Machine Learning and Geographic Information System synergies in AEC domain: A Literature Review and Future Opportunities

Yifeng Sun[1]; Xianfei Yin[1,2]; Jianyu Yin[1]; Chi Chiu Lam[1]

[1] Department of Civil and Environmental Engineering, Faculty of Science and Technology, University of Macau, Avenida da Universidade, Taipa, Macau, China
2 xfyin@um.edu.mo

Abstract. The recent rapid advancements of Machine Learning (ML) have significantly empowered and transformed many research domains, including architecture, engineering, and construction (AEC) area. Meanwhile, in AEC industry, Geographic Information System (GIS) is also a commonly used tool. In order to better understand the state-of-the-art ML and GIS synergies, this research conducts a bibliometric analysis to reveal hidden information from relevant literature and explore the integration cases of ML and GIS in the AEC domain. As a result, a dataset of 3387 relevant articles (including articles and conference papers) published from 2010 to 2023 was retrieved from Scopus and further analyzed. The research employs visualization techniques to highlight key publications, active research institutions, key researchers, influential journals, etc. Several research themes have been identified as the most studied areas. As a result, research gaps are captured, and several promising future research directions are suggested.

Keywords: Machine Learning, Geographic Information System, literature review, AEC, Bibliometric Analysis.

1 Introduction

The architecture, engineering, and construction (AEC) industry has undergone rapid digital transformation with the advancement of information technologies over the past years [1]. The integration of advanced computational methods including ML and spatial analytics with GIS have provided further capabilities to enhance processes across the AEC workflow [2].

ML offers intelligent algorithms that can learn from data to make predictions, classifications, and discoveries [3]. Techniques such as neural networks, support vector machines and random forests have enabled new applications in defect detection [4], schedule optimization [5], cost estimation [6] and safety planning [7]. In the meantime, GIS provides platforms to capture, store, analyze and visualize spatial information that is critical for AEC operations. The combination of ML and GIS has the potential to enable intelligent geospatial analytics to support data-driven decision-making related to site selection [8], infrastructure monitoring [9], resource allocation [10], disaster resilience [11], and etc.

While existing reviews have examined ML techniques for GIS-based land-use classification [12], transportation modeling [13], and urban planning [14], there lacks a systematic, bibliometric analysis focused on the general development and trends in AEC domain targeting the synergies between ML and GIS. bibliometric review employs quantitative methods to analyze academic publications and examine the structure and evolution of a research field [15]. A bibliometric review can provide insights into knowledge themes, influential works, active institutions, and collaboration networks that characterize the progress on ML-GIS integration in AEC [16]. The findings offer strategic guidance on research gaps and future directions.

This study conducts a bibliometric analysis of literature retrieved from Scopus focusing on ML and GIS in the AEC domain from 2010 to 2023. Growth trends, contributing countries/regions, pivotal publications and journals, collaboration networks, and hot topics are examined to map the knowledge landscape. The results provide an evidence-based perspective on the development trajectory, knowledge structure, and priorities of this interdisciplinary domain. The identified research gaps reveal opportunities to advance ML-GIS synergies and applications in AEC. Overall, this review contributes a comprehensive evaluation and outlook of the topic, which potentially benefits researchers and practitioners in their future endeavors.

2 Methodology

The objective of this review is to delineate the current research progress in integrating ML with GIS within the AEC field, while precisely identifying existing research gaps and proposing future research directions. Bibliometric analysis [17] [18] has been employed to achieve this goal (Fig. 1).

In conducting this research, literature relevant to the topic was identified through a carefully designed search strategy in the Scopus database using the following logic: (TITLE-ABS-KEY (ml OR (machine AND learning) OR (deep AND learning)) AND TITLE-ABS-KEY (gis OR (geographic AND information AND system) OR (geographic AND information AND systems))) AND (LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "cp")). Using dataset obtained from Scopus, it can be imported into visualization software for bibliometric analysis.

