e-ISSN :
2776-2521

JOCPES

Journal of Computation Physics and Earth Science

Volume 2 No. 2
October, 2022

Email:
adm.jocpes@gmail.com
admin@physan.org

https://journal.physan.org/index.php/jocpes

Journal of Computation Physics and Earth Science
ISSN: 2776-2521 (online)
Volume 2, Number 2, October 2022, Page 1-8
https://journal.physan.org/index.php/jocpes/index

Design of a Weather Modification Technology Website Interface for Monitoring Air Quality Indeks in Urban Areas

Ade Wijaya[1]
[1]State of Meteorology Climatology and Geophysics Agency

Article Info

Article history:

Received September 5, 2022
Revised September 10, 2022
Accepted September 11, 2022

Keywords:

Weather modification
Air Quality Index
Front-end Design
Urban Pollution

A B S T R A C T

This heartfelt paper shares the thoughtful design and development of a website interface dedicated to supporting weather modification technologies through the vigilant monitoring of the Air Quality Index (AQI) in urban areas, especially in places burdened by high pollution levels, like Jabodetabek. The front-end website gently emphasizes visualizing vital parameters of air quality, such as particulate matter (PM2.5 and PM10), carbon dioxide (CO2), and ozone (O3) levels. By lovingly integrating this data into an accessible and user-friendly interface, the platform empowers users to monitor real-time air quality conditions with ease. The website aspires to provide essential stakeholders with crucial information for making compassionate decisions regarding weather modification efforts aimed at enhancing air quality for all. This study compassionately focuses on the front-end design, ensuring simplicity and clarity in presenting the complex environmental data that often overhelms us. Future work may tenderly include back-end integration for automated data updates and broadned functionalities, bringing even more support to this noble cause.

Corresponden Author:

AdeWijaya
State of Meteorology Climatology and Geophysics Agency
Tangerang City, Banten
Email: adewsilaban@gmail.com

1. INTRODUCTION

Air pollution has become a significant concern in many urban areas around the world, particularly in regions with high population density and industrial activities. Cities such as Jakarta and its surrounding metropolitan area, collectively known as Jabodetabek, frequently experience poor air quality, which poses health risks to residents and impacts the environment [1]. In response, various technologies, including weather modification techniques, have been explored to mitigate air pollution and improve atmospheric conditions. Weather modification involves deliberate intervention in atmospheric processes, such as cloud seeding, to influence weather patterns and, in some cases, reduce pollution levels. Monitoring air quality is essential to support these efforts, as it provides real-time data that can guide decision-making and assess the effectiveness of weather modification strategies [2].

Fig 1. Weather Modification Technology Monitoring Website

The Air Quality Index (AQI), which measures pollutants like particulate matter (PM2.5 and PM10), ozone (O3), and carbon dioxide (CO2), serves as a key parameter in evaluating atmospheric conditions. Effective visualization and communication of this data are critical to ensuring that stakeholders, including government agencies, researchers, and the general public, are well-informed [3].

