OPEN ACCESS
Jurnal Ilmiah Ilmu Terapan Universitas Jambi
Vol. 8 No. 2. December 2024, pp. 398-408 DOI. 10.22437/jiituj.v8i2.32769
Review Article

DATA ANALYSIS AND MACHINE LEARNING APPLICATIONS IN ENVIRONMENTAL MANAGEMENT

Dilovan Asaad Majeed[1], Hawar Bahzad Ahmad[1], Ahmed Alaa Hani[1,2,*], Subhi R. M. Zeebaree[3], Saman Mohammed Abdulrahman[1], Renas Rajab Asaad[1], Amira Bibo Sallow[4]

[1] Department of Computer Science, Nawroz University, Duhok, Iraq
[2] Department of Information Technology, Duhok Technical College, Duhok Polytechnic University, Duhok, KRG-Iraq
[3] Energy Engineering Department, Technical College of Engineering, Duhok Polytechnic University, Duhok, Iraq
[4] Department of Information Technology, Duhok Technical College, Duhok Polytechnic University, Duhok, KRG-Iraq
Corresponding author email: ahmed.alaa@dpu.edu.krd

Article Info
Recieved: Apr 26, 2024
Revised: Jun 21, 2024
Accepted: Sep 20, 2024
OnlineVersion: Sep 23, 2024

Abstract
The rapid expansion of data on air contaminants and climate change, particularly concerning public health, presents both opportunities and challenges for traditional epidemiological methods. This study aims to address these challenges by exploring advanced data collection, pattern identification, and predictive modeling techniques in the context of air pollution research. The focus is leveraging data mining and computational methods to enhance the understanding of air pollution's impact on public health, specifically ozone exposure. A comprehensive review of the scientific literature was conducted, utilizing databases such as Professor, Scholar, Embl, and Nih to identify relevant studies on air pollution epidemiology. The review highlights the integration of data mining, machine learning, and spatiotemporal modeling to improve the detection, analysis, and forecasting of air pollution-related health issues. The findings reveal a growing trend in applying data mining techniques within the field of air pollution epidemiology. Advanced methods, such as spatiotemporal analysis and geographic data mining, enable more precise tracking and forecasting of pollution-related health risks. Continuous advancements in artificial intelligence and the development of more sophisticated sensors and data storage technologies are enhancing the accuracy and reliability of air quality monitoring and public health predictions. This study highlights the transformative potential of integrating data mining and AI techniques into air pollution epidemiology. Exploring emerging technologies like spatiotemporal mining and next-generation sensors paves the way for more accurate, timely, and scalable solutions to monitor air quality and predict its impact on public health, opening new avenues for research and policy interventions.

Keywords: Air Pollution Epidemiology, Data mining, Machine Learning, Predictive Modelling

INTRODUCTION

Since the cost of using satellites to measure airborne contaminants has decreased and more natural and therapeutic data is available, the choice of contamination datasets has expanded. Mainstream epidemiological and ecological wellness models are challenged by such datasets, which frequently are defined by a high range of samples and various points of information with different levels of reliance (Bobb et al., 2015; Kusuma, 2020; Gulsen, & Yalcin, 2024; Raman et al., 2024). New analytical methods are becoming more and more necessary to improve our comprehension of such complex information. This has led to the development of a new branch in the field of ecological health called "data mining processes," which focuses on the effective and efficient processing of large amounts of current data in pollution epidemiological (Villeneuve, & Goldberg, 2020; Hajat et al., 2021; Kirwa et al., 2021; Sari, Omeiza, & Mwakifuna, 2023; Fitriana & Waswa, 2024; Zakiyah, Boonma, & Collado, 2024). Finding patterns, obtaining useful information, and forecasting the course of unknown perhaps upcoming developments are all part of mining data, an algorithmic method often used to assess large datasets. Sophisticated machines, statistical analysis, mathematical learning algorithms, other database systems are just a few of the computer fields it encompasses. Before executing the mining procedure, the info mining technique may involve preprocessing steps. To show the analysis results in a comprehensible and straightforward way, an additional processing stage frequently used (Al-Yasiri & Szabo, 2021; Suwarni, 2021; Wankhade, Rao, & Kulkarni, 2022; Saputro et al., 2023; Yohanie et al., 2023; Asmororini, Kinda, & Sen, 2024; Asrial et al., 2024; Habibi, Jiyane, & Ozsen, 2024). This research restricts its scope to basic data analysis techniques that have been published in the academic literature on the epidemiology of air pollution and explicitly applied to this field.

