Nationwide Air Quality Analysis

Malhar Kakade, Swanand Mahajan, Avishkar Mulay, Sumit Chougale, Shubham Choramle

School of Computer Engineering and Technology, MIT WPU University

Address: MIT World Peace University, Sr No 124, Ex Serviceman Colony, Paud Road, Kothrud, Pune-411038

swanandmahajan12@gmail.com

avishkar9122@gmail.com

malharkakade942@gmail.com

shubham.choramle1330@gmail.com

stchougale09@gmail.com

***Abstract*—Monitoring air quality is essential to comprehending environmental pollution and how it affects public health. We analyze data on air quality that was gathered from 2010 to 2023 from 453 Indian cities in this study. The Central Control Room for Air Quality Management is the source of the dataset, which offers detailed information on a number of pollutants, including PM2.5, PM10, NO2, SO2, CO, and ozone. We conduct a thorough analysis of the trends in air quality, their spatial distribution, and seasonal variations, concentrating on a subset of 17 cities from 5 states. To extract insights from the dataset, our methodology consists of statistical modeling techniques, exploratory data analysis, and data preprocessing. The findings show notable differences in the quality of the air in various states and cities, with potential health effects. Policymakers, researchers, and environmental advocates can benefit greatly from this study's insightful analysis of India's air pollution patterns.**

***Keywords—Air Quality, Big Data, Statistical Modeling, Pollutants.*.**

I. Introduction

Worldwide, air pollution is a serious threat to both public health and the environment, having a negative impact on cardiovascular disease, respiratory health, and general wellbeing. The problem of managing air quality has grown more urgent in nations like India that are rapidly becoming more urbanized and industrialized. The Central Control Room for Air Quality Management was established by the Indian government in order to gather, process, and distribute data on air quality from monitoring stations throughout the nation. This move was made in response to the government's recognition of the significance of tracking and managing air pollution.

An extensive analysis of air quality data gathered from 453 Indian cities between 2010 and 2023 is presented in this research paper. Comprehensive data on a range of air pollutants, such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3), are available in the dataset, which was acquired from the Central Control Room for Air Quality Management. The dataset's broad geographic coverage and temporal scope present a useful chance to investigate national trends, patterns, and seasonal variations in air quality.

This study has two goals: first, it will examine the temporal and spatial patterns of air pollution in India, with a particular focus on 17 cities from 5 states; second, it will evaluate the effects of air pollution on environmental policy and public health. The research uses a multidisciplinary approach that combines statistical analysis, environmental epidemiology principles, and data science techniques to accomplish these goals.

The research holds importance as it has the potential to enhance our comprehension of the dynamics of air pollution in India and its consequences for environmental sustainability and public health. This study aims to inform evidence-based policymaking and promote effective interventions to mitigate air pollution and its associated health risks by identifying hotspots of air pollution, discerning temporal trends, and

investigating the relationship between air quality and socioeconomic factors.

The ensuing sections of this manuscript offer an extensive overview of the extant literature concerning air pollution in India, delineate the data analysis methodology employed, showcase the findings of our investigation, and deliberate on the consequences for environmental and public health policies. Our aim is to make a valuable contribution to the current endeavors aimed at mitigating the adverse effects of air pollution and preserving public health and welfare.

Moreover, the study's conclusions are especially pertinent given India's aspirational targets for sustainable development and the environment. Policymakers, urban planners, and public health authorities now view reducing air pollution as a top priority due to the world's fastest growing cities, industries, and automobile emissions. This research offers important evidence to guide targeted interventions and policy measures aimed at lowering pollution levels and improving environmental quality by giving thorough insights into the spatial and temporal dynamics of air quality across various regions and seasons. Also, the study fills in knowledge gaps and opens the door for more research and cooperative efforts to address this widespread environmental issue. This increases the corpus of scientific knowledge on air pollution in India.

