and immediate responses to threats. However, it is important to emphasize that these benefits rely on data

*Artículo inédito: "Scientometric Analysis of the Relationship Between Artificial Intelligence and Data Engineering: Trends, Collaboration, and Evolution".

**Estudiante de Ingeniería Mecatrónica, semillero SciMetrix, abritto@unal.edu.co, ORCID: 0009-0004-4863-6283, Universidad Nacional de Colombia Sede de La Paz, Cesar, Colombia.

**** Estudiante de Ingeniería Mecatrónica, semillero SciMetrix, decordobar@unal.edu.co, ORCID: 0009-0008-4156-4212, Universidad Nacional de Colombia Sede de La Paz, Cesar, Colombia.

quality and the proper implementation of analysis systems, which ensure effective processing and reliable results.

Keywords: Artificial intelligence, Data Engineering, Data Preprocessing, Deep learning, Machine learning, Scientometrics,

Resumen

La ingeniería de datos se ha convertido en un paso fundamental para la inteligencia artificial (IA) porque la calidad del entrenamiento depende de la calidad de los datos. Esta conexión temática es relativamente nueva y la literatura sobre el tema está bastante dispersa. Por lo tanto, el propósito de esta revisión es identificar los principales aportes en esta área mediante la aplicación del algoritmo del Árbol de la Ciencia. Este algoritmo procesa los resultados de consultas de Scopus y Web of Science para clasificar los artículos en raíz, tronco y ramas. El principal resultado del estudio fue la identificación de tres áreas emergentes. El primero se centra en el uso de la IA para analizar y mejorar datos científicos complejos, haciendo hincapié en la resolución de desafíos de optimización específicos. El segundo aborda la implementación práctica de la IA, abordando cuestiones como la limpieza de datos y la mejora de la eficiencia operativa. El tercero destaca el desarrollo de la IA desde su creación hasta su mantenimiento una vez implementada, aprovechando la ingeniería de datos como herramienta clave para mejorar la formación y el rendimiento de la IA. Estos hallazgos son dignos de mención porque arrojan luz sobre el uso actual de aplicaciones de IA para optimizar procesos en diversos sectores. La capacidad de procesar grandes volúmenes de datos rápidamente mejora la eficiencia y acelera la toma de decisiones en sectores como la salud, la industria y la ciberseguridad. Esto permite diagnósticos personalizados, optimización de procesos y respuestas inmediatas a las amenazas. Sin embargo, es importante enfatizar que estos beneficios dependen de la calidad de los datos y de la adecuada implementación de sistemas de análisis, que aseguren un procesamiento efectivo y resultados confiables.

Palabras clave: Inteligencia artificial, Ingeniería de datos, Preprocesamiento de datos, Aprendizaje profundo, Aprendizaje automático, Cienciometría.

1. Introduction

Artificial Intelligence (AI) has experienced exponential growth in recent years, impacting a wide range of knowledge areas. In parallel, Data Engineering (DE) has become an essential pillar for AI development, providing the infrastructure, methods, and tools necessary to collect, transform, and manage large volumes of reliable data [1].

The relationship between AI and DE is symbiotic, as AI relies on well-structured, high-quality data to learn and generate more accurate and reliable results. Meanwhile, DE leverages AI to enhance the effectiveness of data management and analysis. This synergy improves both AI's efficiency and DE's capacity to deliver more useful data. DE creates and maintains systems for data collection, storage, and processing, including the development of pipelines and the integration of data from various sources to ensure quality. Additionally, it

focuses on building scalable architectures and managing data quality through techniques such as cleaning, validation, and continuous monitoring.

This research examines how DE has enhanced AI development, highlighting the main research trends, the most relevant authors and articles on the topic, as well as the countries and journals that have contributed the most.

2. Methodology

This article aims to highlight the relationship between Artificial Intelligence (AI) and Data Engineering (DE), focusing on key aspects such as the symbiosis between the two. DE enriches AI through data quality, while AI facilitates data generation. This article explores fundamental concepts that are crucial to the current development of AI and DE.

The foundation of this article is based on the collection of scientific literature related to AI and DE from SCOPUS and Web of Science (WoS), aiming to identify the initial approaches to these topics. It considers related themes such as natural language and data processing, among others, using the search equation shown in Table I.

Table I. Search Parameters Used in SCOPUS and Web of Science

Parameter	Web of Science	Scopus
Range	2003-2024	
Date	October 05, 2024	
Document Type	Paper, book, chapter, conference proceedings	
Words	(TITLE (("data engineering" OR "data infrastructure" OR "data pipelines" OR "data architecture" OR "data integration" OR "data management" OR "big data engineering" OR "data processing" OR "ETL processes" OR "data warehousing" OR "data quality" OR "data transformation" OR "data science engineering")) AND TITLE ("artificial intelligence" OR "AI" OR "machine learning" OR "deep learning" OR "natural language processing")) AND PUBYEAR > 2002 AND (LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "cp")) AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (SRCTYPE, "j") OR LIMIT-TO (SRCTYPE, "p"))	
Results	237	642
Total (Wos+Scopus)		663

These results were filtered to eliminate unrelated content and duplicates. This process produced a new dataset which, once organized, provided information such as title, author, publication date, journal, and other details. A Tree of Science (ToS) was constructed: the roots are linked to foundational concepts like Deep Learning (DL), the trunk represents Machine Learning (ML), and the branches extend to cases where AI and ID are applied in various scientific fields (Figure 1).