Techniques in the field of data mining fall into four distinct subcategories, which are broadly divided into forecasting and the discovery: 1) Aggregation; 2) Mining correlation rules; 3) Inference or categorization; and 4) Outlier/Anomaly Diagnosis (Plotnikova, Dumas, & Milani, 2020; Al-Hashedi & Magalingam, 2021; Khan & Shaheen, 2023). Finding patterns, extracting useful information, and forecasting the course of unknown or future events are all part of data mining, a computer technique often used to assess large datasets as shown in figure 1. Clever machines, statistical analysis, mathematical neural networks for learning, etc the use of databases are just a few of the computer fields it encompasses (Okewu et al., 2021; Sarker, 2021; Aggarwal et al., 2022; Baah, Konovalov, & Tenzin, 2024; Fernande, Sridharan, & Kuandee, 2024; Qairunisa, Daningsih, & Candramila, 2024). Before executing the mining algorithm, the data mining technique may include pretreatment steps. To show the analysis results in a comprehensible and straightforward way, a post-processing stage is often used. This research restricts its scope to basic data analysis techniques that have been discussed in the academic literature on the epidemiology of air pollution and explicitly applied to this field.

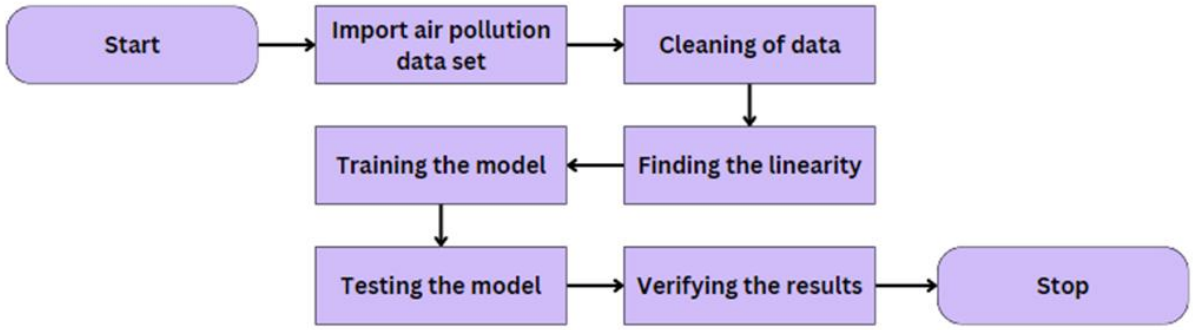

Figure 1. Flowchart of the system

Algorithm in mining data fall into four subcategories, which are broadly divided into prediction and the discovery: 1) Clustering; 2) Mining association rules; 3) Regression and classifications; and 4) To Outlier/Anomaly Detection. The fundamental elements of data mining methods have made it easier to explore some of the most interesting and current data analysis topics, such graph and spatial data mining. The surprising thing is that no previous study has looked at how widely data mining methods are used in the field of epidemiology of air pollution. In order to close this gap, we will investigate if and under what circumstances data mining methods have been used in the area of air pollution epidemiological. By showcasing previously published use cases in their particular field and adjacent

fields, the goal is to increase domain experts' comprehension of the intriguing possibilities of data mining techniques and encourage them to investigate new research avenues.

LITERATURE REVIEW

Effectively employing data mining solutions requires users to thoughtfully evaluate and define their objectives. The chosen goal plays a crucial role in determining the appropriate learning algorithm paradigm. Users interested in information discovery may favor clustering or association mining techniques, which are well-suited for unveiling hidden groupings or relationships among key variables. On the other hand, those aiming to build a prediction model, such as identifying samples with poor air quality or predicting a real-valued result like the air quality index, will pursue a different strategy.

Both knowledge discovery and prediction paradigms offer a diverse range of algorithms, presenting users with the challenge of selecting the most suitable method within each paradigm. Factors to consider include the quantity of available data and the complexity of the problem at hand. For example, addressing complex non-linear classification issues may necessitate advanced algorithms like deep artificial neural networks, which perform well with large datasets but come with considerations of storage, memory, and training time. Artificial neural networks, a powerful class of learning algorithms inspired by biological information processing, have a long history in pattern recognition. Recent applications showcase advancements in complex network topologies such as convolutional and recurrent networks, particularly successful in deep learning (figure 2). However, the traditional multilayer perceptron, a feedforward network with an input layer, hidden layers, and an output layer, remains a prevalent design. Training often involves gradient descent and backpropagation, adjusting network weights through multiple cycles, proving effective despite the non-convex nature of optimization.

Support vector machines (SVMs) emerge as a potent technique for regression and classification challenges, especially in determining maximum margin hyperplanes (Ahsaan et al., 2022; Rodriguez-Perez & Bajorath, 2022; Guido et al., 2024). Non-linear SVMs use user-specified kernels to map input to a higher-dimensional space, implicitly discovering a non-linear decision boundary. SVM's model generation is a convex optimization problem, ensuring that any local minimum is a global minimum.