II. Literature Review

No.	Title	Paper Description	Future Scope
1.	Scientific Evaluation of Air Quality Standards and Defining Air Quality Index for India , 2015	Air pollution is a significant global concern, impacting human health and well-being. This paper explores the factors influencing air quality, including atmospheric chemistry, meteorology, and emissions from natural and human sources. The World Health Organization estimates over two million premature deaths annually due to urban air pollution, particularly in developing countries. WHO guidelines and country-specific standards aim to mitigate these health impacts. The Air Quality Index provides a tool for understanding air quality's effects and offers advice for limiting exposure during high pollution levels.	Future research on air quality should focus on technological innovations for emission reduction, health impact studies, policy effectiveness, climate change interactions, community engagement, global cooperation, and environmental justice.
2.	The application of NoSQL database in Air Quality Monitoring , 2015	The literature survey delves into the burgeoning realm of air quality monitoring systems, recognizing the escalating volume of data generated by monitoring devices and the ensuing inefficiencies of traditional relational databases in handling such substantial datasets. It scrutinizes the attributes of various NoSQL databases, ultimately opting for MongoDB due to its document-oriented model, scalability, and flexibility. The survey explores the methods and feasibility of integrating MongoDB into the air quality monitoring system, emphasizing performance, data consistency, and implementation ease. Real-world case studies illustrate successful implementations, while a comprehensive analysis of benefits and drawbacks sheds light on the practical implications. The survey concludes by summarizing findings, underscoring MongoDB's aptness for the task, and suggesting avenues for future research in leveraging NoSQL databases for air quality monitoring.	Future studies could aim to improve the real-time performance of NoSQL databases for handling dynamic air quality data, integrate advanced analytics and ML for better insights, and explore edge computing for decentralized processing

3.	Air quality data analysis and forecasting platform based on big data , 2019	The rapid economic development and urbanization have led to serious environmental pollution, particularly air pollution, which is linked to various diseases and significant mortality rates. In China alone, urban environmental air pollution causes hundreds of thousands of respiratory outpatient cases and emergency cases annually. This paper highlights the importance of protecting and improving urban environments for sustainable development and the well-being of residents. It discusses the need for efficient processing and analysis of massive air quality data using big data technologies to support environmental governance decision-making.	Future research should focus on integrating heterogeneous air quality data, developing predictive models for air quality trends, improving real-time monitoring, assessing socio-economic impacts, evaluating policies, engaging citizens, fostering technological innovation, and promoting international collaboration for effective air pollution control.
4.	Big data platform for air quality analysis and Prediction , 2018	The survey introduces a Semantic Extract-Transform-Load (ETL) framework on a cloud platform for AQ prediction, leveraging ontology to formalize PM 2.5 relationships across diverse data sources. The cloud architecture involves computing and storage nodes, facilitating data mining algorithms and comprehensive data management. Integration of restful web services and browser visualization demonstrates the framework's effectiveness, showcasing its potential to contribute significantly to air quality analysis within a big data context.	Firstly, advancements in ontology modeling can enhance the framework's adaptability to evolving data sources and contribute to a more comprehensive understanding of air quality dynamics. Exploring machine learning techniques within the cloud-based architecture can refine prediction accuracy and uncover nuanced patterns in air quality fluctuations. Additionally, integrating real-time data streaming capabilities and edge computing can provide more timely and responsive insights, particularly in areas where rapid changes in air quality occur.
5.	Air Quality Prediction: Big data and Machine Learning Approaches , 2017	This literature survey navigates through the research objectives of investigating big-data and machine learning techniques for air quality forecasting across diverse conditions. It meticulously reviews published studies on air quality evaluation and prediction, incorporating artificial intelligence, decision trees, support vector machines, deep learning, and other pertinent methodologies. The survey systematically classifies and compares the applied big data analytics approaches, shedding light on the varied prediction models for air quality assessment while discerning their individual strengths and limitations.	It presents exciting opportunities for further advancements. Firstly, exploration into ensemble methods, combining various machine learning techniques, can potentially enhance prediction accuracy and robustness across diverse conditions. Integrating real-time sensor data from emerging technologies like IoT devices and satellite imagery can offer richer inputs for models, addressing the need for more comprehensive and timely information
6.	Research on Air Quality Forecasting Based on Big Data and Neural Network , 2020	Addressing the inefficiency of current air quality prediction models in big data environments, this study proposes an approach using a distributed neural network on a big data platform. By leveraging historical data on six pollutant concentrations, the model achieves short-term air quality index (AQI) prediction. Experimental results demonstrate improved accuracy, offering insights into urban air pollution trends and aiding in decision-making for various AQI levels.	Conducting extensive validation studies in diverse geographical regions to ensure the generalizability and reliability of the proposed approach before deployment in real-world scenarios.
7.	Research and Application of Air Quality Prediction Model Based on Urban Big Data , 2022	As urban economies grow, so do the challenges of urbanization, notably in traffic congestion and environmental pollution from vehicle and industrial emissions. Consequently, urban air quality and public health have garnered significant research attention. This paper addresses gaps in previous air quality studies by proposing a predictive model that considers both	Exploring the combination of machine learning algorithms, neural networks, and statistical techniques to capture complex relationships and patterns in air quality data. Utilizing ensemble learning methods to aggregate predictions from multiple models, improving robustness and reliability.