Based on the ToS and the correlation of articles, the analysis highlights the connections between the most experienced authors and countries in using, producing, and integrating AI and ID across different disciplines. Using defined parameters, a coherent chronological structure was created, emphasizing key concepts and the evolution of the research, the ToS construction process, and, importantly, the journals with notable expertise in these areas.

After completing the investigation, which focused on AI, its historical ties with DE, and its role as a fundamental concept in this area, the following research summary is provided.

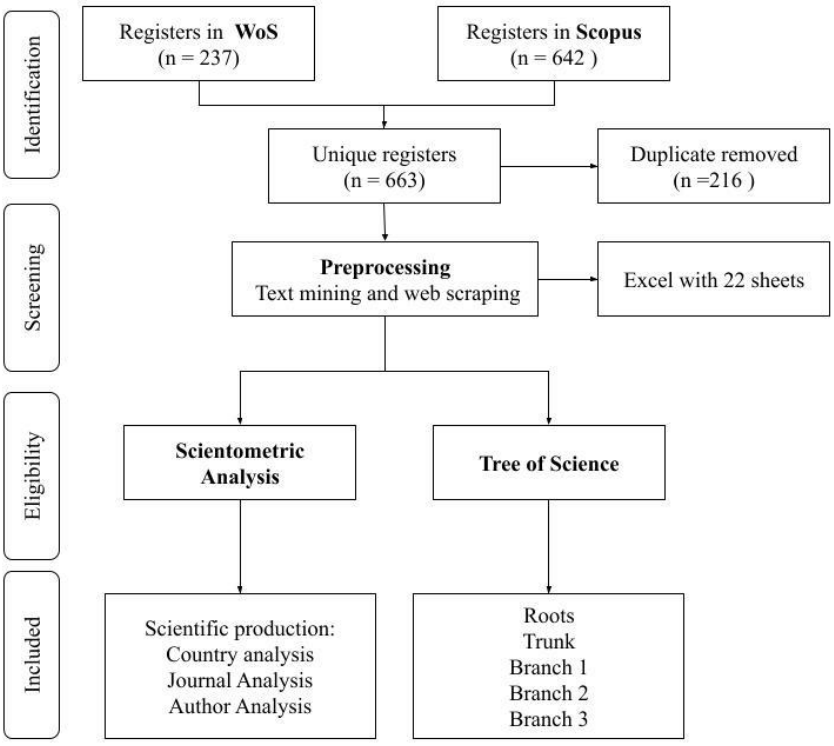

Figure 1. Data Processing Flowchart

3. Results

➢ Scientometric Analysis

Scientific production

It is important to know the annual scientific production as this helps estimate the knowledge production and the interest of the scientific community in exploring and delving into a certain topic. In this case, the topic is related to artificial intelligence and its connection to data engineering. To carry out this analysis of the annual scientific production, an analysis period was designated from 2003 to 2024. In general terms, it is noted that Scopus contains more academic articles compared to WoS in each year, and the annual growth rate over the aforementioned period is 26.62%. Currently, the topic is trending, so research and consequently scientific production have been increasing since 2015(Figure 2).