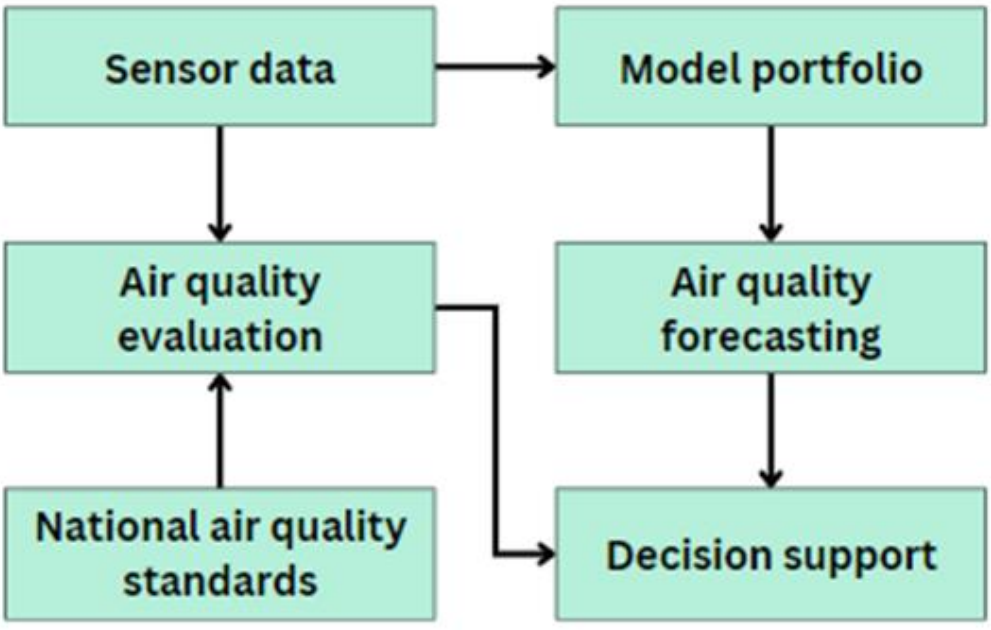

Figure 2. Approach of the System

Hierarchical clustering, a distance-based segmentation, establishes hierarchies within clusters through divisive or agglomerative approaches. Divisive clustering begins with each instance forming its own cluster, merging iteratively, while agglomerative clustering starts with all instances in one large cluster, recursively dividing to create smaller collections. This method effectively visualizes groups at various granularities within the clusters.

RESEARCH METHOD

When faced with intricate datasets containing numerous variables and samples, the application of data mining methods proves highly beneficial. These methods are essential for gaining insights into high-dimensional issues, overcoming limitations often encountered by standard statistical approaches in knowledge discovery. Simultaneously, machine learning algorithms excel at generating reliable predictor functions from complex, high-dimensional data. In contrast, traditional statistical and mathematical approaches, such as regression, may present challenges and be susceptible to errors due to their inherent assumptions.

ALGORITHM:
STEP 1: Gather relevant literature and research papers on data mining and machine learning applications in air pollution epidemiology.
STEP 2: Identify common data sources used in air pollution epidemiology, such as environmental monitoring stations, satellite imagery, and health records.
STEP 3: Review various data mining and machine learning techniques applied to air pollution epidemiology, including regression models, classification algorithms, and clustering methods.
STEP 4: Analyze the performance of different machine learning algorithms in predicting air pollution levels and their impacts on public health outcomes.
STEP 5: Identify challenges and limitations in existing studies, such as data availability, model interpretability, and generalizability across different regions.
STEP 6: Evaluate the potential of advanced machine learning techniques, such as deep learning and ensemble methods, in improving air pollution prediction models.
STEP 7: Assess the impact of air pollution on various health outcomes, including respiratory diseases, cardiovascular problems, and mortality rates.
STEP 8: Explore opportunities for integrating spatial and temporal data into machine learning models to capture complex relationships between air quality and health.
STEP 9: Synthesize findings from the systematic review to provide insights into future research directions and practical applications in air pollution epidemiology.
STEP 10: Publish the systematic review to contribute to the advancement of knowledge in the field and inform policy-making efforts aimed at reducing air pollution and protecting public health.

