Indonesian Journal of Electrical Engineering and Computer Science
Vol. 37, No. 2, February 2025, pp. 1317~1327
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v37.i2.pp1317-1327

Research trends in spatial modeling of PM2.5 concentration using machine learning: a bibliometric review

Retno Tri Wahyuni[1,2], Dirman Hanafi[1], M. Razali Tomari[1], Dadang Syarif Sihabudin Sahid[3]
[1]Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia
[2]Department of Industrial Technology, Politenik Caltex Riau, Pekanbaru, Indonesia
[3]Department of Information Technology, Politenik Caltex Riau, Pekanbaru, Indonesia

Article Info

Article history:

Received Jun 5, 2024
Revised Sep 24, 2024
Accepted Sep 30, 2024

Keywords:

Bibliometric
Machine learning
PM2.5
Spatial modeling
VOSViewer

ABSTRACT

Spatial modeling is commonly used to map research variables, including particulate matter 2.5 (PM2.5) concentrations, in specific areas. The article that surveys publications on the application of machine learning in spatial modeling of PM2.5 using bibliometric methods has not been identified yet. This paper aims to analyze trends in applying machine learning in the spatial modeling of PM2.5 using bibliometric methods. The review was conducted on publications indexed in the Scopus database over the decade (2014–2023) comprising 335 articles. The analysis included co-authorship and co-occurrence using VOSviewer. From the two stages of analysis, it can be concluded that research on this topic has constantly increased over the past 10 years, with the highest productivity coming from researchers in China. This research topic is multidisciplinary, with most publications appearing in environmental science. The research also shows a very high collaboration rate of 0.98. A deeper examination of the keywords reveals the most commonly used machine learning techniques by researchers. The random forest method is the most frequently found in the analyzed documents, followed by deep learning, long short-term memory (LSTM), extreme gradient boosting (XGBoost), and ensemble model.

Corresponding Author:

Retno Tri Wahyuni
Department of Industrial Technology, Politeknik Caltex Riau
Pekanbaru, Indonesia
Email: retnotri@pcr.ac.id

1. INTRODUCTION

Particulate matter 2.5 (PM2.5) concentration describes the amount of fine particulate aerosol particles with diameters up to 2.5 microns produced by various sources, which can result in different chemical compositions and physical characteristics. Among the prevalent components found within PM2.5 are sulfates, nitrates, black carbon, and ammonium, which collectively form a significant portion of these particles [1]. PM2.5 can be generated from anthropogenic activities such as traffic emissions, industrial processes, agricultural practices, and natural sources such as dust storms, sandstorms, and wildfires.

PM2.5 is one of the dangerous pollutants whose concentration is monitored and regulated as an air quality standard by the World Health Organization (WHO), along with five other parameters such as PM10, ozone (O_3), nitrogen dioxide (NO_2), dioxide sulfur (SO_2), and carbon monoxide (CO). Several studies discuss the impact of PM2.5 on health and causes of mortality. As a tiny particulate matter, PM2.5 hurts human health. PM2.5 can penetrate the lungs and bloodstream without being filtered, causing respiratory diseases [2]–[4]. Several epidemiological studies have proven that PM2.5 increases mortality and morbidity rates [5]–[7].

Spatial modeling is widely applied in various fields, one of which can be found in the modeling concentration distribution of pollutants such as PM2.5. One background for developing spatial modeling for the distribution of pollutant concentrations in an area is the limited number of air regulatory monitoring networks (ARMN) due to the high investment and operations costs. On the other hand, several studies have indicated that pollutant concentrations exhibit small-scale spatial variations that are not adequately captured by these regulatory networks. It is often impractical to collect samples from all locations throughout a geographic area at all times [8].

In spatial modeling, many aspects have been researched and published, including those related to the methods and parameters involved in the modeling process. Publications on spatial modeling of PM2.5 using various methods have also been extensively conducted by previous researchers. Wang *et al.* [9] utilized the weather research and forecasting chemistry (WRF-Chem) method, which integrates weather prediction with atmospheric chemical processes, including aerosol and pollutant transport. Other researchers used meteorological data [8], [10], as well as satellite data such as aerosol optical depth (AOD) and normalized difference vegetation index (NDVI) to predict PM2.5 concentrations in certain regions [11]–[13]. Another group of researchers [14]–[16] applied interpolation methods in spatial modeling to predict PM2.5 values in areas lacking measurement data. Some researchers conducted spatial modeling by exploring spatial variables such as land use and topographic conditions of a region regarding the pattern of pollutant distribution using regression models [10]. On the other hand, more complex studies combine regression, interpolation, and other methods such as machine learning to generate a model capable of effectively and accurately predicting values in surrounding measurement areas, as evidenced in the study [17]–[20]. Currently, machine learning algorithms are continuously evolving to address increasingly diverse and complex challenges. One of these developments is the ensemble method, which combines multiple machine-learning models. Some studies on spatial modeling of PM2.5 present performance comparisons between single models and ensemble models, as observed in [17], [20].

Meanwhile, over the past 10 years, publications on PM2.5 in the form of article surveys or literature reviews on the topic of PM2.5 have been found in several publications. A systematic analysis of articles on a specific topic will help researchers identify research gaps and develop topics for further investigation. One of the methods increasingly used by authors is bibliometric analysis. Some of these articles include bibliometric reviews on general trends in PM2.5 research, bibliometric analysis on publications related to PM2.5 and its health impacts, and also bibliometric analysis on implementation of modeling method for PM2.5.

Meng *et al.* [21] uses bibliometric methods to offer a comprehensive overview of PM2.5 research from 1993 to 2017 by analyzing articles retrieved from the Web of Science database. Simultaneously, this study [22] examines publications to reveal collaboration patterns among countries/territories, institutions, and authors, gain deeper insights into global trends, and pinpoint research frontiers related to PM2.5 from 1997 to 2016, using data sourced from the Web of Science database.

Other articles conduct bibliometric analysis on publications related to PM2.5 and its health impacts. Study [23] investigates the association between cancer and particulate matter through related publications from 2012 to 2021 from a macroscopic perspective using the Web of Science database. Study [24] performs a bibliometric analysis of publications on PM2.5 exposure from 1992 to 2022 using the open-access tool, KNIME. Meanwhile, study [25] uses the bibliometric software CiteSpace to analyze relevant literature, covering all years up to 2020 from the Web of Science database, to gain a deeper and more structured understanding of the progress and frontiers of research related to the impact of PM2.5 on health.

Furthermore, a bibliometric analysis was performed to explore the use of machine learning in PM2.5 modeling. Jain *et al.* [26] examined the application of machine learning in air pollution modeling. This study aims to assess the current status and recent advancements in scientific research on using machine learning algorithms to address air pollution challenges. The Web of Science database was used for this study, covering the period from 1990 to 2022. The papers published on this topic were analyzed using VOSViewer and biblioshiny software.