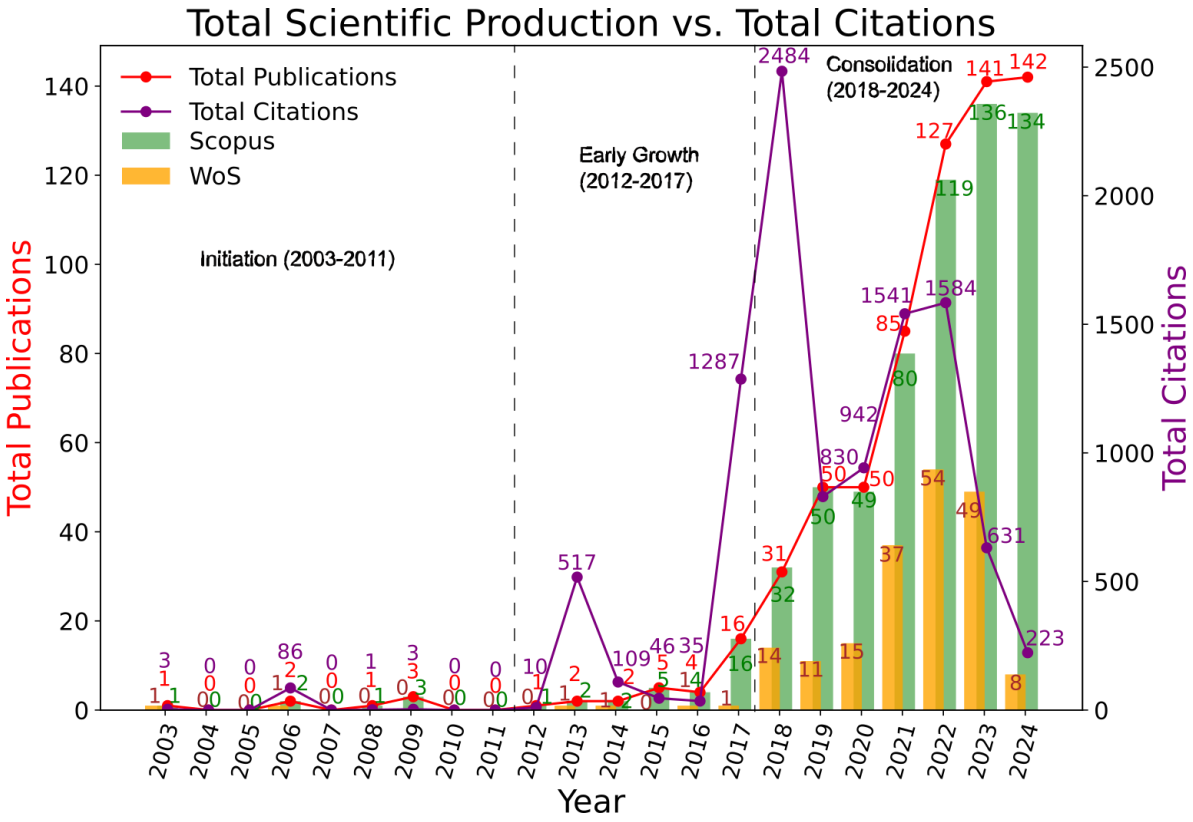

Figure 2. Annual Trends in Scientific Production and Citations in Scopus and WoS (2003–2024): Focus on Data Engineering and Artificial Intelligence

In the previous figure, three periods have been established based on the behavior of the bar chart. In the first interval, there are only a few publications related to the applications of AI and DI in specific sectors such as the governmental sector. In the second interval, more relevant publications begin to emerge as the number of citations starts to increase slightly, compared to the citations in the first interval. In the third interval, the scientific production and citations increase dramatically, generating a significant amount of valuable knowledge and making the AI and DI topic a trend.

Initiation (2003-2011)

In this interval, scientific production was low, with 7 articles representing 1% of the total publications, and years with no publications in either WoS or Scopus. The most cited article in this interval was by Sessions, V., & Valtorta, M. [2], in their research where they mention that data quality is a crucial factor for machine learning tools and that efforts should be focused on its design. Scopus was the database that consistently had more articles in this interval. Another relevant article in this interval was by Holloway et al. [3], who used a Support Vector Machine (SVM), a machine learning algorithm, to make data-driven predictions. In their case, the SVM analyzes patterns to predict relationships between TF (Transcription Factors) and the genes that regulate them. Additionally, they optimized their model by validating the results with known databases.

Early Growth (2012-2017)

In this interval, the topic of AI and DI began to increase scientific production and citations, with a total of 30 articles, representing 4.5% of the total production. Scopus was also the database with the highest number of articles in this interval. The year with the most citations was 2017, with 1287 citations (the third highest value), and the most cited article was by Yao S. et al. [4], who introduced DeepSense, a deep learning framework that combines CNN and RNN. This proposal aims to improve the accuracy of data from mobile device sensors, as these data often have a lot of noise. Another highly recognized article in this interval was by Napolitano, F. et al. [5], which presents a machine learning algorithm for drug repositioning, a process to identify new therapeutic uses for drugs with compounds approved by the FDA. This algorithm achieved a high accuracy of 78%, improving the efficiency and precision of identifying new uses for existing drugs.

Consolidation (2018-2024)

In this period, 626 publications were made, which corresponds to 94.4%, and there were 8,235 citations. The highest number of publications occurred in 2024, while the year with the most citations was 2018, with 2,484 citations, demonstrating that the topic is currently trending and on the rise. Scopus continued to lead with the highest number of articles compared to WoS. Additionally, in this interval, one of the most cited articles is by Manogaran et al. [6], who proposed a data processing framework for large volumes of data

based on machine learning, aiming to detect DNA copies, thereby facilitating cancer detection.

Country Analysis

Table II shows the top 10 countries with the highest scientific production in the field of AI and ID, as well as certain metrics such as citations and quality (measured by Scimago quartiles). The country with the highest productivity is China, with 21.96%, followed by the USA with 15.32%. However, although China leads in scientific production, it is the USA that leads in citations with 23.72%, followed by China, India, and Korea. The latter, despite being in sixth place in scientific production, maintains fourth place in citations with 5.48% of the citations. As for quality, China surpasses other countries in the first three quartiles, only being outdone in quartile 4 by India. AI and ID have positioned themselves in leading research countries and first-world nations like China, the USA, Germany, Italy, etc.

Table II. Breakdown of scientific production and citations by country, including quartile distribution of articles in ranked journals

Country	Production		Citation		Q1	Q2	Q3	Q4
China	139	21.96%	1481	21.14%	54	22	12	4
Usa	97	15.32%	1662	23.72%	23	12	0	2
India	70	11.06%	601	8.58%	12	8	6	12
Germany	31	4.9%	260	3.71%	11	3	2	0
Italy	21	3.32%	117	1.67%	7	2	4	0
Korea	20	3.16%	384	5.48%	5	5	4	0
United Kingdom	18	2.84%	240	3.43%	7	7	0	1
Canada	13	2.05%	236	3.37%	6	3	1	0
Spain	13	2.05%	60	0.86%	7	2	0	1
Australia	12	1.9%	219	3.13%	7	4	0	0

One of China's most recent articles is by Wang J. et al. [7], where they propose a prediction system for wind speed, useful for smart electrical grids dependent on wind energy. They designed and implemented two stages for data preprocessing and used deep learning techniques. This resulted in a more effective prediction model compared to others used in their study.