To maintain a systematic and organized approach, we followed the PRISMA guidelines and implemented Kitchenham's methodologies during the execution of this survey (Bartova, B., Bina, V., & Vachova, 2022; Belle & Zhao, 2023; Hernandez Aros et al., 2024). The subsequent sections offer detailed information on the survey's methodology. Additionally, inspiration for the survey design was drawn from a previous study on dengue illness monitoring, contributing to a robust and structured framework for this investigation. Research Questions: This survey focuses on addressing several key research questions to comprehensively explore the landscape of data mining applications in air pollution epidemiology. The primary research inquiries include:

R1: To what extent has data mining been utilized in the realm of air pollution epidemiology?
R2: Are there specific focal points or concentrated areas within the broader field of air pollution epidemiology where data mining research is particularly active?
R3: In which sub-fields or specialized areas of air pollution epidemiology has data mining been effectively applied?
R4: What specific data mining methods have been employed in the context of air pollution epidemiology?
R5: What are the existing limitations and challenges associated with the current body of work in this field?
R6: What unexplored and potentially promising directions exist for future research endeavors in the intersection of data mining and air pollution epidemiology?

Regarding R1, our investigation delved into the epidemiological literature to identify studies leveraging data mining methods. While we imposed no temporal restrictions, it became evident that the active period in this domain is relatively brief. Notably, there is a discernible trend towards increased

frequency as awareness of the benefits of data mining grows, coupled with enhanced accessibility to technological resources. Following R1, R2 sought to ascertain whether the distribution of current research is uniform across nations and academic institutions globally or if certain countries or organizations exhibit a more pronounced emphasis on this particular field of study.

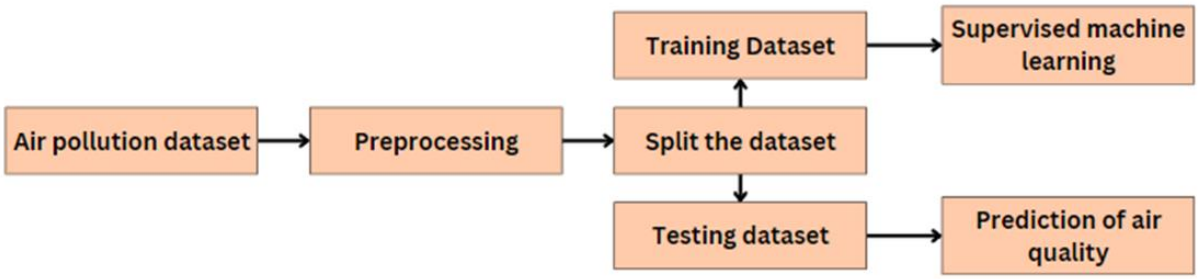

Figure 3. Workflow Methodology of the system

Addressing R3, our examination of recognized articles aimed to logically categorize epidemiological research according to distinct fields of application. Employing data mining, our methodology identified three discernible types of epidemiological research related to atmospheric pollution in the literature. In relation to R4, we scrutinized the paradigms and techniques employed in the literature on air pollution epidemiology (figure 3). Our analysis revealed four distinct categories of techniques that have been utilized. For research question R5, our focus was on evaluating whether there were limitations in the data and/or mining techniques concerning the specified objectives. Our scrutiny particularly emphasized the data used, the algorithms applied, and the procedures employed to assess these approaches, drawing upon our experiences in data mining.

Data extraction from each selected article involved gathering information such as the source and full reference, a concise overview of the study's objectives, the specific air pollutants under investigation, the data mining technique employed to achieve the study's goals, and a brief summary of the study's findings. CB and MSMJ were responsible for conducting the data extraction, while AOV and OZ handled the validation process. Any discrepancies were resolved through collaborative discussion and consensus. Once the fundamental data from the articles was tabulated, data synthesis was undertaken.

Environmental setting, This section encompasses identified areas of interest categorized into three groups: general, outdoor, and indoor. Indoor studies involve the examination of air pollution in confined spaces, such as assessing pollutants in homes or workplaces. In contrast, outdoor studies concentrate on air pollution in open environments, involving measurements at specific intersections or examinations of contaminant dispersion across predetermined areas. The outdoor category is further segmented into rural, metropolitan, and urban settings; however, for the purposes of this discussion and the limited focus of our research, we emphasize a higher level of abstraction.

Three main study aims were recognized from the collected publications: forecasting and prediction, source identification, and hypothesis formulation. The majority of the papers we found were about forecasting or estimating pollution levels based on different pollutant and climate characteristics. Three primary situations were examined in these investigations: a) projecting future pollution levels at a particular location based on data; b) projecting present pollution levels at a specific location based on data from the area; and c) projecting the dispersion of pollutants or the geographic spread of air quality.