In order to analyze the literature on the application of ML in GIS, two leading visualization analysis tools were utilized: VOSviewer [19] and Citespace [20]. These tools enabled the creation of diverse visualizations tailored to specific demands, effectively mapping the distribution and evolution of literature within the knowledge domain [21].

The study incorporated three types of bibliometric analyses:

- Statistics of the dataset: This step involves analyzing basic statistical features of the dataset, including the number of publications, distribution by publication year, types of publications, to gain an overview of the target research field.

- Analyze of countries/regions and institute: Detecting patterns of co-occurrence among countries and institutions. Two analytical approaches are employed to demonstrate the geographical distribution and influence of the publications within the target research field: 'Publication trends' and 'Citations and citations per publication'.
- Co-authorship network: Detecting patterns of co-authorship. Generate a co-authorship network using Citespace and export the centrality of author collaboration.
- Author co-citation network: Identifying frequently cited authors and literature, revealing academic influence and research hotspots. Generate an author co-citation network using Citespace and identify key co-cited authors.
- Keyword co-occurrence analysis: Exploring trends in the co-occurrence of keywords to understand the conceptual structure of target research field. Visualizing with VOSviewer helps reveal trends in the development of hot points over the years and the interrelationships between various research directions within AEC domain.

This method section outlines the framework for the bibliometric analysis of the current state and prospects of the combined use of ML and GIS in the AEC field.

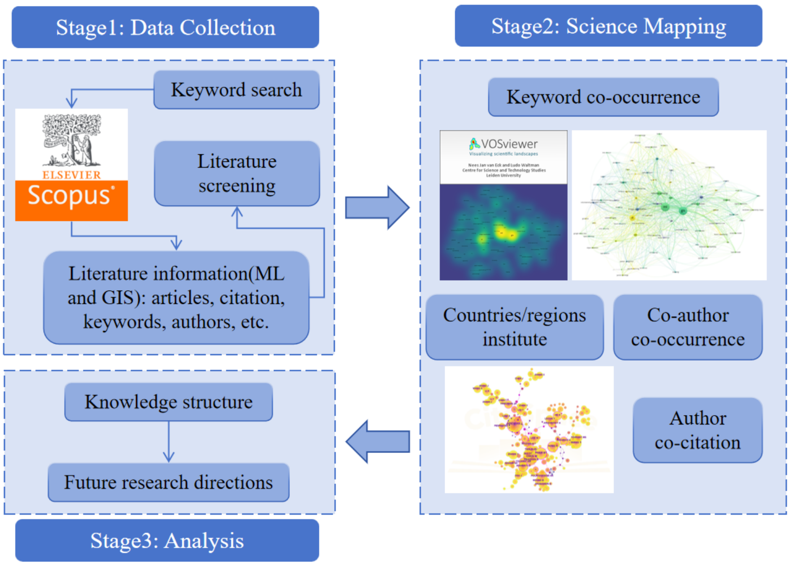

Fig. 1. Overview of the proposed research methodology.

3 Results

3.1 Statistics of the dataset

The study topic is gaining traction within both journals and conference proceedings. As shown in Fig. 2, there has been a slight increase in publications on ML-GIS within AEC domain from 2010 to 2016. A notable surge in publication numbers was witnessed from 2017 through 2023. Even the number of papers published in 2023 reached 700, indicating the widespread application of ML-GIS in the AEC domain.