From the USA side, one of the most recent articles is by Surehali, S. et al. [8], who focus on the use of AI in a prediction system for the performance of sustainable construction binders based on mining tailings and ID for handling complex and limited experimental data.

The most recent study from India is by Jyothi, V., et al. [9], who present a method for data management in smart cities, proposing an AI-based system to optimize security, privacy, and efficiency in handling large volumes of data.

Finally, the most recent article from Korea, by Kim, J. et al. [10], deals with a deep learning algorithm capable of detecting lung nodules in chest X-rays. To achieve this, Kim and his team emphasize the importance of data quality for the success of this algorithm, as they performed pixel-level labeling, improving the algorithm’s accuracy.

Figure 3 shows the scientific collaboration of countries based on co-author affiliations of the same article. The collaboration network presents two large groups led by China and the USA. A relevant piece of work developed by researchers Huang, Molisch, and others [11] affiliated with China and the USA, respectively, uses machine learning techniques to improve the modeling and analysis of vehicle-to-vehicle communication channels, essential for autonomous vehicles, intelligent transport systems, etc. Another significant study was done in cooperation with Germany, the Netherlands, and Italy [12], resulting in SalaciaML, a deep learning algorithm that significantly improves the quality control of ocean data, specifically temperature measurements. This algorithm is characterized by its high precision, successfully classifying more than 90% of the data.

On the other hand, a recent article by researchers Gharieb, Algarhy, Darraj, and others [13], affiliated with Egypt, the USA, and the United Kingdom, implemented an Internal Data Integration Platform (IDIP) to dramatically transform operational efficiency and data management in oil and gas exploration and production operations. With the creation of IDIP, data from various sources was integrated, processes were automated, and discrepancies were identified in real-time. This technological solution was designed to manage, centralize, and structure data. Another collaborative work from the UK but this time with China was by researchers Antonio and Yongsheng [14], who focused on the application of 3D laser scanning technology and deep learning to protect and preserve old building sites. They did this by constructing detailed 3D models, processing image data through deep learning, and automatically removing errors in the data. This demonstrated that the combination of technologies like 3D laser scanning and deep learning can significantly improve the way architectural heritage is documented and preserved.

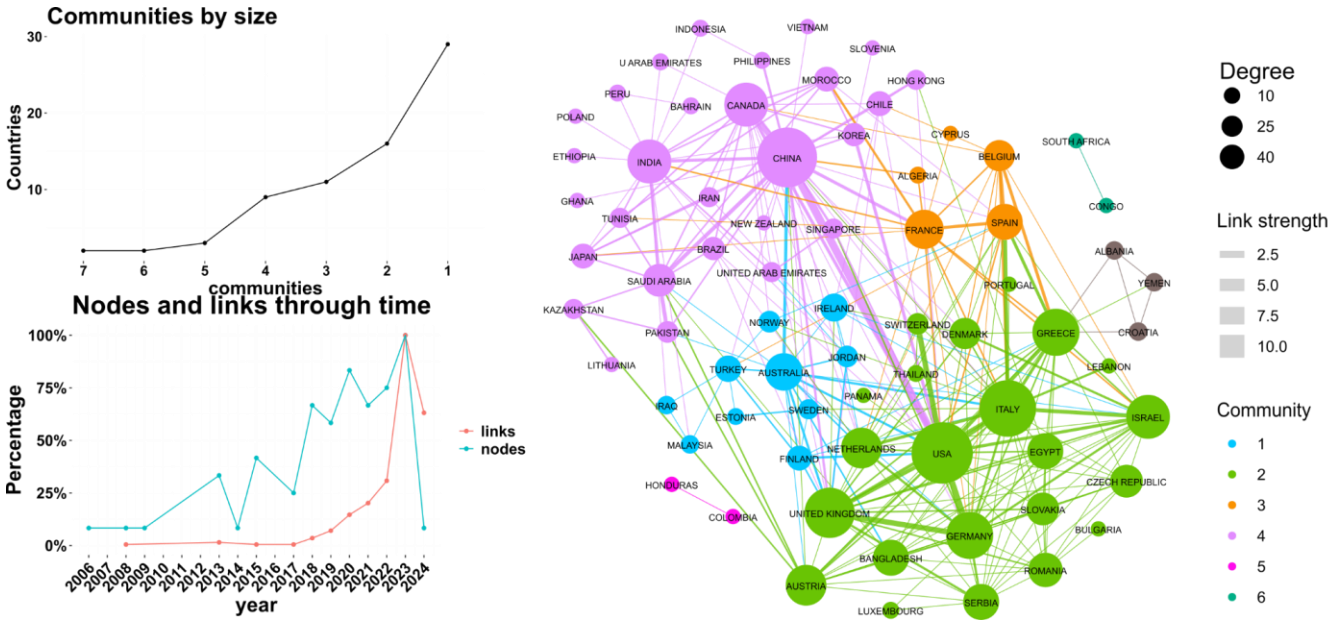

Figure 3. Global Network of Scientific Research Collaboration Between Countries Journal Analysis