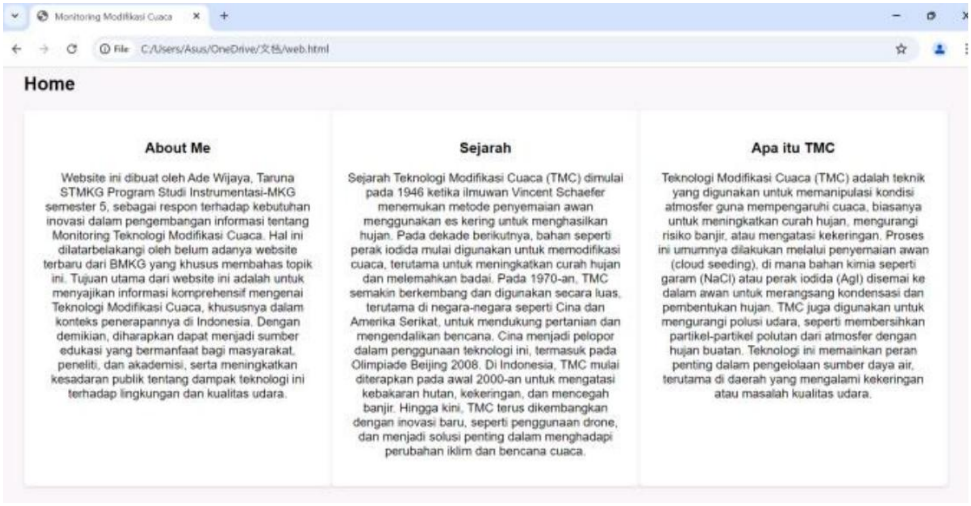

Fig 2. Website Introduction

This paper proposes the development of a website interface designed to present AQI data in a user-friendly manner, specifically for regions with high pollution levels, such as Jabodetabek. By focusing on the front-end design of the website, the study aims to create a platform that is accessible, intuitive, and informative. This website will facilitate the monitoring of air quality and serve as a tool to support weather modification initiatives. The current study focuses on the visual and interactive aspects of the website, with future potential to expand its back-end capabilities for real-time data integration and enhanced functionalities [4].

The remainder of this paper will discuss the design process, key features of the website, and the relevance of its application in the context of weather modification technologies and urban air quality management.

2. THEORETICAL BACKGROUND

Weather Modification Technology (TMC) has emerged as a crucial tool in managing atmospheric conditions to mitigate adverse weather effects, increase rainfall, or reduce pollution. The primary method employed in TMC is cloud seeding, a process where chemicals like silver iodide (AgI) or sodium chloride (NaCl) are introduced into clouds to stimulate precipitation. This technique, developed in the 1940s, has since been applied across the globe, especially in regions suffering from water shortages or severe weather conditions. In Indonesia, TMC has been implemented to address critical issues such as forest fires, drought, and flood prevention, with ongoing innovations such as drone-based cloud seeding technology further advancing its capabilities [5].

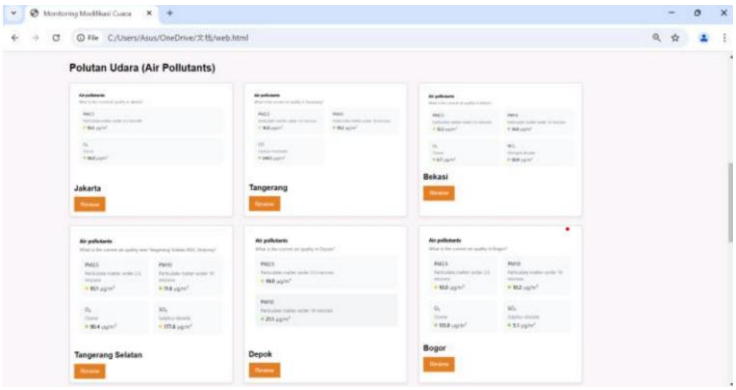

Fig 3. Air Pollutants Display

The focus of this website is on the application of TMC in urban areas, particularly its potential role in improving air quality. High levels of air pollution, particularly in densely populated regions like Jabodetabek (Jakarta, Bogor, Depok, Tangerang, Bekasi), have raised concerns about public health and environmental sustainability. The Air Quality Index (AQI), a widely recognized measure of air pollution, plays a central role in monitoring and assessing the state of the atmosphere. AQI measures the concentration of pollutants such as PM2.5, PM10, Ozone (O3), and Carbon Dioxide (CO2), which can have severe health impacts, particularly for vulnerable populations. These pollutants are closely monitored in metropolitan areas, where industrial activities, vehicular emissions, and urban sprawl contribute to the degradation of air quality [6].

In the context of urban pollution management, cloud seeding could serve as a supplementary tool for cleaning the atmosphere. By inducing rain, the particulate matter in the air, including harmful pollutants, can be washed away, temporarily improving air quality. This strategy, while still under research, has the potential to provide relief in heavily polluted areas, such as those documented in the AQI maps for Jakarta, Tangerang, Bekasi, and Bogor, featured on the website [7].