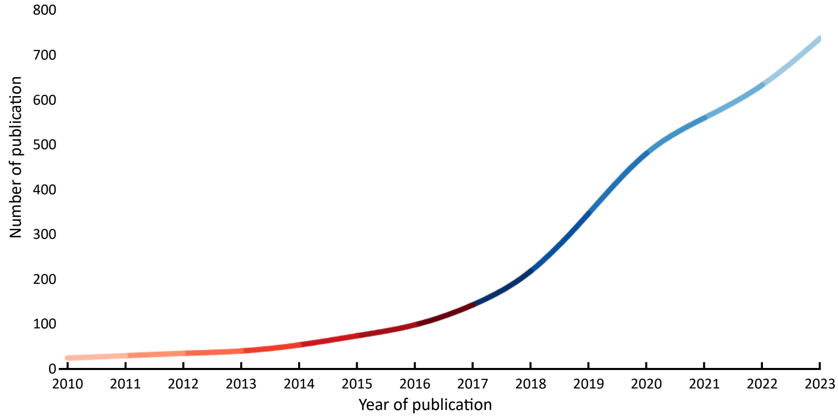

Fig. 2. Annual publication distribution in ML-GIS from 2010 to 2023.

Using the keyword search strategy outlined in Section 2, distribution of number of papers regarding journal and conference sources are tabulated in Table 1 and 2. The majority of academic publications regarding the ML-GIS in AEC domain were published in journals including Remote Sensing, ISPRS Annals of Photogrammetry Remote Sensing and Spatial Information Sciences, ISPRS International Journal of Geo Information, and etc. Among these journals, Remote Sensing published the largest number of papers on this topic. Moreover, conference proceeding that have made significant contributions to the field include the IEEE Conference on International Society for Optical Engineering and International Geoscience and Remote Sensing Symposium Proceedings.

Table 1. List of academic journals that published related to ML-GIS (2010-2023).

Journal title	Number of relevant articles	% Total publications
Remote Sensing	138	5.77%
ISPRS Annals of the Photogrammetry Remote Sensing and Spatial Information Sciences	59	2.47%
ISPRS International Journal of Geo Information	56	2.34%
Sustainability	48	2.01%
International Journal of Geographical Information Science	45	1.88%
Science of the Total Environment	44	1.84%
IEEE Access: Practical Innovations, Open Solutions	43	1.80%
Geocarto International	42	1.76%

Table 2. List of conference proceedings that published related to ML-GIS (2010-2023).

Conference title	Number of relevant articles	% Total publications
International Society for Optical Engineering	40	4.17%
International Geoscience and Remote Sensing Symposium Proceedings	29	3.02%
Conference Series Earth and Environmental Science	26	2.71%
International Conference Proceeding Series	24	2.50%
International Proceedings in Informatics	16	1.67%
Journal of Physics Conference Series	15	1.56%
Ceur Workshop Proceedings	10	1.04%

3.2 Analyze of countries/regions and institute

Leveraging bibliometric analysis, the extracted data reflects the publication distribution across various countries and regions in the field of ML-GIS. The data reveals that China (877 articles), the United States (686 articles), India (369 articles), Iran (236 articles), Australia (205 articles), South Korea (187 articles), the United Kingdom (161 articles), Vietnam (150 articles), Germany (146 articles), and Italy (136 articles) stand out as the predominant contributors in the field. The volume of articles published by a country or region underscores the emphasis placed on ML-GIS integration within the AEC domain. Fig. 3 delineates the annual publication trends of the top five countries from 2010 to 2013, illustrating a consistent upward momentum in publication numbers for China and India, while the United States, Iran, and Australia have exhibited a stable publishing activity in recent years.

Citation count serves as another pivotal indicator reflecting the recognition of research by a country or region. Higher citation counts suggest a better quality of research. As shown in Fig. 4, the top five countries by total number of citations include the United States (12,019 citations), China (10,390 citations), Australia (3,102 citations), Iran (2,895 citations), and United Kingdom (2323 citations). Furthermore, calculating the Citation Per Publication (CPP) by dividing the total number of citations by the total number of publications serves as an effective measure of the average quality of research papers. The United States has the highest CPP at 17.52, which means each paper is cited by more than 17 papers on average.

The contributions from various institutions towards ML-GIS integration within AEC domain have been evaluated. These institutions demonstrate substantial activity within the targeted research area, as evidenced by the Chinese Academy of Sciences (99 articles), the University of Technology Sydney (94 articles), Duy Tan University (79 articles), and Sejong University (75 articles). However, no institution can be considered the main research center around the world as they represent a low percentage of the world-wild research in the target research field.