From the multitude of publications on survey articles using bibliometrics related to PM2.5 and the use of machine learning, to the best of the author's knowledge, no researcher has systematically and specifically reviewed publications that apply machine learning in spatial modeling of PM2.5. However, research on spatial modeling, specifically for PM2.5 parameters, has significant potential due to the high health and environmental risks posed by such pollutants. Similarly, the use of machine learning, which is increasingly popular in spatial modeling, addresses challenges related to data nonlinearity and other computational issues. Thus, this paper will examine the publication trends in spatial modeling of PM2.5 concentrations employing machine learning methods through bibliometric analysis. The survey of articles on the use of machine learning in spatial modeling will assist researchers in understanding trends in the application of machine learning methods for PM2.5 modeling.

This study aims to obtain a broad understanding of the use of machine learning algorithms in the context of spatial modeling of PM2.5. To gain a comprehensive understanding of the field's development status, the authors examined the available literature from the Scopus online database. These papers were exported to the VOSviewer software for analysis. Thus, this study provides researchers with broad insights into unique research questions about spatial modeling of PM2.5, including: a) what is the annual scientific publication growth related to this topic?, b) How has the amount of research on "spatial modeling and machine learning" evolved?, c) Which journals are the most productive?, d) How are collaborations among renowned researchers conducted?, e) Which countries are involved in collaborations on different aspects of PM2.5 spatial modeling?, f) What are the key terms associated with "spatial modeling, pm2.5 and machine learning" found in the literature?

2. MATERIALS AND METHODS

2.1. Literature search strategy

Extensive searching was conducted through the Scopus database using several keywords. The first strategy employed was to search for documents using the keywords "spatial modeling" and "PM2.5", excluding “machine learning”. This initial search yielded a total of 5,970 documents. The purpose of this search was to gain a more general overview of the trends in spatial modeling of PM2.5. Subsequently, a more specific search was conducted by adding the keyword “machine learning”. This was done to assess the trend of machine learning usage in spatial modeling, which is the investigated topic of this paper. Detailed criteria for publication search are outlined in Table 1. The data retrieval date was March 5, 2024, and English was chosen as the preferred language. Initially, 335 documents were obtained. All records were retrieved in .csv format and processed using Microsoft Excel. Through Microsoft Excel, the process involved completing author keyword records, enhancing, and refining bibliography data, and handling exceptions for inaccessible documents.

Table 1. Summary of important information source retrieval

Description	Value
Search query	TS = ("spatial model" OR "spatial modelling" OR "spatial modeling" OR "spatial") AND ("pm2.5" OR "pm 2.5" OR "particulate matter 2.5") AND ("machine learning" OR "machine-learning")
Timespan	2014 to 2023
Type of document	Articles, conference papers, reviews, conference reviews, book chapter

2.2. Bibliometric analysis

According to [24], the bibliometric analysis process consists of five steps: study design, data collection, data pre-processing, analysis, visualization, and interpretation of results. In this study, we conducted a co-author and co-occurrence analysis. Co-author analysis is conducted to determine the level of collaboration among authors on papers and to identify the most productive countries of the authors. Meanwhile, co-occurrence analysis aims to identify the keywords most frequently found in the publication documents. These keywords are a crucial part of the paper to understand the core of the research topic in the publication. Co-occurrence analysis of keywords refers to two or more keywords appearing together in some document. Through keyword analysis, various critical research topics and characteristics of a research field can be revealed.

We used VOSviewer to conduct the co-author and co-occurrence analysis. VOSviewer is a freely available computer program for constructing and visualizing bibliometric maps. This tool was published by its creators in 2009 [27]. In the publication, it is explained that VOSviewer differs from most computer programs for bibliometric mapping. It is further explained that VOSviewer pays special attention to the graphical representation of bibliometric maps. The functionality of VOSviewer is very useful for displaying large bibliometric maps in an easily interpretable way. In addition to keyword analysis, this study also identifies other bibliographic data related to countries, sources, and authors. The analysis process is carried out with the assistance of Microsoft Excel.

3. RESULTS AND DISCUSSION

3.1. Bibliometric analysis result

As explained in the previous section, the document publication search strategy was conducted for the types of documents: articles, reviews, and proceeding papers. From the collected publications, totaling 335 publications, on the topics of "spatial modeling," "PM2.5," and "machine learning," an overview of the distribution of publication types can be obtained, as presented in Figure 1. The Scopus website has summarized the trend of these publication types. It can be seen that out of the total documents, 89% are articles, 6% are conference papers, 2% are review articles, 2% are conference reviews, and 1% are book chapters.

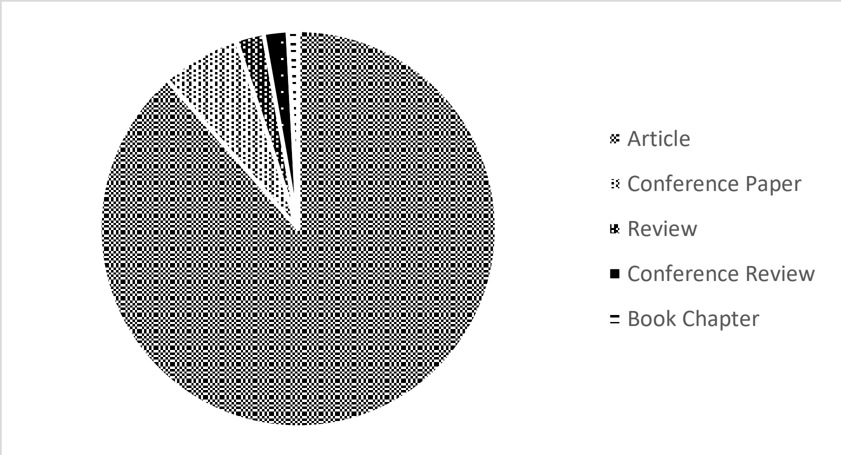

Figure 1. Type of publication of the analyzed documents

The publication trend on this topic from 2014 to 2023 can be seen in Figure 2. The graph shows a significant increase in publications on this topic, as evidenced by the growing number of publications released each year. This increase becomes more pronounced starting from 2020, peaking at 99 publications in 2022. Publications related to PM2.5 modeling are often associated with environmental issues, health, climate change, and pollution from urban and industrial activities. Research on PM2.5 from 2019 to 2022 is also strongly linked to the health impacts of the COVID-19 pandemic.