Research that aimed to classify human diseases according to assessments of air quality or forecast increases in hospitalizations or illnesses by using pollution and climatic data were closely related to each other. Under the source identifying category, studies attempted to use a variety of pollutant and weather factors to establish a connection between a specific rise or decrease in air quality and the source of the pollution. A significant amount of the literature produced conjectures. These research, which took into account factors like the evolution of air pollution in a particular place, its spread across a region or internationally, indicators for wellness variables, and more, used data mining techniques to unearth hidden connections within the large dataset. Next, new hypotheses were developed and tested using the relationships that had been found. A correlation that is found, for example, may suggest that some chemical composition (X, Y, Z) is linked to an increased number of ER visits at a particular healthcare facility, or that certain weather conditions (W, R) in conjunction with

increased maritime traffic result in worsened readings of the air quality index. These correlations might motivate focused investigations to delve further into the found connections.

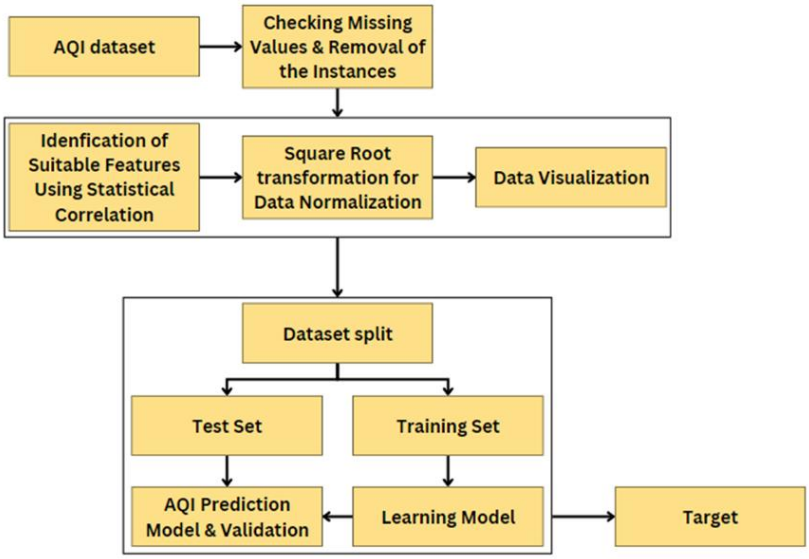

Figure 4. Implementation of the System

RESULTS AND DISCUSSION

The following is a summary statistics compilation that shows how many articles were found using our search method. Four hundred items came up in the first search, and one more was suggested for evaluation. After the preliminary screening and eligibility evaluation, 47 articles were considered appropriate for this survey.

Regional and temporal overview 18 of the studies came from Germany and the UK, 16 from the USA, 10 from China, and 4 from other Asian countries, according to our analysis of the research. The publications were published between October 20, 2017, and 2000. We credit this trend to better data ownership, more powerful processing, and more people outside the data mining industry being aware of and able to utilize data mining technology (table 1). The data mining software Weka, which enables users to directly apply data mining methods to their data using user-friendly graphical or Java interfaces, is one noteworthy technology supporting this trend.

Table 1. System Feature and Their Description, Benefits and Challenges

Feature	Description	Benefits	Challenges
Data Sources:	Utilizes diverse data sources like air quality monitoring stations, satellite imagery, meteorological data, population demographics, and health records.	Provides comprehensive information about air pollution exposure, health outcomes, and potential influencing factors.	Data quality and availability inconsistencies across different sources, potential privacy concerns regarding health records.
Data Mining:	Performs exploratory analysis to identify patterns and trends in air pollution and health data, generating hypotheses for further investigation.	Discovers hidden relationships and potential risks not readily apparent through traditional methods.	Requires careful interpretation and validation to avoid overfitting and spurious correlations.
Machine Learning:	Implements algorithms to predict air pollution levels, assess exposure risks, estimate health impacts, and identify vulnerable populations.	Enables personalized risk assessment, proactive interventions, and targeted resource allocation.	Model complexity and interpretability challenges, potential biases in algorithms and data, need for robust evaluation and

Feature	Description	Benefits	Challenges
			validation.
Spatial Analysis:	Integrates geographical information systems (GIS) and spatial statistics to examine the spatial distribution of air pollution and its association with health outcomes.	Identifies hotspots, vulnerable areas, and potential pollution sources, informing targeted interventions.	Requires accurate geospatial data and advanced computational resources for complex spatial analyses.
Temporal Analysis:	Analyzes time series data to understand trends, seasonality, and short-term fluctuations in air pollution and their impact on health.	Provides insights into potential triggers and patterns of health effects informing preventive measures.	Challenges in handling missing data. outliers. and long-term dependencies in time series data.