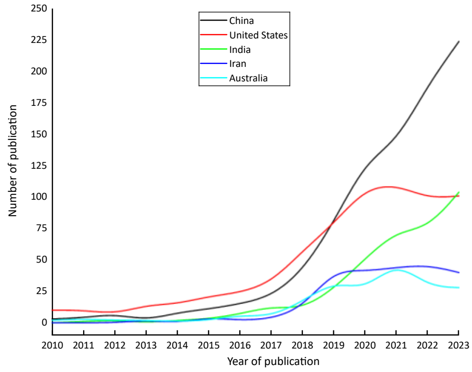

Fig. 3. Publication trends for countries (2010-2023).

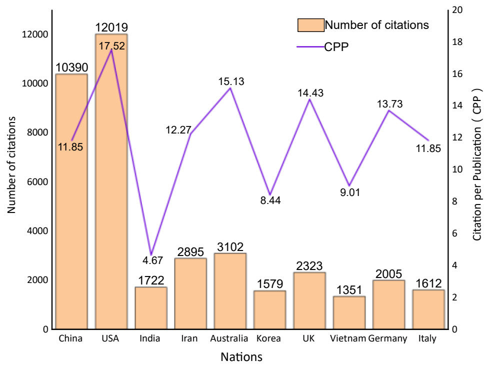

Fig. 4. Citations and citations per publication by countries (2010-2023).

3.3 Co-authorship network

Information about authors can be extracted from bibliographic records, facilitating the identification of leading researchers and their collaborative network in the targeted research area. This process enables the creation of a co-authorship network.

The top 10 most productive authors were initially identified based on the volume of published papers (Table 3). The leading positions are held by B. Pradhan from the University of Technology Sydney, D. Tien Bui from the University of South-Eastern Norway, and B.T. Pham from the University of Transport Technology.

A co-authorship network can be generated using Citespace. As shown in Fig. 5, a node possessing high centrality typically links two or more significant clusters of nodes. In this Figure, authors demonstrating the highest centrality include BA. Baharin

(Centrality= 0.12), D. Tien Bui (Centrality= 0.07), W. Chen (Centrality= 0.06), S. Himan (Centrality= 0.06), S. Li (Centrality= 0.05), C. Kamran (Centrality= 0.05).

Table 3. List of the top 10 most productive authors (2010-2023).

Author	Institution	Country	Count	Percentage
B. Pradhan	University of Technology Sydney	Australia	94	2.78%
D. Tien Bui	University of South-Eastern Norway	Norway	62	1.83%
B.T. Pham	University of Transport Technology	Vietnam	41	1.21%
S. Lee	Institute of Geoscience and Mineral Resources	Korea	32	0.94%
A. Arabameri	Tarbiat Modares University	Iran	30	0.86%
H.R. Pourghasemi	Shiraz University	Iran	27	0.80%
W. Chen	Xi'an University of Science and Technology	China	26	0.77%
R. Costache	National Institute of Hydrology and Water Management	Romania	26	0.77%
H. Shahabi	University of Kurdistan	Iran	26	0.77%
I. Prakash	Geological Survey of India	India	24	0.71%