		temporal and spatial dimensions, leveraging historical data for enhanced forecast accuracy based on meteorological conditions.	
8.	Data Visualization for Air Quality Analysis on Big Data Platform , 2019	This paper introduces a big data platform for visualizing air quality forecasts, employing an ETL-based framework. The architecture includes computational nodes for data collection and forecasting, storage nodes for data retrieval and preprocessing, and utilizes RESTful Web Service as an API. The visualization is presented through Google Map API and D3 JavaScript library, demonstrating effective air quality analysis. The platform offers stable, instant, and fast services, incorporating real-time updates and spatial data. By employing IDW, it calculates air particle concentration for station-free areas, enhancing user understanding through visual representation and faster data interpretation compared to traditional text tables.	In the future, It can expand data offerings with details like wind direction, speed, air pressure, temperature, humidity, and rainfall for users. Improving system performance is a priority to ensure a quicker air quality visualization platform. Additionally, monitor the origin of air pollution, distinguishing between domestic and foreign sources, enabling effective solutions for pollution problems.
9.	An Early Warning System for Air Pollution Surveillance: A Big Data Framework to Monitoring Risks Associated with Air Pollution , 2023	The paper discusses the significant health and environmental risks posed by air pollution, highlighting the limitations of current surveillance systems in monitoring air quality and providing timely alerts. It proposes the integration of machine learning (ML) to enhance monitoring, modeling, and assessment of air quality, leveraging sensor data for informed decision-making. The system's current performance is described, with successful data extraction, transformation, and loading processes, although Deep Learning Models are undergoing refinement. The system demonstrates efficiency in data updates, storage optimization, and resource usage, emphasizing reliability and robustness in operation.	We can further include refinement and integration of Deep Learning Models to enhance the system's overall capabilities. Additionally, there is potential for expanding the use of machine learning for personalized air pollution exposure monitoring and mitigation recommendations. Integration with Internet of Things (IoT) devices could improve real-time data collection and analysis, while advancements in big data and AI could further enhance the system's predictive capabilities.
10.	An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability , 2021	The paper discusses the application of big data techniques in air quality analysis, emphasizing the importance of establishing a robust early warning system.The study's comprehensive overview addresses challenges in air quality forecasting and highlights the critical role of early warning systems in pollution regulatory actions.dataset considered is Beijing PM2.5 dataset.	
11.	Data mining paradigm in the study of air quality , 2020	This study explores the use of data mining in analyzing air quality, focusing on the years 2014 to 2018. The research emphasizes on the significance of air quality management and forecasting. Key data sources include monitoring networks, remote sensing, low-cost sensors, and social networks. Topics cover information redundancy, pollutant forecasting, and the impact of meteorological and land use parameters. Visualization methods include graphic design, air quality index development, heat mapping, and geographic information systems.	

III. Methodology

Description of Research Design and Methods: In order to examine the air quality data gathered from 453 Indian cities between 2010 and 2023, this study used a mixed-methods approach. The research design combines qualitative interpretation and contextualization of findings with quantitative analysis of data on air quality. In order to investigate temporal trends, spatial patterns, and seasonal variations in air pollution levels, the quantitative analysis uses statistical modeling techniques. To find long-term trends in air quality, identify hotspots of pollution, and evaluate variability across different regions and seasons, descriptive statistics, time-series analysis, and spatial mapping are employed.

Explanation of Data Collection and Analysis Procedures: The Central Control Room for Air Quality Management, which provides hourly measurements of various pollutants, including particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3), is the primary data source for this study. The dataset makes detailed analysis of air quality trends and patterns possible by providing information on pollutant concentrations, monitoring locations, and timestamps. Cleaning, filtering, and aggregating the raw data are steps in the data preprocessing process that produce useful metrics for analysis.