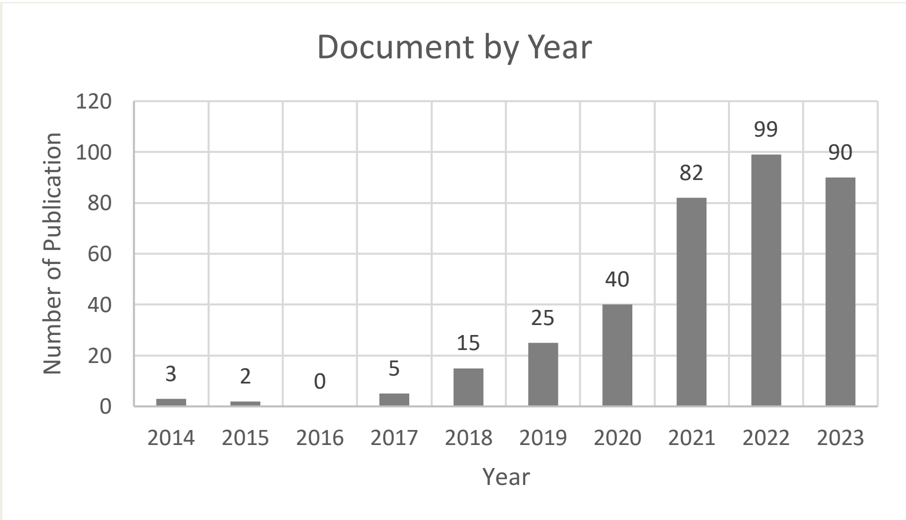

Figure 2. Number of publications on spatial modeling of PM2.5 using machine learning from 2014 to 2023

From Figure 3, it can be seen that the five subject areas contributing the most publications are "environmental science" at 34.99%, "earth and planetary sciences" at 18.82%, "computer science" at 7.47%, "Engineering" at 6.53%, and "Medicine" at 5.29%. Additionally, there are other subject areas totaling 26.6%, including "social sciences," "pharmacology, toxicology, and pharmaceutics," "chemistry," "mathematics," and others. The diversity of subject areas resulting from the search indicates that the use of machine learning in modeling the spatial distribution of PM2.5 concentrations is a multidisciplinary research field.

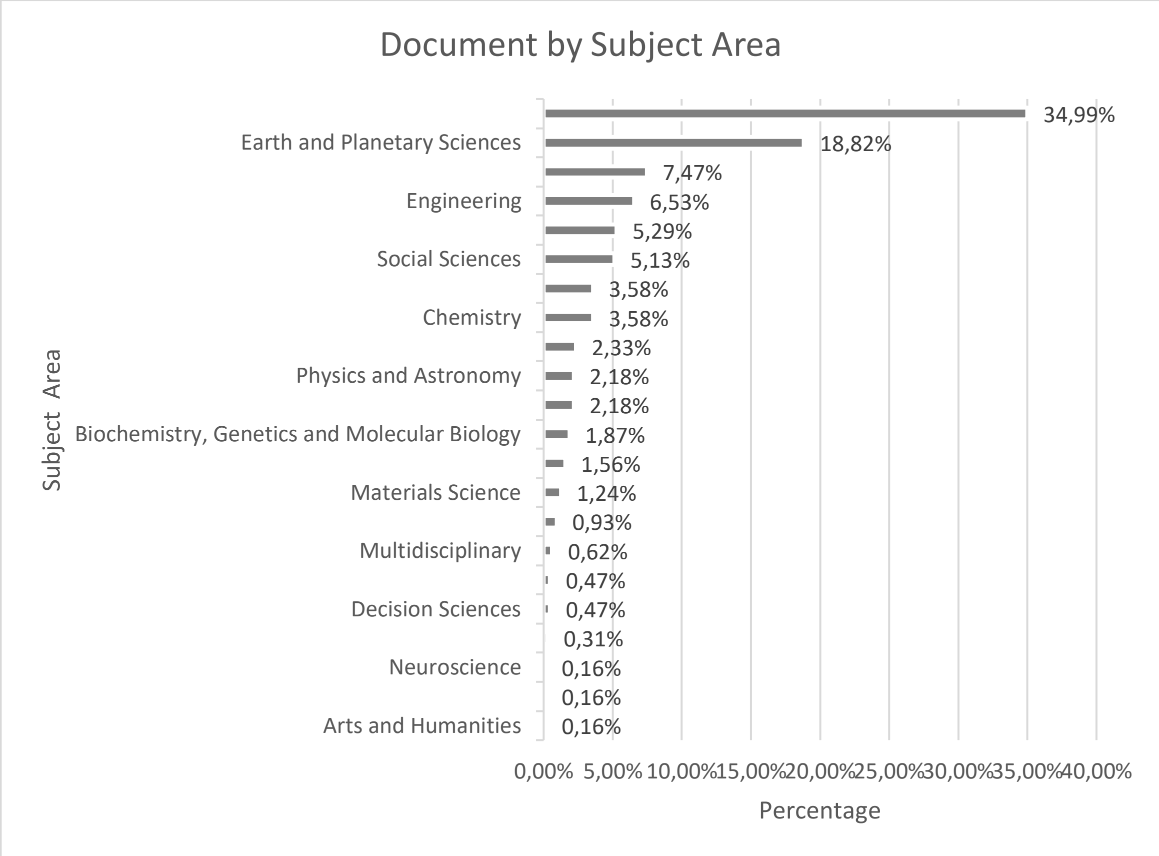

Figure 3. Subject area for publications on spatial modeling of PM2.5 using machine learning from 2014 to 2023

A summary of basic information about the analyzed documents is presented in Table 2. The table shows that the publication documents originate from various publishers, totaling 121 sources. The number of authors contributing to these publications is 1,762 individuals, with an average citation of 56 per document. The collaboration rate in these publications is calculated using a formula referenced in [28]. The value DC indicates the degree of collaboration, where N_m is the total number of collaborative research documents, and N_i is the total number of individual research documents. The DC value in the range of $0.5 < DC < 1$ indicates powerful collaboration in the writing of the papers. This figure shows that the number of collaborative research outputs exceeds those conducted individually.

$$DC = \frac{N_m}{N_m + N_s}$$

$$DC = \frac{329}{329+6} = 0.98 \tag{1}$$

Table 2. Summary of important information

Description	Results
Timespan	2014:2023
Sources (journals, books)	121
Documents	335
Keywords plus (ID)	3035
Authors' keywords (DE)	863
Average citations per document	56
Authors	1762
Authors of single-authored documents	5
Authors of multi-authored documents	1757
Single-authored documents	6
Collaboration rate (degree of collaboration)	0.98

The co-authorship analysis was also conducted based on the researchers' countries of origin, which can indicate which countries are active in research related to spatial modeling of PM2.5 concentration distribution using machine learning. From the results of the unit analysis "countries" using VosViewer, there are 51 countries of origin of researchers conducting research on this topic, with 5 countries having the highest number of researchers presented in Table 3.