Journal Analysis

Table III shows the top 10 journals with the highest scientific production on AI and ID, including their impact factor, h-index, and Scimago quartile. The journal with the most articles is IEEE Access, with an h-index of 242 and located in Q1. One of the most relevant articles from this journal is the work by Zheng, S. et al. [15], in which they propose a new architecture to process and manage radio signals through deep learning. One of the most recent papers is by Zhang, K. [16], who proposes a pipeline that not only focuses on the design and optimization of the model but also on aspects related to data management and preparation. This addresses the need for a broader and more comprehensive perspective in the development of deep learning models. Among its results, it demonstrates that this pipeline improves organization, efficiency, and evaluation capacity compared to traditional approaches.

The journal ACM International Conference Proceeding Series is the second with the highest number of articles. One of the most relevant papers mentions the importance of data management in the context of AI applied to government (public policy). It is pointed out that without proper data management, simple collection or acquisition of systems is not enough

to achieve good results. This paper presents the idea that for AI to be beneficial to society, it is necessary to invest in processes that build trust in governmental data [17].

In conclusion, Table 3 demonstrates the rise and maturity of AI and ID, as they are positioned in high-quality journals.

Table III. High-Impact Journals in AI and DE: Country, h-index, and Quartile Metrics

Journal	N.º	%	Country	Impact factor	Quartile	H. Index
IEEE Access	16	0.02%	USA	0.96	Q1	242
ACM International Conference Proceeding Series	15	0.02%	USA	0.253	-	151
Proceedings Of The Acm Sigmod International Conference On Management Of Data	12	0.02%	USA	2.64	-	177
Ceur Workshop Proceedings	9	0.01%	Germany	0.191	-	66
Journal Of Physics: Conference Series	8	0.01%	United Kingdom	0.18	-	99
Sensors	6	0.01%	Switzerland	0.786	Q1	245
Plos One	5	0.01%	USA	0.839	Q1	435
Proceedings Of Spie - The International Society For Optical Engineering	5	0.01%	USA	0.152	-	193
Proceedings Of The Vldb Endowment	5	0.01%	USA	2.666	Q1	153
Applied Sciences	4	0.01%	Switzerland	0.508	Q2	130

Figure 4 shows the citation network between journals, highlighting the 3 main groups. In the first group, studies related to AI and ID applied to the fields of medicine and biology are emphasized, where research focuses on improving tools to make certain processes more efficient [18], [19]. The second group deals with advanced data processing with AI in the field of computer science, specifically in IoT, cybersecurity, data management, etc. [20], [21]. Finally, the third group focuses on optimizing processes in various fields, such as the use of AI for structuring and analyzing medical data related to pancreatic cancer [22]. Another case is the article by Shen, Y. et al. [23], where they optimize an industrial process for wastewater treatment. In their case, the use of IoT was limited by the volume of data being generated. With the help of deep learning (DL), they were able to preprocess the data and eliminate redundant data, transmitting and storing only the most relevant. The "nodes and links through time" figure shows that the proportion of nodes was initially higher compared to the links, except in the last two years (2023 and 2024), where the links slightly outnumbered the nodes. This means that, initially, various journals contributed to the AI and ID theme, but lately, cooperation and relationships between international journals have been strengthening.