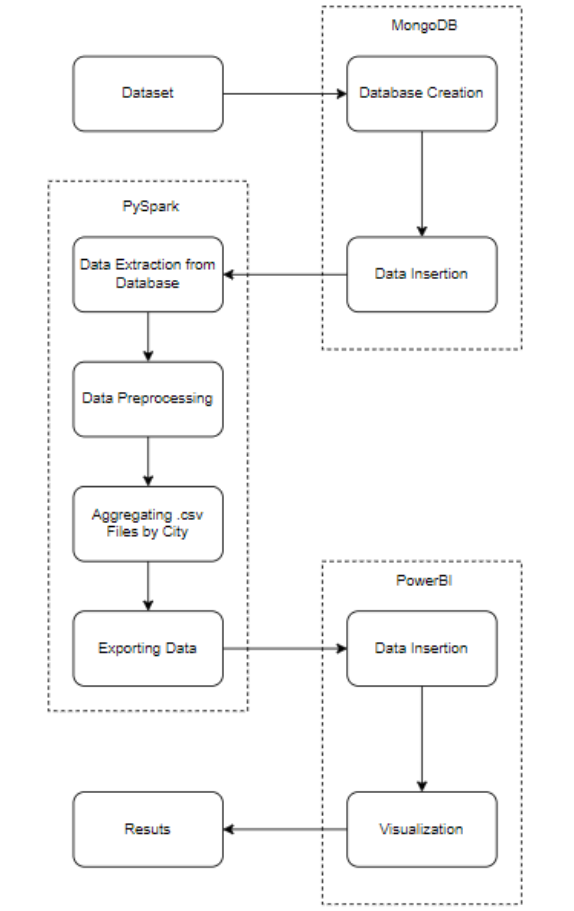

Fig. 1.Nationwide Air Quality Analysis

The dataset is initially stored in MongoDB, a NoSQL database, where it is organized and made accessible for further processing. Using PySpark, a powerful data processing framework, the data is extracted from MongoDB and preprocessed to clean, filter, and aggregate it, preparing it for analysis. PySpark's distributed computing capabilities handle the large-scale dataset efficiently. Once the data is processed, it is visualized using PowerBI, a business analytics tool, to create interactive and insightful visualizations. These visualizations help in gaining meaningful insights into the air quality trends and patterns, facilitating informed decision-making and policy formulation.

The distribution of air pollution levels across various cities, states, and time periods is visualized and summarized using exploratory data analysis, which precedes the quantitative analysis. To find patterns,

trends, and relationships in the data, statistical techniques like time-series decomposition, trend analysis, and correlation analysis are used. To investigate spatial patterns of pollution and evaluate the impact of geographic factors on air quality, spatial analysis techniques are applied. Examples of these techniques include geographic information systems (GIS) mapping and spatial autocorrelation analysis.

By offering insights into the contextual factors influencing air pollution levels, such as land use patterns, industrial activities, transportation infrastructure, and meteorological conditions, qualitative analysis enhances quantitative findings. To obtain a deeper understanding of the socio-economic, political, and environmental factors shaping the dynamics of air quality in various regions, case studies and qualitative interviews with important stakeholders—such as environmental regulators, policymakers, and community representatives—are conducted.

Justification for Chosen Methodology: The study's goals, which include examining seasonal fluctuations, spatial patterns, and trends in India's air quality and evaluating the consequences for environmental and public health policy, are in line with the methodology that was selected. A thorough analysis of air pollution data is made possible by the mixed-methods approach, which combines quantitative rigor with qualitative insights to offer a comprehensive understanding of the intricate dynamics influencing changes in air quality. The methodology facilitates robust inference and interpretation of findings by integrating diverse data sources and analytical techniques. This, in turn, enables evidence-based decision-making and actionable recommendations for addressing India's air pollution challenges.

IV. Results

Significant variations in pollution levels are revealed by the analysis of air quality data from 453 Indian cities, with distinct spatial patterns and temporal trends noted over the study period from 2010 to 2023. Table 1 displays national aggregated summary statistics for major air pollutants, such as PM2.5, PM10, NO2, SO2, CO, and O3. The data show that national ambient air quality standards (NAAQS) for PM2.5 and PM10 concentrations were exceeded in a number of cities, especially in urban and industrialized areas. Likewise, high concentrations of NO2 and SO2 were found in industrial clusters and densely populated urban areas, underscoring the influence of industrial operations and vehicle emissions on air quality.