Table 3. Countries of origin of researchers most active in research

No	Country	Number of document
1	China	197
2	United States	126
3	South Korea	21
4	India	20
5	United Kingdom	17

Research on spatial modeling of PM2.5 is mostly conducted in China, followed by the United States, South Korea, India, and the United Kingdom. According to the world air quality report released by IQAir [1], [29], from 2022 to 2023, three of these five countries saw an increase in their rankings for countries with the highest average PM2.5 concentrations, while two countries saw a decrease in their rankings. China ranked 25^{th} in 2022 and rose to 19^{th} in 2023; South Korea from 56^{th} to 50^{th}; India from 8^{th} to 3^{rd}; while the United Kingdom fell from 101^{st} to 112^{th} and the United States from 99^{th} to 102^{nd}. The high level of research related to PM2.5 modeling is driven by several factors, including the increasing pollution levels in these countries, rapid urbanization and industrialization, government policies and regulations, technological advancements, public awareness, and the collaboration among researchers from different countries.

In addition to these countries, there are also several countries of origin of researchers that are quite active, such as Canada, Taiwan, Hong Kong, Italy, Japan, Australia, Israel, Germany, Iran, and Sweden, with a range of research document numbers between 8 and 14 documents. Furthermore, several countries are starting to actively conduct research, such as Indonesia, Ghana, Norway, Bangladesh, Spain, and Poland.

Research on the topic of spatial modeling of PM2.5 using machine learning still has significant opportunities for exploration, especially for researchers in countries with minimal research or those yet to delve into this topic. Another consideration for why this topic is compelling to investigate is its relevance to global environmental issues such as climate change and zero net carbon emissions. Additionally, the varying characteristics of different regions such as climate conditions, geography, and emission sources are strong reasons why spatial modeling for a specific area needs to be conducted.

3.2. Co-occurrence analysis of keywords

Keywords are terms or phrases that succinctly reflect the topic of a paper. These keywords are useful in helping readers or researchers identify the content of the paper. The presence of keywords allows a paper to be more easily found in databases or search engines when someone is seeking information on a similar topic. Choosing good keywords will increase the chances of being cited and wider dissemination. Additionally, keywords help in categorizing and indexing research, making it easier for databases to group studies under the appropriate subject areas.

In co-occurrence analysis of keywords, the trends of a topic as well as the relationships between keywords across papers can be observed from the analysis of the simultaneous appearance of keywords. By using the VOSviewer application, analysis can be conducted on the co-occurrence of both authors and keywords in terms of frequency and their associations. In this analysis, some common functional terms such as pronouns and prepositions are excluded. The results of this analysis provide deeper insights into how keywords relate to one another and can help identify patterns or themes that emerge in the research.

After obtaining documents from the Scopus database and performing data pre-processing using Microsoft Excel, analysis was conducted using VOSviewer. By using co-occurrence analysis for author keywords, a total of 863 keywords were obtained. By applying a threshold of 5 documents where a keyword must appear, a total of 37 keywords were obtained. Figure 4 shows the main keyword map focused mostly on the analyzed documents. This keyword map can be divided into 6 clusters visualized with different colors. These clusters are based on the co-citation patterns among the analyzed journals. The clusters highlight distinct themes and trends within the research, illustrating how various keywords interrelate within the broader context of the study.

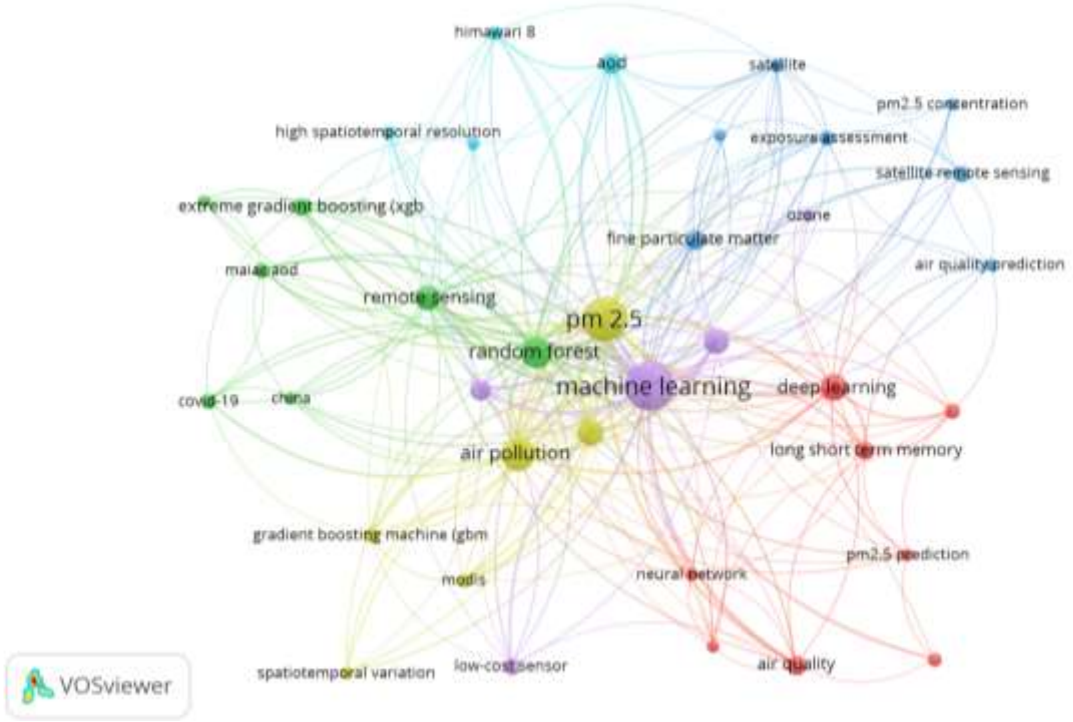

Figure 4. Network visualization

Moreover, to identify the trend of using machine learning methods in the analyzed documents, co-occurrence analysis was conducted on the author keywords unit, which indicates the types of machine learning methods. From 335 documents, a total of 770 keywords were obtained using a threshold of 1 document. Subsequently, the keywords were processed using Microsoft Excel. During the data processing, keywords that represent types of machine learning methods were selected, resulting in 82 keywords. Keywords that represent the same type of machine learning method but have different spellings were grouped into one method type. The results of this grouping show that there are 21 machine learning methods found in the documents, with the 5 most frequently occurring methods presented in Table 4.