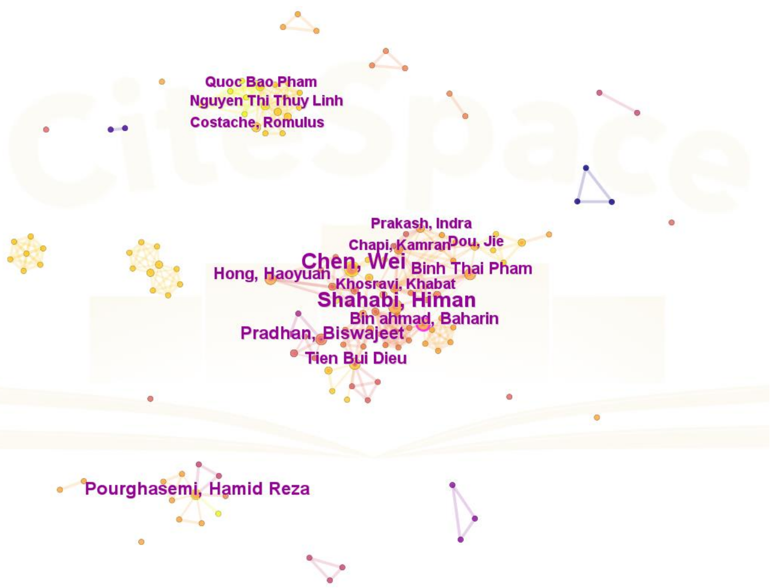

Fig. 5. Co-authorship network.

3.4 Author co-citation network

A co-citation network is a widely utilized bibliometric method that reveal the interconnected relationships among authors. Fig. 6 shows the co-citation network, comprising 121 nodes and 381 links (Fig. 6). The network's node size represents the frequency of an author being co-cited with others, where larger nodes denote a higher number of co-citations. The connecting lines between authors symbolize co-citation relationships, with thicker lines representing stronger relevance between the research themes of the linked authors. Nodes with high centrality typically connect two or more large nodes, and they are positioned centrally. As shown in Fig. 6, nodes exhibiting high centrality are distinguished by darker exterior rings (in purple). Consequently, the authors with the highest centrality include L. Atakew (Centrality= 0.52), B.T. Pham (Centrality= 0.49), W. Chen (Centrality= 0.4), D. Tien Buid (Centrality= 0.38), H.R. Pourghasemi (Centrality= 0.35), I.D. Moore (Centrality= 0.31).

The analysis conducted on the dataset comprising 3387 articles imported into Citespace ensured computation, resulting in co-citation frequencies. The authors with the highest co-citation frequencies include L. Breiman (Frequency= 560), W. Chen (Frequency= 335), Y. Wang (Frequency= 291), B.T. Pham (Frequency= 286), S. Lee (Frequency= 255), H.I. Hong (Frequency= 244), B. Pradhan (Frequency= 233), H.R. Pourghasemi (Frequency= 231), and D. Tien Buid (Frequency= 219). Several of these frequently cited authors also rank high in the number of publications. Moreover, the diversity of their affiliated countries and institutions reflects the global interest and diverse institutional engagement in this field. The intersection of authors with high centrality and those frequently co-cited indicates a correlation between higher co-citation frequency and a stronger central role in the network.

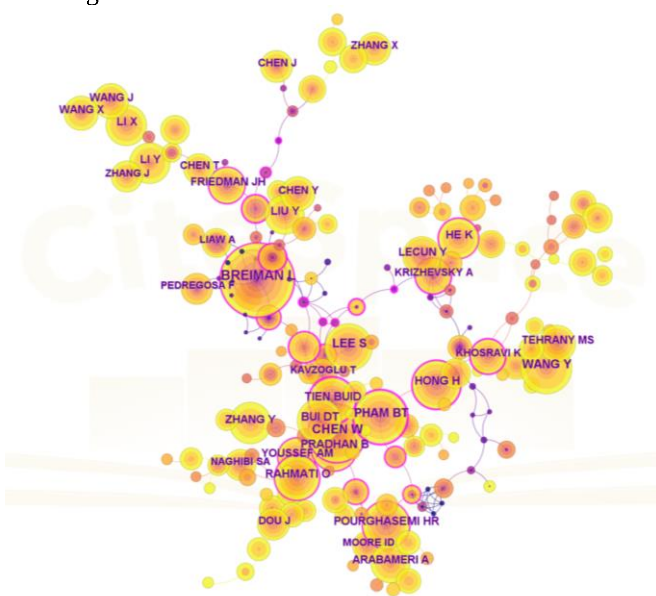

Fig. 6. Author co-citation network in Citespace.