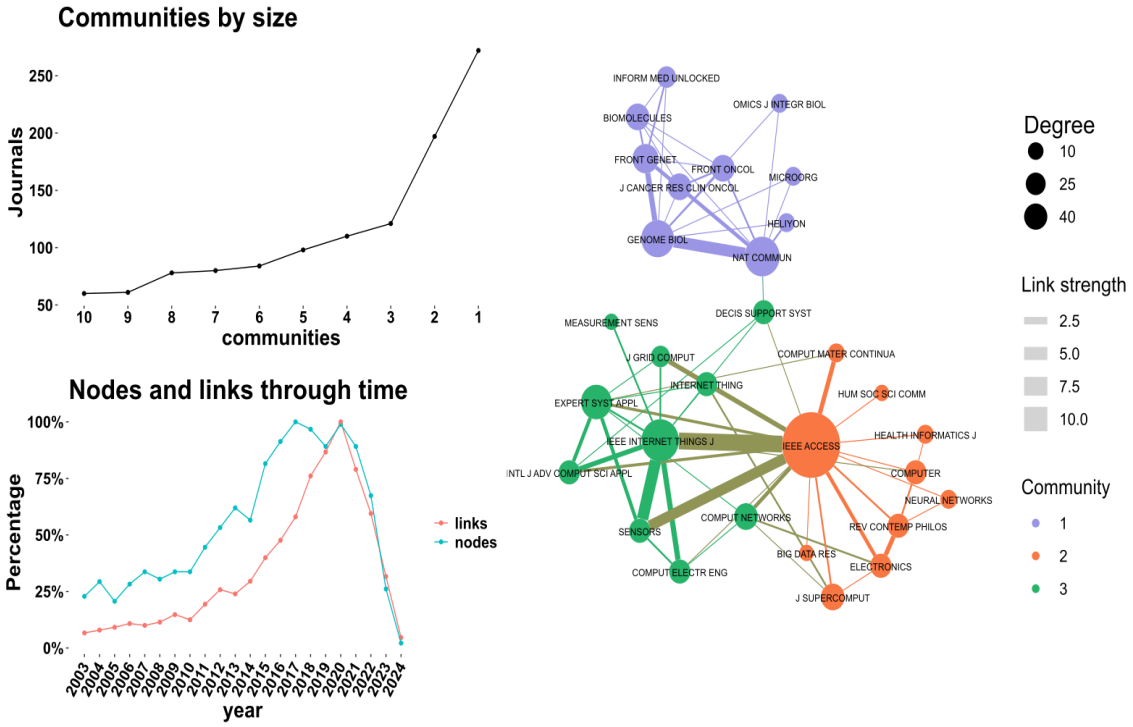

Figure 4. Network of Journal Citations Highlighting Collaborative Communities

Author Collaboration Network

Table IV lists the top 20 researchers with the highest productivity in the field of AI and ID. It can be seen that Wang J is the researcher who has written the highest number of publications on this topic. Among his most recent articles is a study focused on using AI for teaching in a vocational course, utilizing a structured model of blended learning, i.e., before, during, and after the class [24]. Another prominent author is Li Y, and one of his standout studies is a comprehensive review of data integration techniques from a machine learning perspective, emphasizing that the incorporation of data from multiple sources is key for successful analysis [25].

Table IV. Author Productivity and Institutional Affiliations

Author	Total Publications	Affiliation
WANG J	13	School of Artificial Intelligence, Guangzhou, China
LI Y	9	Digital Technologies Research Center, Ottawa, Canada
LIU J	9	University of Sydney, Sydney, Australia
WANG Z	9	University-Hong Kong Baptist University United International College, Zhuhai, China
ZHANG Y	9	Sun Yat-Sen University, Guangzhou 510275, China
LI J	8	Capital Medical University, Beijing, China
LIU X	7	Shandong University of Science and Technology, Qingdao, China
LIU Y	7	Taiyuan University of Technology, Taiyuan, China
LIU Z	7	Shenzhen University, Shenzhen, China
WANG Y	7	Central South University, Changsha, China
WANG W	6	University of Waikato, New Zealand
WANG X	6	IBM T. J. Watson Research Center, NY, USA
GUPTA N	5	Delhi Technological University, New Delhi, India
LI M	5	Toronto Metropolitan University, Toronto, Canada
SUN L	5	China University of Geosciences, Wuhan, China
ZHANG H	5	Research Institute of Safety Technology, Beijing, China
ZHANG X	5	Xi'an Jiaotong University, Xi'an, China
BORDAWEKAR R	4	IBM T.J. Watson Research Center, NY, USA
GUPTA A	4	Bharathidasan Institute of Management, Tiruchirappalli, India
JIN Y	4	China University of Petroleum., Beijing, China

Figure 5 shows the scientific collaboration network between the main authors. This network is built from their personal networks or egonetworks. In this case, the collaboration network shows a single group, as this group is much larger than the others, so it is sufficient to review this one group. One of the main researchers is Zhang Y, with his research where he develops a three-dimensional single molecule localization algorithm (3D-SMLM), which allows observing the exact position of molecules in 3D. This tool is very useful in biomedical research and is characterized by the complexity of the data [26].