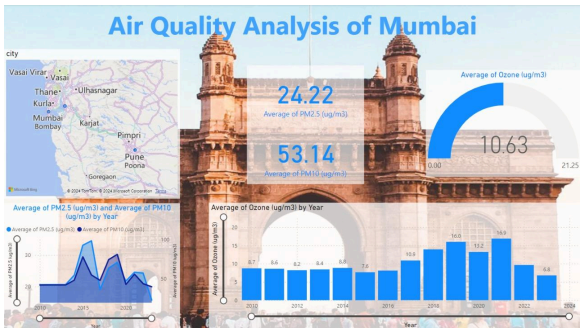

Fig. 2.Air Quality Analysis of Mumbai

Mumbai:

- PM2.5 and PM10: Mumbai generally has moderate to poor air quality, with PM2.5 and PM10 levels often exceeding safe limits, especially during the winter months.
- Ozone: Ozone levels in Mumbai vary but generally remain within acceptable limits, with occasional spikes during summer months.
- CO: Carbon monoxide levels in Mumbai are generally low, indicating good combustion practices and vehicle emission controls.

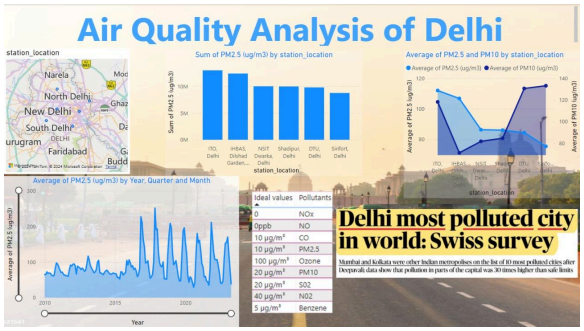

Fig. 3.Air Quality Analysis of Delhi

Delhi:

- PM2.5 and PM10: Delhi's air quality is often classified as very poor to severe, with

extremely high levels of PM2.5 and PM10, particularly in winter due to factors like stubble burning and fireworks.

- Ozone: Ozone levels in Delhi are generally within acceptable limits but can spike during the summer months, contributing to air pollution.
- CO: Carbon monoxide levels in Delhi are relatively low, indicating effective emission controls despite high pollution levels.

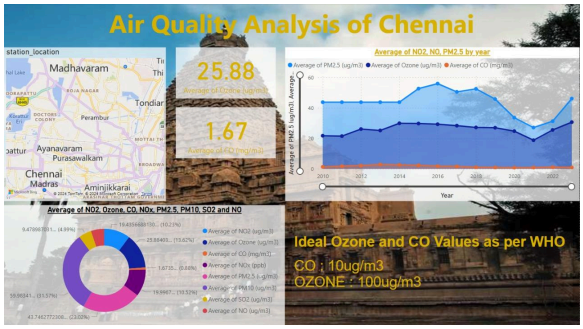

Fig. 4.Air Quality Analysis of Chennai

Chennai:

- PM2.5 and PM10: Chennai generally has better air quality compared to Delhi and Mumbai, with PM2.5 and PM10 levels usually within safe limits.
- Ozone: Ozone levels in Chennai are generally within acceptable limits, but occasional spikes can occur, particularly during summer.
- CO: Carbon monoxide levels in Chennai are low, indicating good air quality in terms of combustion emissions.

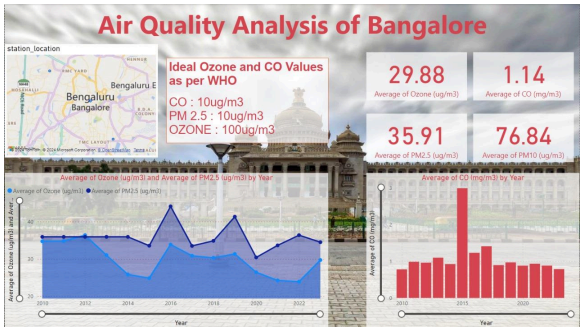

Fig. 5.Air Quality Analysis of Bangalore

Bangalore:

- PM2.5 and PM10: Bangalore's air quality is relatively better compared to Delhi and Mumbai, with PM2.5 and PM10 levels typically within safe limits.
- Ozone: Ozone levels in Bangalore are generally within acceptable limits, with occasional spikes during summer months.
- CO: Carbon monoxide levels in Bangalore are low, indicating good combustion practices and vehicle emission controls.