3.5 Keyword co-occurrence analysis

The keywords of a published article represent its core themes and can serve to illuminate the research focus. As shown in Fig. 7, a co-occurring keywords network in the ML-GIS integration within AEC domain was constructed by utilizing VOSviewer. The network embodies the crucial content of the research and its associated subfields. Within this network, distances between two keywords describe their relationship. The closer the distance, the closer the relationship [22]. The network uses the size of keyword to indicate their frequency of occurrence. The larger the size of keyword, the higher the frequency of keyword occurrence [23].

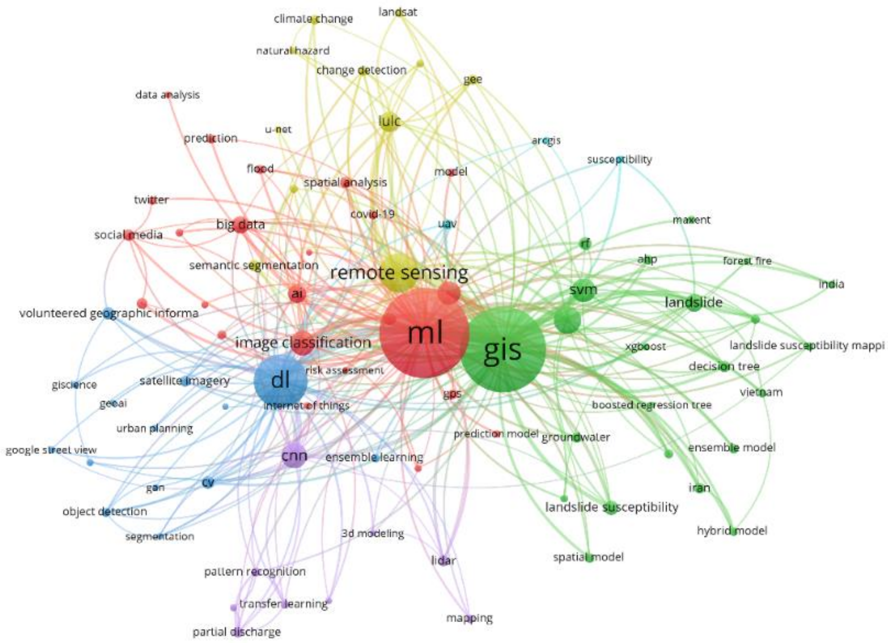

Fig. 7. Author keywords co-occurrence of ML-GIS in VOSviewer.

In this study, data from 3,387 publications indexed in Scopus were compiled to unravel the intricate web of keyword relationships in the field. After importing this data into VOSviewer and setting a threshold for keyword occurrence at 18 so that 86 meet the threshold. This threshold selection was based on multiple experiments with other parameters to generate the optimal clusters. As shown in Table 4, the number of keyword occurrences signifies the frequency of each keyword's appearance in the 3387 articles. For instance, ML, GIS, Deep Learning, Remote Sensing, Random Forest, and Image Classification emerged as particularly frequent. This suggests that these themes are widely explored in existing research, or these tools and methods are extensively utilized in the field.

Research themes in the integration of ML-GIS have evolved over time. As shown in Table 4, research on Image Classification [24], Data Mining [25], Support Vector

Machines [26], and Big Data [27] garnered significant attention in the early stages, with the average appearance year of keywords being 2018. Subsequently, researchers shifted their focus to Artificial Neural Networks [28], Remote Sensing [29], and Random Forest [30], with the average appearance year of keywords being 2020 at that time. The emergence of neural networks led to significant changes in computer technology, prompting more researchers to propose the construction of intelligent neural networks during that period. From 2021 to the present, attention has shifted towards Convolutional Neural Network [31], Deep Learning [32], and Artificial Intelligence [33], indicating a convergence of DL and GIS.