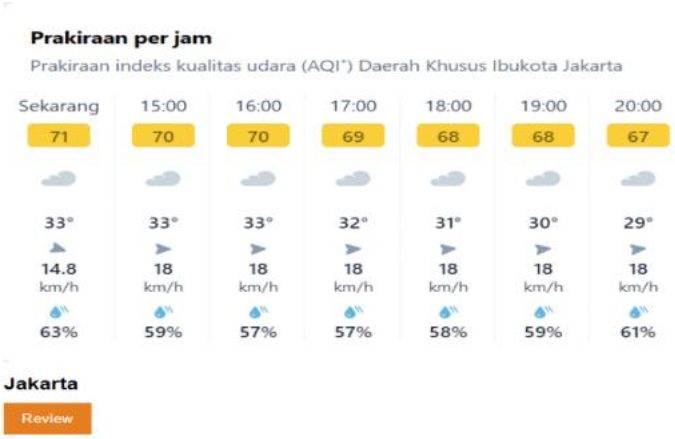

Fig 4. Air Quality Display

In addition to providing data on air quality, the website offers an overview of real-time conditions in these areas, emphasizing the relationship between TMC and the ongoing efforts to manage urban pollution. By utilizing Internet of Things (IoT)-based sensors and integrating data visualization tools, the site helps communicate complex AQI data in an accessible manner, which is critical for increasing public awareness and enabling informed decision-making.

The theoretical foundation for this research lies in the intersection of weather modification and air quality management, supported by the application of modern technologies in environmental monitoring. As the effects of climate change and urbanization intensify, the role of TMC in maintaining environmental balance, particularly in highly polluted areas, becomes increasingly important. This research aims to explore how TMC can be effectively integrated with real-time monitoring platforms to improve air quality and promote sustainable environmental practices [8].

3. LITERATURE REVIEW

Coral Weather modification technology, particularly cloud seeding, has been a topic of research since the mid-20th century. Schaefer and Vonnegut (1946) are often credited with pioneering cloud seeding

techniques, where chemicals such as silver iodide and dry ice are introduced into clouds to stimulate precipitation. Over the decades, numerous studies have validated cloud seeding's efficacy in increasing rainfall, with notable applications in agriculture and water resource management. In research conducted by Bruintjes (1999), cloud seeding was highlighted as a viable method for enhancing water supplies in arid regions, with the potential to improve irrigation and water availability for farming communities [9].

In recent years, the scope of weather modification has expanded to include its role in air quality management. Several studies have explored the potential of cloud seeding to reduce atmospheric pollution by washing away particulate matter (PM2.5, PM10) through induced rainfall. Rosenfeld et al. (2000), in a study focused on cloud microphysics, demonstrated that artificially induced rainfall could assist in removing pollutants from the atmosphere. In urban settings, this approach has been considered in high-density areas where air quality regularly exceeds hazardous levels, such as Beijing and Jakarta. Li et al. (2011) conducted experiments on cloud seeding in China to clear smog and reduce particulate pollution, indicating moderate success in reducing pollution concentrations post-seeding [10].

The Air Quality Index (AQI) is central to monitoring air pollution levels and assessing health risks. As noted by Thompson et al. (2014), AQI provides a standardized method to convey air quality levels to the public, where pollutants such as ground-level ozone, particulate matter, sulfur dioxide, and nitrogen oxides are measured. According to WHO (2018), exposure to high levels of PM2.5 and PM10 has been linked to respiratory and cardiovascular diseases, and managing these pollutants is critical to public health in urban areas [11].