Table 4. The most frequently found types of machine learning methods in the documents

No	Method	Number of occurrence
1	Random forest (RF)	49
2	Deep learning	33
3	Long short-term memory (LSTM)	17
4	Extreme gradient boosting (XGBoost)	13
5	Ensemble model	12

From the results of the bibliometric search, it is evident that random forest is the most frequently found machine learning method as a keyword in the analyzed documents, followed by deep learning, LSTM, XGBoost, and ensemble model. RF is an ensemble model of the bagging type that combines several decision tree models. As an ensemble method, it has the advantage of higher accuracy and tends to be more robust against overfitting. Faster computation and the requirement for less data are also reasons why the RF method is widely used by researchers. Another ensemble method that is increasingly being used for spatial modeling is XGBoost, which uses the boosting principle. Besides ensemble methods, deep learning methods, including LSTM, are also often used in solving spatial modeling, especially for their ability to handle large, complex, and unstructured data. In research published in 2023, the RF method can be seen in studies [30]–[34]. Meanwhile, in the past 5 years, the deep learning method was found in studies [35]–[38], the LSTM network method in [39]–[41], the XGBoost method in [42]–[47], and the ensemble model method, which combines several varied models, was found in [48]–[54].

In addition to the aforementioned 5 methods, there are several machine learning methods that are widely used by researchers, such as convolutional neural network (CNN), gradient boosting machine (GBM), support vector machine (SVM), neural networks, and light gradient boosting machine (LGBM). With the variety of these methods, several research papers also discuss performance comparisons of various machine learning methods in spatial modeling of PM2.5, as found in [44], [49], [52], [53].

The visualization map of the density of machine learning method usage can be seen in Figure 5. The prevalence of these five methods is evident from their positions on the map, which are located in the orange and yellow-colored areas with larger label sizes. The red areas on the map indicate keywords that previous researchers have widely used, while colors further from red indicate keywords that researchers have not extensively used.

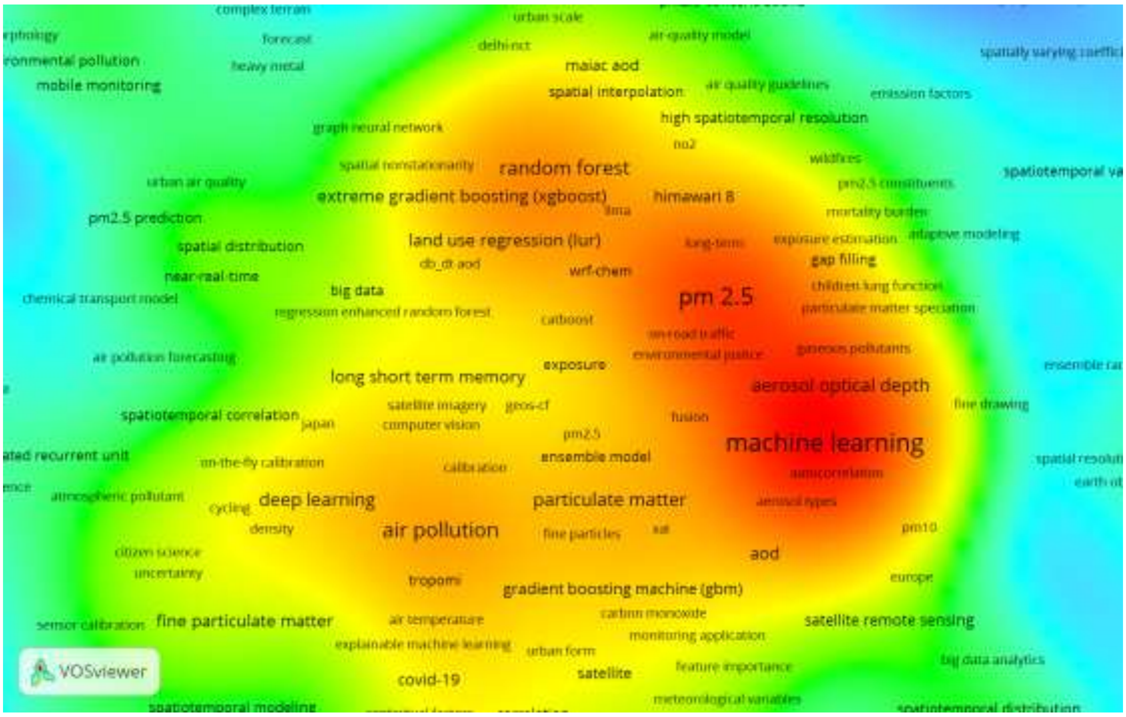

Figure 5. Density visualization

The development of machine learning methods opens opportunities for researchers to improve the performance of models that have been widely used, conduct comparative studies of existing methods, develop new methods, or investigate methods that are still underutilized by other researchers. Research can also be conducted by applying machine learning methods to case studies of specific geographical areas and comparing the results of method application across different regions. Several keywords in Figure 5 can be used to identify research gaps in existing studies. In research on spatial modeling of PM2.5, collaboration among researchers across regions and countries can also be conducted to exchange ideas regarding model transfer and factors influencing the model.

4. CONCLUSION

The results of the bibliometric study conducted indicate that research on "spatial modeling," "PM2.5," and "machine learning" from 2014 to 2023 has experienced an increasing trend, with the peak of the highest number of publications occurring in 2022. The research results show that various disciplines have been involved, such as "environmental science," "earth and planetary sciences," "computer science," "Engineering," and "Medicine," making it an interdisciplinary study area. The collaboration rate in these publications is also very high, at 0.98, indicating that most publications are written collaboratively rather than individually. Based on the results of the co-authorship analysis, research documents on spatial modeling of PM2.5 using machine learning come from various countries, with the highest number of documents being the result of researchers from China, followed by the United States, South Korea, India, and the United Kingdom. Many machine learning algorithms in studies of spatial modeling of PM2.5 have addressed modeling techniques, and the results of co-occurrence analysis conclude that the random forest method is popular among researchers, followed by deep learning, LSTM, XGBoost, and ensemble model. The development of machine learning methods opens opportunities for researchers to conduct further research such as improving model performance, comparative studies, developing new methods, or applying methods to case studies of unexplored regions. Research on this topic still has significant opportunities for exploration, especially for researchers in countries with minimal research, because the relevance of this topic to global environmental issues and its unitary nature depends on the characteristics of different regions.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Politeknik Caltex Riau and Universiti Tun Husein Onn Malaysia for their academic and financial support during the research process on spatial modeling of PM2.5, which began with the literature review presented in this paper.