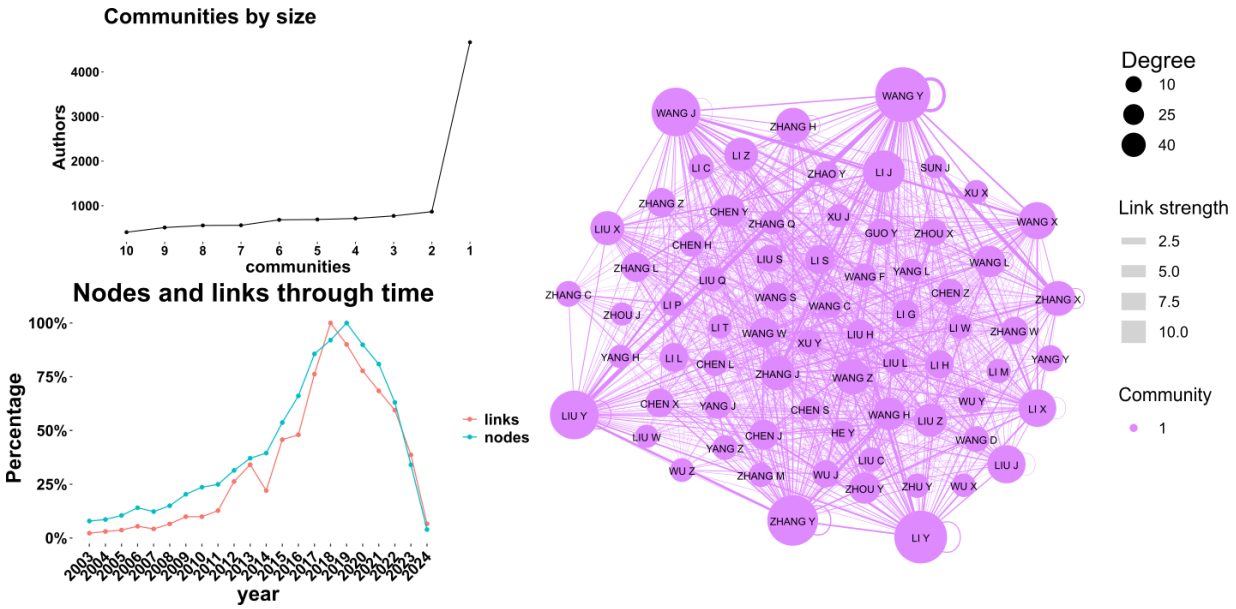

Figure 5. Collaborative Network of Prominent Authors and Their Interconnections

Conclusions

The present study conducts a scientometric review of the close relationship between DE and AI, using the ToS algorithm based on a search in Scopus and WoS. The results show that this topic has been explored from its theoretical foundations, where advances in neural networks and deep learning have enabled solving complex classification, prediction, and pattern recognition problems in large and varied datasets. The scientometric analysis revealed key contributors, such as prominent authors Wang J. and Li Y., as well as influential journals like IEEE Access and Sensors, which have been central to repeated collaborations by these authors. Among the countries with the highest production, China and the USA stand out, while South Korea, despite having fewer publications compared to others, exhibits a significant number of citations, reflecting the impact of its contributions.

Therefore, these developments have facilitated the creation of architectures for biomedical image processing and sequence data analysis in areas such as natural language processing and real-time prediction. Subsequently, the growth in practical applications highlights the use of these technologies in areas such as precision medicine, improving clinical predictions through the integration of omics data; in IoT network security, optimizing real-time anomaly detection; and in the automotive industry, where monitoring systems enable automatic detection of production failures. Together, these advances underline the relevance of a solid DE infrastructure to maximize the potential of AI in extracting valuable information and making decisions in increasingly complex and demanding environments.

Three major trends were also identified. The first refers to the use of deep learning (DL) models in processing complex data for various scientific and engineering applications. These models have proven particularly effective in extracting precise details from microscopic images, such as in the analysis of advanced materials using electron microscopy, and in the segmentation of structural images, where the quality and variety of training data are crucial for model accuracy. Additionally, DL approaches are applied in noise removal from geophysical data, improving subsurface structure interpretation, and in structural health monitoring of infrastructure, enabling early identification of potential failures. Finally, the use of AI in spectrometric analysis speeds up the classification of compounds in consumer products, optimizing safety and efficiency in these processes.

The second trend highlights the advanced applications of AI and DL across various sectors such as healthcare, education, and manufacturing, focusing on data management and optimization. In the healthcare sector, innovations are emphasized in the accessibility and organization of oncological data, adopting a holistic approach to care and monitoring. This involves not only disease treatment but also considers preventive aspects, long-term follow-up, and the patient's quality of life. In education, DL is used to analyze musical patterns, enabling the development of adaptive and personalized instruction tailored to each individual's learning process. This suggests a shift from traditional teaching methods, providing new, dynamic, and more effective educational experiences. In manufacturing, predictive maintenance stands out by using patterns developed through DL. This ensures data reliability and prevents critical operational failures, improving efficiency by minimizing downtime, reducing costs, and enhancing industrial effectiveness, thus enabling operational continuity.

The third trend focuses on optimizing data management for end-to-end ML and AI applications in business and industrial environments. This approach addresses challenges such as preparing and processing large volumes of data, highlighting the combination of classical algorithms with AI to improve business analysis and decision-making. Data quality and comprehensive management are fundamental throughout the AI systems lifecycle, especially in cloud infrastructures that handle massive data. In the industrial field, the use of

ML for fault detection in critical systems, such as aircraft landing gear, demonstrates how integrating intelligent sensors and efficient data management can improve operational safety and precision in preventive diagnostics.

The research identified some common limitations in the field of scientometrics, such as the database query (WoS and Scopus) being limited to 2003 onwards and English as the language. However, ToS helps reduce this bias by integrating references, which allows for finding articles from earlier dates. Additionally, few articles other than in English were found in the search, so these limitations are not expected to significantly affect the final results.

For future research, it is suggested to include the keyword "Big Data," as it is a recurring theme that, although not explicitly mentioned, is always implied when discussing large volumes of data and DE. This is a topic that has been pointed out even in root articles. And although it is a relatively new term, it is gaining strength and could broaden this scientometric perspective.

References

[1]C. Ordonez, W. Macyna, and L. Bellatreche, "Data engineering and modeling for artificial intelligence", Data Knowl. Eng., vol. 153, p. 102346, 2024.

[2]V. Sessions and M. Valtorta, "The effects of data quality on machine learning algorithms", in Proceedings, pp. 485–498, 2006.

[3]D. T. Holloway, M. A. Kon, and C. DeLisi, "Machine learning methods for transcription data integration", IBM J. Res. Dev., vol. 50, pp. 631–643, 2006.