Table 4. Quantitative measurements of cooccurrence keywords in ML-GIS.

keywords	Occurrences	Total link strength	Average citations	Average year
Machine learning	1015	1461	18.43	2020
Geographic information system	943	1544	24.96	2020
Deep learning	441	607	14.02	2021
Remote sensing	267	584	22.64	2020
Random forest	146	264	27.99	2020
Convolutional neural network	134	251	13.82	2021
Image classification	122	266	16.80	2018
Support vector machine	119	247	42.60	2018
Artificial neural network	113	220	20.78	2019
Land use and land cover	90	184	22.49	2019
Artificial intelligence	77	149	11.17	2021
Landslide	75	161	59.05	2020
Big data	70	128	18.39	2019
Landslide susceptibility	52	106	40.37	2020
Computer vision	43	107	11.74	2021
Data mining	43	79	23.84	2017
Volunteered geographic information	41	75	23.71	2019
Random forest	40	71	24.15	2020

4 Conclusion

This study provides a bibliometric review by conducting quantitative bibliometric analyses of articles, researchers, and countries/regions. A total of 3387 essential journal articles and conference proceedings were collected from the Scopus database, with co-authorship analysis, co-occurrence analysis, and co-citation analysis employed to visualize the status and trends in the field. The study focused on integrating ML and GIS within the AEC domain. Based on the keyword co-occurrence network analysis reveals specific applications, such as landslide detection, Earth insights, and geospatial risk assessment and monitoring. Furthermore, bibliometric analysis also demonstrates its utility in identifying research trends across various other topics.

Despite the contributions provided in this study, its results should be considered in certain limitations. The scope of the research was bounded by the initial selection of keywords, which could enlarge the coverage of the studies. For example, a small number of articles are also relevant to the domains of medicine, biology, agronomy, etc. Additionally, a deeper analysis of the data, such as more detailed centrality and clustering summaries, is lacking and could be expanded upon in future research.

Reference

1. Barlish, K., Sullivan, K.: How to measure the benefits of BIM — A case study approach. Automation in Construction 24, 149-159 (2012).
2. Akinci, B., Fischer, M., Levitt, R., Carlson, A.: Formalization and automation of time-space conflict analysis. Journal of Computing in Civil Engineering 16(2), 124-134 (2002).
3. Jordan, M.I., Mitchell, T.M.: Machine learning: Trends, perspectives, and prospects. Science 349(6245), 255-260 (2015).
4. Cao, W., Liu, Q., He, Z.: Review of pavement defect detection methods. IEEE Access 8, 14531-14544 (2020).
5. Zhou, J., Love, P. E., Wang, X., Teo, K. L., Irani, Z.: A review of methods and algorithms for optimizing construction scheduling. Journal of the Operational Research Societ 64, 1091-1105 (2013).
6. Adeli, H., Wu, M.: Regularization neural network for construction cost estimation. Journal of construction engineering and management 124(1), 18-24 (1998).
7. Fang, Q., Ding, L., Luo, H., Love, P.E.D.: Computer vision for behavior-based safety in construction: A review and future directions. Advanced Engineering Informatics 46, 101260 (2021).
8. Azhar, S.: Role of visualization technologies in safety planning and management at construction jobsites. Procedia engineering 171, 215-226 (2017).
9. Fiorentini, N., Maboudi, M., Leandri, P., Losa, M., Gerke, M.: Surface motion prediction and mapping for road infrastructures management by PS-InSAR measurements and machine learning algorithms. Remote Sensing 12(23), 3976 (2020).
10. Yun, J., Ryu, K. R., & Ham, S.: Spatial analysis leveraging machine learning and GIS of socio-geographic factors affecting cost overrun occurrence in roadway projects. Automation in Construction 133, 104007 (2022).
11. Hussain, M., Tayyab, M., Ullah, K., Ullah, S., Rahman, Z. U., Zhang, J., Al-Shaibah, B.: Development of a new integrated flood resilience model using machine learning with GIS-based multi-criteria decision analysis. Urban Climate 50, 101589 (2023).
12. Maxwell, A.E., Warner, T.A., Fang, F.: Implementation of machine-learning classification in remote sensing: An applied review. International Journal of Remote Sensing 39(9), 2784-2817 (2018).
13. Iamtrakul, P., Chayphong, S., Kantavat, P., Hayashi, Y., Kijsirikul, B., Iwahori, Y.: Exploring the Spatial Effects of Built Environment on Quality of Life Related Transportation by Integrating GIS and Deep Learning Approaches. Sustainability 15(3), 2785 (2023).

14. Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., Zhang, W.: Assessing street-level urban greenery using Google Street View and a modified green view index. Urban Forestry & Urban Greening 14(3), 675-685 (2015).
15. Zupic, I., Čater, T.: Bibliometric methods in management and organization. Organizational Research Methods 18(3), 429-472 (2015).
16. Merigó, J.M., Yang, J.B.: A bibliometric analysis of operations research and management science. Omega 73, 37-48 (2017).
17. Zupic, I., Čater, T.: Bibliometric methods in management and organization. Organizational Research Methods 18(3), 429-472 (2015).
18. Yin, X., Liu, H., Chen, Y., Al-Hussein, M.: Building information modelling for off-site construction: Review and future directions. Automation in construction 101, 72-91 (2019).
19. van Eck, N.J., & Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84(2), 523-538 (2010).
20. Chen, C.M., Hu, Z., Liu, S., Tseng, H.: Emerging trends in regenerative medicine: A scientometric analysis in CiteSpace. Expert opinion on biological therapy 12(5), 593-608 (2012).
21. Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science 24(4), 265-269 (1973).
22. Zhao, A., Chen, Y., Zhang, Y., Xiao, B., Li, L., Yin, X.: Understanding Age-Friendliness in Urban Built Environments: A Bibliometric Analysis and Visualization. In ICCREM 2023. 135-148 (2023).
23. Li, L., Luan, H., Yin, X., Dou, Y., Yuan, M., Li, Z.: Understanding sustainability in off-site construction management: State of the art and future directions. Journal of Construction Engineering and Management 148(11), 03122008 (2022).
24. Lu, D., Weng, Q.: A survey of image classification methods and techniques for improving classification performance. International journal of Remote sensing 28(5), 823-870 (2007).
25. Spielman, S. E., Thill, J. C.: Social area analysis, data mining, and GIS. Computers, Environment and Urban Systems 32(2), 110-122 (2008).
26. Yao, X., Dai, F. C.: Support vector machine modeling of landslide susceptibility using a GIS: A case study. IAEG2006 793, 1-12 (2006).
27. Yue, P., Jiang, L.: BigGIS: How big data can shape next-generation GIS. IEEE In 2014 The third international conference on Agro-Geoinformatics, 1-6 (2014).
28. Grekousis, G., Manetos, P., Photis, Y. N.: Modeling urban evolution using neural networks, fuzzy logic and GIS: The case of the Athens metropolitan area. Cities 30, 193-203 (2013).
29. Singh, A.: Remote sensing and GIS applications for municipal waste management. Journal of environmental management 243, 22-29 (2019).
30. Chen, W., Zhang, S., Li, R., Shahabi, H.: Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. Science of the total environment 644, 1006-1018 (2018).
31. Yu, J., Zeng, P., Yu, Y., Yu, H., Huang, L., Zhou, D.: A Combined Convolutional Neural Network for Urban Land-Use Classification with GIS Data. Remote Sensing 14(5), 1128 (2022).
32. Pepe, M., Costantino, D., Alfio, V. S., Vozza, G., Cartellino, E.: A novel method based on deep learning, GIS and geomatics software for building a 3D city model from VHR satellite stereo imagery. ISPRS International Journal of Geo-Information 10(10), 697 (2021).
33. Choi, Y.: GeoAI: Integration of Artificial Intelligence, Machine Learning, and Deep Learning with GIS. Applied Sciences 13(6), 3895 (2023).