REFERENCES

[1] IQ Air, "2022 World Air Quality Report," 2022. Accessed: Oct. 06, 2023. [Online]. Available: https://www.iqair.com/dl/pdf-reports/2022/2022_World_Air_Quality_Report_INT.pdf?utm_source=dynamics&utm_medium=email&utm_campaign=WAQR&utm_content=email_1#msdynmkt_trackingcontext=e0621245-2c11-45f2-beaf-d1738c6fb911.

[2] F. Wang *et al.*, "Respiratory diseases are positively associated with PM2.5 concentrations in different areas of Taiwan," *PLoS ONE*, vol. 16, no. 4 April, p. e0249694, Apr. 2021, doi: 10.1371/journal.pone.0249694.

[3] L. Yang, C. Li, and X. Tang, "The impact of PM2.5 on the host defense of respiratory system," *Frontiers in Cell and Developmental Biology*, vol. 8, Mar. 2020, doi: 10.3389/fcell.2020.00091.

[4] G. P. Bălă, R. M. Râjnoveanu, E. Tudorache, R. Motișan, and C. Oancea, "Air pollution exposure the (in)visible risk factor for respiratory diseases," *Environmental Science and Pollution Research*, vol. 28, no. 16, pp. 19615–19628, 2021, doi: 10.1007/s11356-021-13208-x.

[5] C. Liu *et al.*, "Ambient particulate air pollution and daily mortality in 652 cities," *New England Journal of Medicine*, vol. 381, no. 8, pp. 705–715, Aug. 2019, doi: 10.1056/nejmoa1817364.

[6] R. B. Hayes *et al.*, "PM2.5 air pollution and cause-specific cardiovascular disease mortality," *International Journal of Epidemiology*, vol. 49, no. 1, pp. 25–35, Feb. 2020, doi: 10.1093/ije/dyz114.

[7] V. C. Pun, F. Kazemiparkouhi, J. Manjourides, and H. H. Suh, "Long-term PM2.5 exposure and respiratory, cancer, and cardiovascular mortality in older US adults," *American Journal of Epidemiology*, vol. 186, no. 8, pp. 961–969, Oct. 2017, doi: 10.1093/aje/kwx166.

[8] S. Jain, A. A. Presto, and N. Zimmerman, "Spatial modeling of daily PM2.5, NO_2, and CO concentrations measured by a low-cost sensor network: comparison of linear, machine learning, and hybrid land use models," *Environmental Science and Technology*, vol. 55, no. 13, pp. 8631–8641, Jul. 2021, doi: 10.1021/acs.est.1c02653.

[9] P. Wang, X. Qiao, and H. Zhang, "Modeling PM2.5 and O_3 with aerosol feedbacks using WRF/Chem over the Sichuan Basin, southwestern China," *Chemosphere*, vol. 254, p. 126735, Sep. 2020, doi: 10.1016/j.chemosphere.2020.126735.

[10] L. K. Widya *et al.*, "Comparison of spatial modelling approaches on PM10 and NO_2 concentration variations: a case study in Surabaya city, Indonesia," *International Journal of Environmental Research and Public Health*, vol. 17, no. 23, pp. 1–15, Nov. 2020, doi: 10.3390/ijerph17238883.

[11] M. Unik, I. S. Sitanggang, L. Syaufina, and I. N. S. Jaya, "PM2.5 estimation using machine learning models and satellite data: a literature review," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, pp. 359–370, 2023, doi: 10.14569/IJACSA.2023.0140538.

[12] Q. Cui, F. Zhang, S. Fu, X. Wei, Y. Ma, and K. Wu, "High spatiotemporal resolution PM2.5 concentration estimation with machine learning algorithm: a case study for Wildfire in California," *Remote Sensing*, vol. 14, no. 7, p. 1635, Mar. 2022, doi: 10.3390/rs14071635.

[13] P. Gupta *et al.*, "Machine learning algorithm for estimating surface PM2.5 in Thailand," *Aerosol and Air Quality Research*, vol. 21, no. 11, p. 210105, 2021, doi: 10.4209/AAQR.210105.

[14] P. C. Chen and Y. T. Lin, "Exposure assessment of PM2.5 using smart spatial interpolation on regulatory air quality stations with clustering of densely-deployed microsensors," *Environmental Pollution*, vol. 292, p. 118401, Jan. 2022, doi: 10.1016/j.envpol.2021.118401.

[15] H. Zhang *et al.*, "Using Kriging incorporated with wind direction to investigate ground-level PM2.5 concentration," *Science of the Total Environment*, vol. 751, p. 141813, Jan. 2021, doi: 10.1016/j.scitotenv.2020.141813.

[16] K. I. Solihah, D. N. Martono, and B. Haryanto, "Identifying the best spatial interpolation method for estimating spatial distribution of PM2.5 in Jakarta," *IOP Conference Series: Earth and Environmental Science*, vol. 893, no. 1, p. 012043, Nov. 2021, doi: 10.1088/1755-1315/893/1/012043.

[17] H. Dai, G. Huang, J. Wang, H. Zeng, and F. Zhou, "Spatio-temporal characteristics of PM2.5 concentrations in China Based on multiple sources of data and LUR-GBM during 2016–2021," *International Journal of Environmental Research and Public Health*, vol. 19, no. 10, p. 6292, May 2022, doi: 10.3390/ijerph19106292.

[18] S. Yang, H. Wu, J. Chen, X. Lin, and T. Lu, "Optimization of PM2.5 estimation using landscape pattern information and land use regression model in Zhejiang, China," *Atmosphere*, vol. 9, no. 2, p. 47, Feb. 2018, doi: 10.3390/atmos9020047.

[19] C. Da Wu, Y. T. Zeng, and S. C. C. Lung, "A hybrid kriging/land-use regression model to assess PM2.5 spatial-temporal variability," *Science of the Total Environment*, vol. 645, pp. 1456–1464, Dec. 2018, doi: 10.1016/j.scitotenv.2018.07.073.

[20] P. Y. Wong *et al.*, "Using a land use regression model with machine learning to estimate ground level PM2.5," *Environmental Pollution*, vol. 277, p. 116846, May 2021, doi: 10.1016/j.envpol.2021.116846.

[21] S. Meng, B. Ding, and J. Chen, "The research and citation landscape of PM2.5 in 1993–2017: a retrospective analysis with bibliometric," in *Proceedings of the 4th International Symposium on Business Corporation and Development in South-East and South Asia under B&R Initiative (ISBCD 2019)*, Paris, France: Atlantis Press, 2020. doi: 10.2991/aebmr.k.200708.072.

[22] S. Yang *et al.*, "Trends on PM2.5 research, 1997–2016: a bibliometric study," *Environmental Science and Pollution Research*, vol. 25, no. 13, pp. 12284–12298, May 2018, doi: 10.1007/s11356-018-1723-x.