We discovered that association extractive industries, regression analysis, clustering, and classification approaches were used in our research. Classification and regression techniques were related to forecasting goals, whereas correlation mining and clustering techniques were usually tied to the creation of hypotheses and source allocation. Regression and classification, which both concentrate on producing numerical predictions, were the most often used data mining approaches, accounting for 59% of the study. In 26% of the research, clustering techniques were used, and in 15% of the papers, association mining. These investigations focus on understanding the interaction between pollution, meteorological factors, and various pollutants in the air. The primary goal is to establish connections between specific airborne pollutants and their potential sources, such as industrial areas, regions, and significant intersections. The research primarily emphasizes outdoor and indoor air quality, employing principal component analysis to analyze the relative contributions of meteorological conditions, transportation, fuel-fired equipment, and industry to air pollution. To achieve source apportionment, researchers combine correlation analysis with clustering-based algorithms.

Numerous recurring challenges are evident in the examined publications, mainly revolving around the concept of data. Many studies rely on data collected from limited locations and time frames, hindering the generalization of results to other settings. Localized data poses a notable challenge for prediction models. Preprocessing real-world data involves merging sources, noise elimination, and effective organization, impacting the efficacy of models. The lack of comprehensive preprocessing strategies for epidemiological datasets on air pollution in existing literature is apparent. Social media data, despite its volume, requires screening due to unrelated content, necessitating novel approaches like feature extraction.

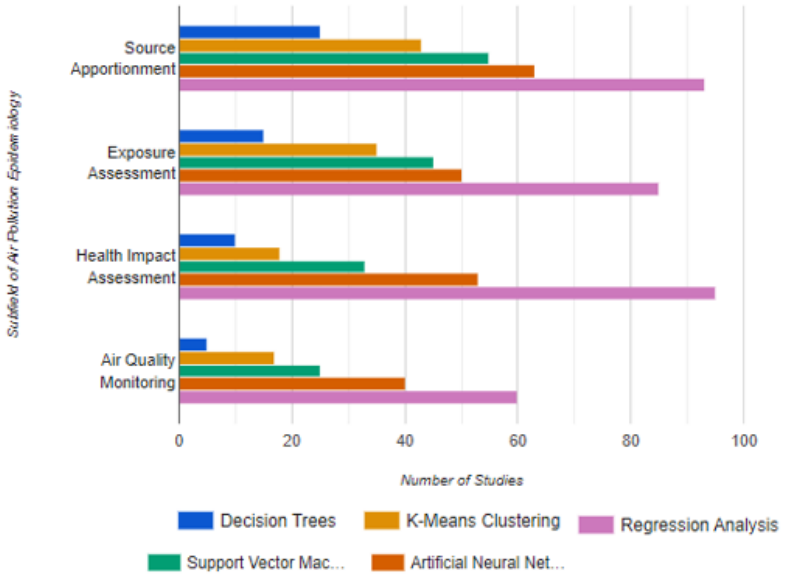

Figure 5. Data Mining and ML Techniques in Air Pollution Epidemiology of the System Application

Understanding data limitations is crucial, as issues with representativeness and granularity can restrict insights. Some studies overlook the cumulative nature of health consequences in their data, emphasizing the need for comprehensive consideration in future research. Deploying data mining techniques introduces challenges related to algorithmic parameters, requiring careful selection of clusters and metrics. The literature may lack appropriate assessment measures in certain situations, necessitating the development of novel metrics. Improving the interpretability of predictions from complex algorithms could benefit the health professions community.

CONCLUSION

The analyzed publications reveal several persistent challenges, primarily centered around the concept of data. Many studies depend on data collected from limited locations and time frames, posing difficulties in generalizing results to diverse settings. The use of localized data presents a significant obstacle for prediction models. The preprocessing of real-world data involves tasks such as merging sources, noise elimination, and effective organization, impacting the efficiency of models. Notably, there is a noticeable absence of comprehensive preprocessing strategies for epidemiological datasets on air pollution in existing literature. Social media data, despite its abundance, necessitates screening due to irrelevant content, prompting the exploration of novel approaches like feature extraction.

Understanding the limitations of data is crucial, as issues related to representativeness and granularity can constrain insights. Some studies overlook the cumulative nature of health consequences in their data, underscoring the importance of comprehensive consideration in future research. The deployment of data mining techniques introduces challenges related to algorithmic parameters, demanding careful selection of clusters and metrics. In certain situations, the literature may lack appropriate assessment measures, indicating the need for the development of novel metrics. Enhancing the interpretability of predictions from complex algorithms could prove beneficial for the health professions community.