[4]S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "DeepSense", in Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2017. doi: 10.1145/3038912.3052577.

[5]F. Napolitano et al., "Drug repositioning: a machine-learning approach through data integration", J. Cheminform., vol. 5, p. 30, 2013.

[6]G. Manogaran et al., "Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering", Wirel. Pers. Commun., vol. 102, pp. 2099–2116, 2018.

[7]J. Wang, X. Niu, L. Zhang, Z. Liu, and X. Huang, "A wind speed forecasting system for the construction of a smart grid with two-stage data processing based on improved ELM and deep learning strategies", Expert Syst. Appl., vol. 241, p. 122487, 2024.

[8]S. Surehali, T. Han, J. Huang, A. Kumar, and N. Neithalath, "On the use of machine learning and data-transformation methods to predict hydration kinetics and strength of alkali-activated mine tailings-based binders", Constr. Build. Mater., vol. 419, p. 135523, 2024.

[9]D. V. Jyothi, D. T. Sreelatha, D. T. M. Thiyagu, R. Sowndharya, and N. Arvinth, "A data management system for smart cities leveraging artificial intelligence modeling techniques to enhance privacy and security", J. Internet Serv. Inf. Secur., vol. 14, pp. 37–51, 2024.

[10]J. Y. Kim, W.-S. Ryu, D. Kim, and E. Y. Kim, "Better performance of deep learning pulmonary nodule detection using chest radiography with pixel level labels in reference to computed tomography: data quality matters", Sci. Rep., vol. 14, p. 15967, 2024.

[11]C. Huang et al., "Machine-learning-based data processing techniques for vehicle-to-vehicle channel modeling", IEEE Commun. Mag., vol. 57, pp. 109–115, 2019.

[12]S. Mieruch, S. Demirel, S. Simoncelli, R. Schlitzer, and S. Seitz, "SalaciaML: A deep learning approach for supporting ocean data quality control", Front. Mar. Sci., vol. 8, 2021.

[13]A. Gharieb et al., "In-house integrated big data management platform for exploration and production operations digitalization: From data gathering to generative AI through machine learning implementation using cost-effective open-source technologies - experienced mature workflow", in Day 2 Tue, April 23, 2024, SPE, 2024. doi: 10.2118/218560-ms.

[14]Y. Yin and J. Antonio, "Application of 3D laser scanning technology for image data processing in the protection of ancient building sites through deep learning", Image Vis. Comput., vol. 102, p. 103969, 2020.

[15]S. Zheng et al., "Big data processing architecture for radio signals empowered by deep learning: Concept, experiment, applications and challenges", IEEE Access, vol. 6, pp. 55907–55922, 2018.

[16]K. Zhang, "Incorporating deep learning model development with an end-to-end data pipeline", IEEE Access, vol. 12, pp. 127522–127531, 2024.

[17]T. Harrison et al., "The data firehose and AI in government", in Proceedings of the 20th Annual International Conference on Digital Government Research, ACM, New York, NY, USA, 2019. doi: 10.1145/3325112.3325245.

[18]X. Xiang et al., "Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning", Nat. Commun., vol. 12, p. 3238, 2021.

[19]Y. Yu et al., "Improved prediction of bacterial CRISPRi guide efficiency from depletion screens through mixed-effect machine learning and data integration", Genome Biol., vol. 25, p. 13, 2024.

[20]O. P. Olawale and S. Ebadinezhad, "Cybersecurity anomaly detection: AI and Ethereum blockchain for a secure and tamperproof IoHT data management", IEEE Access, vol. 12, pp. 131605–131620, 2024.

[21]R. Koulali, H. Zaidani, and M. Zaim, "Image classification approach using machine learning and an industrial Hadoop-based data pipeline", Big Data Res., vol. 24, p. 100184, 2021.

[22]G. Manias et al., "Advanced data processing of pancreatic cancer data integrating ontologies and machine learning techniques to create holistic health records", Sensors (Basel), vol. 24, p. 1739, 2024.

[23]Y. Shen et al., "A deep-learning-based data-management scheme for intelligent control of wastewater treatment processes under resource-constrained IoT systems", IEEE Internet Things J., vol. 11, pp. 25757–25770, 2024.

[24]J. Wang, "Practical research on blended teaching in higher vocational institutes based on ICVE platform: Take artificial intelligence data processing course as an example", in Proc. 2022 3rd Int. Conf. Educ., Knowl. Inf. Manag. (ICEKIM), IEEE, 2022. doi: 10.1109/ICEKIM55072.2022.00047.

[25]Y. Li and A. Ngom, "Data integration in machine learning", in Proc. 2015 IEEE Int. Conf. Bioinf. Biomed. (BIBM), IEEE, 2015. doi: 10.1109/BIBM.2015.7359925.

[26]S. Guo, J. Lin, Y. Zhang, and Z.-L. Huang, "Enhancing the data processing speed of a deep-learning-based three-dimensional single molecule localization algorithm (FD-DeepLoc) with a combination of feature compression and pipeline programming", J. Innov. Opt. Health Sci., 2024. doi: 10.1142/S1793545824500251.