[23] X. Luo *et al.*, “A bibliometric and visualization analysis on the association between chronic exposure to fine particulate matter and cancer risk,” *Frontiers in Public Health*, vol. 10, Dec. 2022, doi: 10.3389/fpubh.2022.1039078.

[24] N. Jia, Y. Li, R. Chen, and H. Yang, “A review of global PM2.5 exposure research trends from 1992 to 2022,” *Sustainability (Switzerland)*, vol. 15, no. 13, p. 10509, Jul. 2023, doi: 10.3390/su151310509.

[25] M. Han, F. Yang, and H. Sun, “A bibliometric and visualized analysis of research progress and frontiers on health effects caused by PM2.5,” *Environmental Science and Pollution Research*, vol. 28, no. 24, pp. 30595–30612, Jun. 2021, doi: 10.1007/s11356-021-14086-z.

[26] S. Jain, N. Kaur, S. Verma, Kavita, A. S. M. S. Hosen, and S. S. Sehgal, “Use of machine learning in air pollution research: a bibliographic perspective,” *Electronics (Switzerland)*, vol. 11, no. 21, p. 3621, Nov. 2022, doi: 10.3390/electronics11213621.

[27] N. J. van Eck and L. Waltman, “Software survey: VOSviewer, a computer program for bibliometric mapping,” *Scientometrics*, vol. 84, no. 2, pp. 523–538, Aug. 2010, doi: 10.1007/s11192-009-0146-3.

[28] T. Dhar, “Bibliometric analysis of the Journal of Librarianship and Information Science during 2012-2021,” *International Journal of Information Studies and Libraries*, vol. 7, no. 2, pp. 1–9, 2022.

[29] IQ Air, “2023 World Air Quality Report,” 2023. Accessed: Aug. 09, 2024. [Online]. Available: https://www.iqair.com/dl/2023_World_Air_Quality_Report.pdf?utm_source=pdf&utm_medium=download&utm_campaign=waqr23&utm_id=waqr23&utm_term=ft#msdynmkt_trackingcontext=463261b7-ccc9-4586-8072-0359d48a99c0

[30] M. S. S. Sunder, V. A. Tikkiwal, A. Kumar, and B. Tyagi, “Unveiling the transparency of prediction models for spatial PM2.5 over Singapore: comparison of different machine learning approaches with explainable artificial intelligence,” *AI (Switzerland)*, vol. 4, no. 4, pp. 787–811, Sep. 2023, doi: 10.3390/ai4040040.

[31] X. Ma, T. Chen, R. Ge, F. Xv, C. Cui, and J. Li, “Prediction of PM2.5 concentration using spatiotemporal data with machine learning models,” *Atmosphere*, vol. 14, no. 10, p. 1517, Sep. 2023, doi: 10.3390/atmos14101517.

[32] A. Dhandapani, J. Iqbal, and R. N. Kumar, “Application of machine learning (individual vs stacking) models on MERRA-2 data to predict surface PM2.5 concentrations over India,” *Chemosphere*, vol. 340, p. 139966, Nov. 2023, doi: 10.1016/j.chemosphere.2023.139966.

[33] F. Wen, L. Liu, P. Zhang, W. Ma, and Z. Li, “Prediction of spatial distribution of PM2.5 pollution based on machine learning random forest optimized LUR model,” in *Proceedings - 2023 International Conference on Mechatronics, IoT and Industrial Informatics, ICMIII 2023*, IEEE, Jun. 2023, pp. 170–173. doi: 10.1109/ICMIII58949.2023.00037.

[34] Z. Wang, P. Chen, R. Wang, Z. An, and L. Qiu, “Estimation of PM2.5 concentrations with high spatiotemporal resolution in Beijing using the ERA5 dataset and machine learning models,” *Advances in Space Research*, vol. 71, no. 8, pp. 3150–3165, Apr. 2023, doi: 10.1016/j.asr.2022.12.016.

[35] R. Son, D. Stratoulias, H. C. Kim, and J. H. Yoon, “Estimation of surface PM2.5 concentrations from atmospheric gas species retrieved from TROPOMI using deep learning: Impacts of fire on air pollution over Thailand,” *Atmospheric Pollution Research*, vol. 14, no. 10, p. 101875, Oct. 2023, doi: 10.1016/j.apr.2023.101875.

[36] X. Yan *et al.*, “Cooperative simultaneous inversion of satellite-based real-time PM2.5 and ozone levels using an improved deep learning model with attention mechanism,” *Environmental Pollution*, vol. 327, p. 121509, Jun. 2023, doi: 10.1016/j.envpol.2023.121509.

[37] Y. Qiu *et al.*, “Regional aerosol forecasts based on deep learning and numerical weather prediction,” *npj Climate and Atmospheric Science*, vol. 6, no. 1, p. 71, Jun. 2023, doi: 10.1038/s41612-023-00397-0.

[38] H. Wang, L. Zhang, R. Wu, and Y. Cen, “Spatio-temporal fusion of meteorological factors for multi-site PM2.5 prediction: a deep learning and time-variant graph approach,” *Environmental Research*, vol. 239, p. 117286, Dec. 2023, doi: 10.1016/j.envres.2023.117286.

[39] J. Ma, Y. Ding, J. C. P. Cheng, F. Jiang, and Z. Wan, “A temporal-spatial interpolation and extrapolation method based on geographic Long Short-Term Memory neural network for PM2.5,” *Journal of Cleaner Production*, vol. 237, p. 117729, Nov. 2019, doi: 10.1016/j.jclepro.2019.117729.

[40] M. A. Bhatti, Z. Song, U. A. Bhatti, and N. Ahmad, “Predicting the impact of change in air quality patterns due to COVID-19 lockdown policies in multiple urban cities of Henan: a deep learning approach,” *Atmosphere*, vol. 14, no. 5, p. 902, May 2023, doi: 10.3390/atmos14050902.

[41] J. Ma, Y. Ding, V. J. L. Gan, C. Lin, and Z. Wan, “Spatiotemporal prediction of PM2.5 concentrations at different time granularities using IDW-BLSTM,” *IEEE Access*, vol. 7, pp. 107897–107907, 2019, doi: 10.1109/ACCESS.2019.2932445.

[42] Y. Song *et al.*, “HPO-empowered machine learning with multiple environment variables enables spatial prediction of soil heavy metals in coastal delta farmland of China,” *Computers and Electronics in Agriculture*, vol. 213, p. 108254, Oct. 2023, doi: 10.1016/j.compag.2023.108254.

[43] M. Liu, X. Luo, L. Qi, X. Liao, and C. Chen, “Simulation of the spatiotemporal distribution of PM2.5 concentration based on GTWR-XGBoost two-stage model: a case study of Chengdu Chongqing economic circle,” *Atmosphere*, vol. 14, no. 1, p. 115, Jan. 2023, doi: 10.3390/atmos14010115.