The analyzed publications notably lack discussions about application-related options. While algorithms are well-described, clarity regarding other design and implementation choices is lacking. Many publications focus on findings from a single data mining method, overlooking details about other considered algorithms and their assessments. Providing information on the data mining packages used and software implementation would enhance the applicability of the findings to the field of air pollution epidemiology.

ACKNOWLEDGMENTS

Thank you to all colleagues who have helped, so that this research can be carried out and completed.

AUTHOR CONTRIBUTIONS

Author 1-2 creates articles and creates instruments and is responsible for research, author 3-4 Analyzes research data that has been collected, author 5-7 assists in research data analysis, instrument validation and input research data.

CONFLICTS OF INTEREST

The author(s) declare no conflict of interest.

REFERENCES

Aggarwal, K., Mijwil, M. M., Al-Mistarehi, A. H., Alomari, S., Gok, M., Alaabdin, A. M. Z., & Abdulrhman, S. H. (2022). Has the future started? The current growth of artificial intelligence, machine learning, and deep learning. *Iraqi Journal for Computer Science and Mathematics*, *3*(1), 115-123. https://doi.org/10.52866/ijcsm.2022.01.01.013

Ahsaan, S. U., Kaur, H., Mourya, A. K., & Naaz, S. (2022). A hybrid support vector machine algorithm for big data heterogeneity using machine learning. *Symmetry*, *14*(11), 2344. https://doi.org/10.3390/sym14112344

Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, *40*, 100402. https://doi.org/10.1016/j.cosrev.2021.100402

Al-Yasiri, Q., & Szabo, M. (2021). Incorporation of phase change materials into building envelope for thermal comfort and energy saving: A comprehensive analysis. *Journal of Building engineering*, *36*, 102122. https://doi.org/10.1016/j.jobe.2020.102122

Asmororini, E., Kinda, J., & Sen, B. (2024). Innovation Learning Geography with ArcGIS Online: The Impact to Skills Collaborative and Achievement Student School Upper Intermediate. *Journal of Educational Technology and Learning Creativity*, *2*(1), 1-12. https://doi.org/10.37251/jetlc.v2i1.969

Asrial, A., Syahrial, S., Kurniawan, D. A., Putri, F. I., Perdana, R., Rahmi, R., Susbiyanto, S., & Aldila, F. T. (2024). E-Assessment for Character Evaluation in Elementary Schools. *Qubahan Academic Journal*, *4*(3), 806-822. https://doi.org/10.48161/qaj.v4n3a595

Baah, R., Konovalov, O., & Tenzin, S. (2024). Effectiveness of e-assessment in science learning: Improving the quality and efficiency of assessment in the digital era. *Integrated Science Education Journal*, *5*(2), 74-81. https://doi.org/10.37251/isej.v5i2.960.

Bartova, B., Bina, V., & Vachova, L. (2022). A PRISMA-driven systematic review of data mining methods used for defects detection and classification in the manufacturing industry. *Production*, *32*, e20210097. http://dx.doi.org/10.1590/0103-6513.20210097

Belle, A. B., & Zhao, Y. (2023). Evidence-based decision-making: On the use of systematicity cases to check the compliance of reviews with reporting guidelines such as PRISMA 2020. *Expert Systems with Applications*, *217*, 119569. https://doi.org/10.1016/j.eswa.2023.119569

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., ... & Coull, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, *16*(3), 493-508. https://doi.org/10.1093/biostatistics/kxu058

Fernande, R., Sridharan, V., & Kuandee, W. (2024). Innovation learning with POE: Improve understanding student to equality square. *Journal of Educational Technology and Learning Creativity*, *2*(1), 20-28. https://doi.org/10.37251/jetlc.v2i1.977.

Fitriana, H., & Waswa, A. N. (2024). The influence of a realistic mathematics education approach on students' mathematical problem solving ability. *Interval: Indonesian Journal of Mathematical Education*, *2*(1), 29-35. https://doi.org/10.37251/ijome.v2i1.979.

Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024). An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. *Information*, *15*(4), 235. https://doi.org/10.3390/info15040235

Gulsen, M., & Yalcin, S. S. (2024). Fostering tomorrow: uniting artificial intelligence and social pediatrics for comprehensive child well-being. *Turkish archives of pediatrics*, *59*(4), 345. https://doi.org/10.5152/TurkArchPediatr.2024.24076

Habibi, M. W., Jiyane, L., & Ozsen, Z. (2024). Learning Revolution: The Positive Impact of Computer Simulations on Science Achievement in Madrasah Ibtidaiyah. *Journal of Educational Technology and Learning Creativity*, *2*(1), 13-19. https://doi.org/10.37251/jetlc.v2i1.976