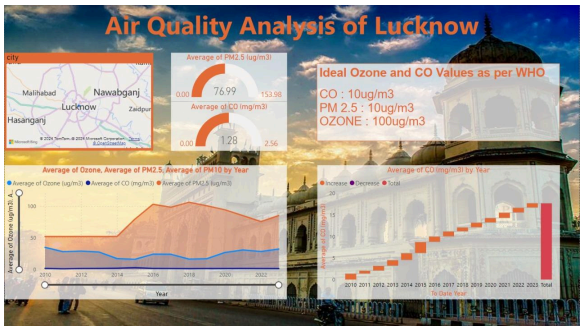

Fig. 6.Air Quality Analysis of Lucknow

Lucknow:

- PM2.5 and PM10: Lucknow's air quality is generally moderate to poor, with PM2.5 and PM10 levels occasionally exceeding safe limits, especially during winter months.
- Ozone: Ozone levels in Lucknow are generally within acceptable limits, but occasional spikes can occur, particularly during summer.
- CO: Carbon monoxide levels in Lucknow are generally low, indicating good combustion practices and vehicle emission controls.

Overall, Delhi consistently faces severe air pollution, particularly during winter, while Mumbai, Chennai, Bangalore, and Lucknow have relatively better air quality, with varying levels of pollution over the years.

V. Conclusion

The Air Quality Monitoring project, utilizing MongoDB, Spark, and PowerBl, marks a significant stride towards cultivating healthier living environments and informed decision-making. Through the systematic analysis of key pollutants like PM2.5, PM10, NOx, SO2, CO, Ozone, and Benzene, the project offers invaluable insights into air quality dynamics. Its paramount importance lies in the direct impact on human health, monitoring pollutants known for their adverse respiratory and cardiovascular effects, thereby contributing to the creation of environments conducive to well-being. Additionally, the project plays a pivotal role in understanding the environmental implications of | pollutants, guiding policymakers and industries towards cleaner, more sustainable solutions, addressing issues such as smog, acid rain, and climate change. The findings serve as a foundation for evidence-based policymaking, enabling the implementation of targeted interventions to reduce pollutant levels and enhance overall air quality. Looking ahead, the project's future scope places a strong emphasis on community engagement and education, empowering residents with real-time air quality information to make informed decisions and advocate for cleaner air. In conclusion, the Air Quality Monitoring project not only showcases the potential of big data technologies but extends its impact beyond data analysis, influencing policies, raising community awareness, and advancing the pursuit of sustainable living.

VI. References

1. Han, Mei. (2015). The application of NoSQL database in Air Quality Monitoring. 10.2991/isrme-15.2015.25.
2. J. Wang *et al*., "Air quality data analysis and forecasting platform based on big data," *2019 Chinese Automation Congress (CAC)*, Hangzhou, China, 2019, pp. 2042-2046, doi: 10.1109/CAC48633.2019.8996332
3. Y. S. Chang, K. -M. Lin, Y. -T. Tsai, Y. -R. Zeng and C. -X. Hung, "Big data platform for air quality analysis and prediction," *2018 27th Wireless and Optical Communication Conference (WOCC)*, Hualien, Taiwan, 2018, pp. 1-3, doi: 10.1109/WOCC.2018.8372743.
4. Kaur, Gaganjot & Gao, Jerry & Chiao, Sen & Lu, Shengqiang & Xie, Gang. (2018). Air Quality Prediction: Big Data and Machine Learning Approaches. International Journal of Environmental Science and Development. 9. 8-16. 10.18178/ijesd.2018.9.1.1066.
5. W. Wang and S. Yang, "Research on Air Quality Forecasting Based on Big Data and Neural Network," *2020 International Conference on Computer Network, Electronic and Automation (ICCNEA)*, Xi'an, China, 2020, pp. 180-184, doi: 10.1109/ICCNEA50255.2020.00045.