[44] L. Lin *et al.*, “Estimating PM2.5 concentrations using the machine learning RF-XGBoost model in Guanzhong Urban Agglomeration, China,” *Remote Sensing*, vol. 14, no. 20, p. 5239, Oct. 2022, doi: 10.3390/rs14205239.

[45] J. Wang *et al.*, “A full-coverage estimation of PM2.5 concentrations using a hybrid XGBoost-WD model and WRF-simulated meteorological fields in the Yangtze River Delta Urban Agglomeration, China,” *Environmental Research*, vol. 203, p. 111799, Jan. 2022, doi: 10.1016/j.envres.2021.111799.

[46] Z. Fan, Q. Zhan, C. Yang, H. Liu, and M. Bilal, “Estimating PM2.5 concentrations using spatially local XGBoost based on full-covered sara aod at the urban scale,” *Remote Sensing*, vol. 12, no. 20, pp. 1–20, Oct. 2020, doi: 10.3390/rs12203368.

[47] J. Ma, Z. Yu, Y. Qu, J. Xu, and Y. Cao, “Application of the XGBoost machine learning method in PM2.5 prediction: a case study of Shanghai,” *Aerosol and Air Quality Research*, vol. 20, no. 1, pp. 128–138, 2020, doi: 10.4209/aaqr.2019.08.0408.

[48] P. Y. Wong, H. J. Su, S. C. C. Lung, and C. Da Wu, “An ensemble mixed spatial model in estimating long-term and diurnal variations of PM2.5 in Taiwan,” *Science of the Total Environment*, vol. 866, p. 161336, Mar. 2023, doi: 10.1016/j.scitotenv.2022.161336.

[49] Q. Di *et al.*, “An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution,” *Environment International*, vol. 130, p. 104909, Sep. 2019, doi: 10.1016/j.envint.2019.104909.

[50] J. T. Kelly *et al.*, “Examining PM2.5 concentrations and exposure using multiple models,” *Environmental Research*, vol. 196, p. 110432, May 2021, doi: 10.1016/j.envres.2020.110432.

[51] Q. Fu *et al.*, “High-resolution PM2.5 concentrations estimation based on stacked ensemble learning model using multi-source satellite TOA data,” *Remote Sensing*, vol. 15, no. 23, p. 5489, Nov. 2023, doi: 10.3390/rs15235489.

[52] S. Sangpradid, T. Uttha, and I. Aroonsri, "Estimates of PM2.5 concentration based on aerosol optical thickness data using ensemble learning with support vector machine and decision tree," *Environmental Research, Engineering and Management*, vol. 79, no. 4, pp. 18–28, Dec. 2023, doi: 10.5755/j01.erem.79.4.33913.

[53] J. M. Carmona, P. Gupta, D. F. Lozano-García, A. Y. Vanoye, I. Y. Hernández-Paniagua, and A. Mendoza, "Evaluation of modis aerosol optical depth and surface data using an ensemble modeling approach to assess PM2.5 temporal and spatial distributions," *Remote Sensing*, vol. 13, no. 16, p. 3102, Aug. 2021, doi: 10.3390/rs13163102.

[54] C. C. Chen, Y. R. Wang, H. Y. Yeh, T. H. Lin, C. S. Huang, and C. F. Wu, "Estimating monthly PM2.5 concentrations from satellite remote sensing data, meteorological variables, and land use data using ensemble statistical modeling and a random forest approach," *Environmental Pollution*, vol. 291, p. 118159, Dec. 2021, doi: 10.1016/j.envpol.2021.118159.

BIOGRAPHIES OF AUTHORS

Retno Tri Wahyuni received bachelor's degree and master in Electrical Engineering from Sepuluh November Institute of Technology, Indonesia. Currently pursuing doctoral studies at the Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia. She has also been a lecturer in the Electronics Systems Engineering Technology program at Politeknik Caltex Riau since 2006. As a lecturer, her activities include teaching, research, and community service. Her current research focuses on environmental topics, both in the design and development of embedded systems and in system modeling using various methods, including machine learning. To support his knowledge of environmental topics, he collaborates with relevant institutions such as the Meteorology and Geophysics Agency and also holds a competency certification in air pollution control. She can be contacted at email: retnotri@pcr.ac.id.

Dirman Hanafi is an Associate Professor in the Faculty of Electrical and Electronic Engineering at the University of Tun Hussein Onn Malaysia (UTHM). He obtained his Ph.D. in Control and Instrumentation Engineering from the University of Technology Malaysia (UTM) in 2006, his M. Eng from the sandwich program between the Instrumentation and Control Program, Institute of Technology Bandung (ITB) Indonesia and FH Karlsruhe Germany in 1998, and his Engineering degree in Electrical from the University of Bung Hatta, Indonesia in 1994. His vast experience and expertise in the field of system identification, control systems, and artificial intelligence. He is active, and until now, he has been publishing 65 papers in conferences, journals, and books. He also published 6 IPs and received an innovation award and best paper award from a conference. He is also a member of IEEE and the Institute of Engineers Indonesia (IEI). He can be contacted at email: dirman@uthm.edu.my.

Dr. M. Razali Tomari received Ph.D. from Saitama University in 2013 and currently is Associate Professor and Deputy Dean (Research, Development and Publication) in Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia. He is a principal researcher at Research Centre for applied electromagnetics (EMCenter), registered professional engineer with Board of Engineer Malaysia and Professional Technologist with Malaysia Board of Technology (MBOT). His research activities cover deep learning for medical image analysis and smart recycling system, robot-robot interaction, low power wide area network (LPWAN) application in agriculture, smart wheelchair, and brainwave entrainment which published more than 80 manuscripts in international journals and proceedings. He can be contacted at email: mdrazali@uthm.edu.my.

Dadang Syarif Sihabudin Sahid received bachelor's degree in mathematics from Bandung Institute of Technology, Indonesia. The M.Sc. degree in Information Technology from University Teknologi Malaysia. Doctoral holder from Universitas Gadjah Mada, Indonesia. He has servised as a lecturer of Information Technology Department of Politeknik Caltex Riau, since July 1st, 2000. As a lecturer, he carries out three main activities, included teaching, research and community service. Currently, he teaches in fields of Information Systems, Application and Systems Development, IT Project Management, Information Technology Concepts, and IT strategic planning. The areas of research that are focused on include IT project management, soft computing, context aware computing, and IT planning. He received many research grants, such as collaboration grants, higher education ministry grants, including grants from abroad. He also has experience as deputy director and director at the Politeknik Caltex Riau. He can be contacted at email: dadang@pcr.ac.id.