Hajat, A., MacLehose, R. F., Rosofsky, A., Walker, K. D., & Clougherty, J. E. (2021). Confounding by socioeconomic status in epidemiological studies of air pollution and health: challenges and opportunities. *Environmental health perspectives*, *129*(6), 065001. https://doi.org/10.1289/ehp7980

Hernandez Aros, L., Bustamante Molano, L. X., Gutierrez-Portela, F., Moreno Hernandez, J. J., & Rodríguez Barrero, M. S. (2024). Financial fraud detection through the application of machine learning techniques: a literature review. *Humanities and Social Sciences Communications*, *11*(1), 1-22. https://doi.org/10.1057/s41599-024-03606-0

Khan, S., & Shaheen, M. (2023). From data mining to wisdom mining. *Journal of Information Science*, *49*(4), 952-975. https://doi.org/10.1177/01655515211030872

Kirwa, K., Szpiro, A. A., Sheppard, L., Sampson, P. D., Wang, M., Keller, J. P., ... & Kaufman, J. D. (2021). Fine-Scale air pollution models for epidemiologic research: insights from approaches developed in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Current environmental health reports*, *8*(2), 113-126. https://doi.org/10.1007/s40572-021-00310-y

Kusuma, R. S. (2020). Improving students' basic asking skills by using the discovery learning model. *Tekno - Pedagogi : Jurnal Teknologi Pendidikan*, *10*(2), 8-13. https://doi.org/10.22437/teknopedagogi.v10i2.32743.

Okewu, E., Adewole, P., Misra, S., Maskeliunas, R., & Damasevicius, R. (2021). Artificial neural networks for educational data mining in higher education: A systematic literature review. *Applied Artificial Intelligence*, *35*(13), 983-1021. https://doi.org/10.1080/08839514.2021.1922847

Plotnikova, V., Dumas, M., & Milani, F. (2020). Adaptations of data mining methodologies: A systematic literature review. *PeerJ Computer Science*, *6*, e267. https://doi.org/10.7717/peerj-cs.267/supp-1

Qairunisa, M. R., Daningsih, E., & Candramila, W. (2024). Development of electronic pocketbook media on plant-like protist for class x senior high school. *Integrated Science Education Journal*, *5*(3), 142-153. https://doi.org/10.37251/isej.v5i3.1067.

Raman, R., Leal Filho, W., Martin, H., Ray, S., Das, D., & Nedungadi, P. (2024). Exploring Sustainable Development Goal Research Trajectories in Small Island Developing States. *Sustainability*, *16*(17), 7463. https://doi.org/10.3390/su16177463

Rodriguez-Perez, R., & Bajorath, J. (2022). Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery. *Journal of Computer-Aided Molecular Design*, *36*(5), 355-362. https://doi.org/10.1007/s10822-022-00442-9

Saputro, H. D., Rustaminezhad, M. A., Amosa, A. A., & Jamebozorg, Z. (2023). Development of e-learning media using adobe flash program in a contextual learning model to improve students' learning outcomes in junior high school geographical research steps materials. *Journal of Educational Technology and Learning Creativity*, *1*(1), 25-32. https://doi.org/10.37251/jetlc.v1i1.621.

Sari, R., Omeiza, I. I., & Mwakifuna, M. A. (2023). The influence of number dice games in improving early childhood mathematical logic in early childhood education. *Interval: Indonesian Journal of Mathematical Education*, *1*(2), 61-66. https://doi.org/10.37251/ijome.v1i2.776.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Suwarni, R. (2021). Analysis the process of observing class iv students in thematic learning in primary schools. *Tekno - Pedagogi : Jurnal Teknologi Pendidikan*, *11*(1), 26-32. https://doi.org/10.22437/teknopedagogi.v11i1.32717.

Villeneuve, P. J., & Goldberg, M. S. (2020). Methodological considerations for epidemiological studies of air pollution and the SARS and COVID-19 coronavirus outbreaks. *Environmental health perspectives*, *128*(9), 095001. https://doi.org/10.1289/ehp7411

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731-5780. https://doi.org/10.1007/s10462-022-10144-1

Yohanie, D. D., Botchway, G. A., Nkhwalume, A. A., & Arrazaki, M. (2023). Thinking process of mathematics education students in problem solving proof. *Interval: Indonesian Journal of Mathematical Education*, *1*(1), 24-29. https://doi.org/10.37251/ijome.v1i1.611.

Zakiyah, Z., Boonma, K., & Collado, R. (2024). Physics learning innovation: Song and animation-based media as a learning solution for mirrors and lenses for junior high school students. *Journal of Educational Technology and Learning Creativity*, *2*(2), 54-62. https://doi.org/10.37251/jetlc.v2i2.1062.