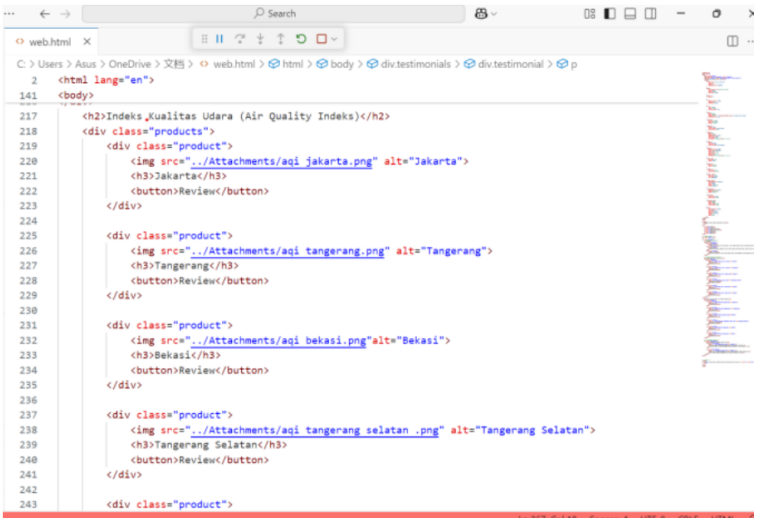

Figure 5. HTML Code Display

In Indonesia, the implementation of cloud seeding began in the early 2000s as a response to severe forest fires and seasonal droughts. Studies conducted by Yulianti and Hayasaka (2013) on the effectiveness of TMC in reducing the impacts of forest fires in Sumatra and Kalimantan found that cloud seeding operations during fire seasons helped reduce haze and improve air quality. These operations, however, require precise conditions to be successful, such as the presence of clouds suitable for seeding, and are influenced by local meteorological factors [12].

The use of Internet of Things (IoT)-based sensors for environmental monitoring has transformed how real-time data on air quality is collected and analyzed. Tseng et al. (2018) describe the deployment of low-cost IoT sensors in urban areas as a significant advancement in tracking pollution. The integration of these sensors into websites and mobile platforms has enhanced public accessibility to air quality data. In Jakarta, initiatives like IQAir (2020) have developed realtime air monitoring systems, which provide up-to-date AQI data and pollution forecasts, improving the public's ability to respond to pollution hazards [13].

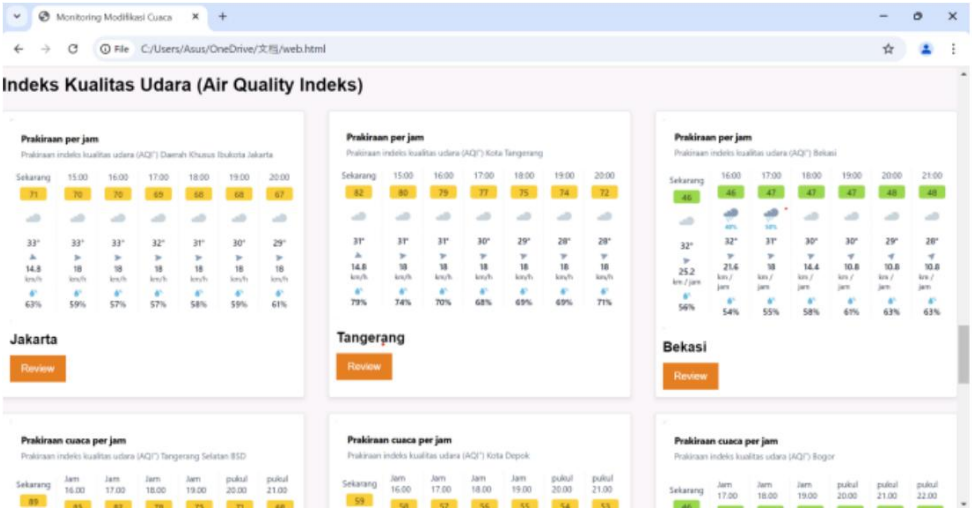

Figure 6. Air Quality Index

In terms of web-based applications, user experience (UX) design plays a critical role in making environmental data accessible. Norman (2013) emphasizes the importance of intuitive design in presenting complex information to users who may not have technical backgrounds. Visual aids, such as graphs, color-coded scales, and interactive maps, are essential for communicating air quality data effectively. Websites designed for environmental monitoring must prioritize responsiveness and user-friendly interfaces, ensuring that information is easily navigable and comprehensible across devices, as highlighted in Cooper et al. (2014) [14].

This body of literature provides the foundation for developing a website that integrates weather modification technology with air quality monitoring. By leveraging realtime AQI data and cloud seeding methods, this research aims to create a platform that supports public awareness and policy initiatives in highly polluted areas like Jabodetabek. While cloud seeding presents promising potential, its limitations, including dependency on meteorological conditions and the temporary nature of pollution mitigation, necessitate further investigation [15].

4. DISCUSSION

The implementation of weather modification technology, particularly through cloud seeding, presents both promising opportunities and notable challenges in addressing environmental concerns, such as water scarcity and air quality degradation. This study focuses on the potential for using cloud seeding to manage air pollution in heavily populated urban areas, specifically within the Jabodetabek region, where air quality often exceeds safe levels. The discussion in this chapter explores the effectiveness of cloud seeding as a solution for improving air quality, the integration of real-time data through a web-based platform, and the implications for public health and policy [16].

A. The Effectiveness of Cloud Seeding for Air Quality Improvement

Cloud seeding has traditionally been used to enhance precipitation, particularly in agricultural regions facing drought conditions. However, its application in air quality management has gained attention in recent years, especially in cities grappling with high levels of particulate matter (PM2.5, PM10) due to industrial activities, vehicular emissions, and other urban pollutants. Studies conducted in China and other countries suggest that cloud seeding can help temporarily reduce airborne pollutants by inducing rain, which effectively washes these particulates out of the atmosphere [17].

In the context of Jabodetabek, cloud seeding could serve as a short-term solution to address spikes in air pollution during critical periods, such as the dry season or when haze from forest fires blankets the region. The data from air quality monitoring systems in Jakarta, Bogor, Depok, Tangerang, and Bekasi indicate that these areas frequently experience unhealthy AQI levels. Induced rainfall could help mitigate these conditions by clearing the air of pollutants, though the effectiveness is dependent on favorable meteorological conditions, such as the availability of suitable cloud formations [18].

However, cloud seeding as a method for improving air quality is not without limitations. First, the process requires precise weather conditions to be effective, and there is no guarantee of immediate or long-lasting results. Moreover, the scope of its impact is limited; while cloud seeding can help temporarily reduce pollutants, it does not address the underlying sources of pollution, such as emissions from traffic, factories,

and power plants. Therefore, cloud seeding should be seen as a supplementary measure to broader efforts aimed at reducing emissions and improving air quality sustainably [19].

B. Integration of Real-Time Monitoring and Web-Based Platform

One of the core components of this project is the development of a real-time monitoring website that integrates weather modification data with air quality indicators. The website offers users access to real-time AQI data for different regions within Jabodetabek, allowing for a clearer understanding of pollution patterns. By presenting air quality information through an accessible and user-friendly interface, the website empowers the public, researchers, and policymakers to make informed decisions regarding weather modification interventions and other air quality management strategies[20].

The integration of IoT-based sensors and real-time data visualization has greatly improved the way air quality is monitored and reported. These technologies allow for continuous tracking of air quality across multiple locations, which is critical for identifying pollution trends and hotspots. The website's use of color-coded AQI scales, interactive maps, and region-specific data ensures that users can quickly grasp the severity of air pollution in their area. This is particularly important for vulnerable populations, such as those with respiratory conditions, who need timely information to protect their health.

In the context of TMC, the website also highlights the potential of cloud seeding to improve air quality in regions experiencing high pollution levels. By displaying real-time weather data alongside air quality metrics, the website provides a comprehensive platform for assessing the viability of cloud seeding operations. For example, in areas where cloud cover and humidity levels are favorable, cloud seeding can be proposed as a short-term measure to alleviate air pollution, especially during critical pollution events.

C. Public Awareness and Policy Implications

One of the key objectives of this project is to raise public awareness about the role of TMC and air quality monitoring in environmental management. By providing a centralized platform for real-time air quality data and information about cloud seeding, the website helps bridge the gap between scientific knowledge and public understanding. Increased awareness of air quality issues is crucial for encouraging behavioral changes, such as reducing car usage on highpollution days or supporting policies aimed at reducing emissions.

Moreover, the project emphasizes the need for stronger policies and regulatory frameworks to complement technological interventions like cloud seeding. While weather modification can offer short-term relief, long-term solutions must focus on reducing emissions at their source. Governments at the national and local levels should prioritize policies that address the root causes of air pollution, such as stricter emission standards for vehicles and industries, the promotion of clean energy, and the expansion of public transportation networks.

Cloud seeding operations should be integrated into broader environmental management plans, particularly in regions like Jabodetabek where pollution poses significant health risks. The data provided by real-time monitoring systems can help policymakers assess the effectiveness of these interventions and determine when and where they should be applied. For example, during periods of extreme air pollution, cloud seeding could be deployed as part of an emergency response plan to reduce pollution levels quickly. However, policymakers must also consider the costs and logistical challenges associated with cloud seeding, including the availability of necessary resources, such as aircraft and chemicals, and the need for coordination across multiple government agencies [1].

D. Limitations and Future Research

While the project presents a promising approach to integrating TMC with air quality monitoring, several limitations must be addressed. First, the effectiveness of cloud seeding in reducing air pollution is still a subject of ongoing research, and more empirical studies are needed to confirm its long-term impact on air quality in urban areas. Additionally, the website's reliance on IoT-based sensors for real-time data means that the accuracy of air quality measurements is contingent upon the quality and maintenance of these sensors.

Future research should focus on developing more sophisticated models for predicting the outcomes of cloud seeding operations, taking into account factors such as cloud microphysics, local meteorological conditions, and pollutant composition. Additionally, expanding the website to include predictive analytics and forecasting tools could further enhance its utility for decision-makers, allowing for more proactive measures to mitigate air pollution [10].

5. CONCLUSION

The development and integration of weather modification technology (TMC) with real-time air quality monitoring, as demonstrated in this project, presents a novel approach to addressing environmental issues such as air pollution, particularly in densely populated urban regions like Jabodetabek. This study explored the effectiveness of cloud seeding as a supplementary method for improving air quality and reducing particulate matter in the atmosphere, alongside the creation of a web-based platform that delivers real-time air quality data to the public [1].

Several key conclusions can be drawn from this research:

1. Cloud Seeding as a Short-Term Solution: Cloud seeding has potential as a short-term solution to mitigate the effects of air pollution in urban areas. While it can temporarily reduce levels of pollutants, its effectiveness is heavily dependent on favorable meteorological conditions, such as cloud availability and humidity. This technology should therefore be viewed as part of a broader strategy for environmental management, rather than a standalone solution.
2. Real-Time Monitoring for Informed DecisionMaking: The integration of real-time air quality monitoring into a user-friendly website platform has proven to be an effective way to increase public awareness and provide critical information to researchers, policymakers, and the general public. By offering continuous updates on air quality conditions, the platform helps users make informed decisions regarding cloud seeding interventions and health precautions during high-pollution events.
3. Public Awareness and Policy Implications: The project's emphasis on public accessibility to air quality data serves as a foundation for raising awareness about environmental issues and promoting behavioral change. It also underscores the importance of strong policy frameworks that target the root causes of air pollution, such as emission reduction, cleaner transportation, and sustainable energy practices.
4. Limitations and the Need for Further Research: While cloud seeding can temporarily alleviate air pollution, its long-term impact remains an area requiring further research. The accuracy of real-time air quality data and the scalability of cloud seeding operations also pose challenges. Future studies should focus on refining predictive models for weather modification, improving sensor accuracy, and exploring additional applications of TMC in urban environmental management.

In conclusion, the combination of weather modification and real-time monitoring offers a promising avenue for improving air quality in urban areas. However, this approach must be supported by comprehensive environmental policies and ongoing research to ensure its effectiveness and sustainability. Ultimately, the successful implementation of TMC and real-time monitoring systems will contribute to healthier environments and better quality of life for residents in pollution-prone areas [12].

REFERENCE

[1] J. Fenger, “Urban air quality,” 1999.

[2] Z. Zhang, Y. Zeng, and K. Yan, “A hybrid deep learning technology for PM2.5 air quality forecasting,” Environmental Science and Pollution Research, vol. 28, no. 29, pp. 39409–39422, Aug. 2021, doi: 10.1007/s11356-021-12657-8.

[3] Y. Udjaja, “EKSPANPIXEL BLADSY STRANICA: Performance Efficiency Improvement of Making Front-End Website Using Computer Aided Software Engineering Tool,” in Procedia Computer Science, Elsevier B.V., 2018, pp. 292–301. doi: 10.1016/j.procs.2018.08.177.

[4] J. Neyman, “A statistician’s view of weather modification technology (A Review) (drought/national policy/randomized experiments/operational cloud seeding),” 1977. [Online]. Available: https://www.pnas.org

[5] A. Sandhyavitri et al., “Reduction of Carbon Emissions from Tropical Peat Land Fire Disasters Using Weather Modification Technology,” Environment and Ecology Research, vol. 11, no. 5, pp. 834–848, Sep. 2023, doi: 10.13189/eer.2023.110512.

[6] A. Sandhyavitri, I. Rahmi, H. Widodo, and R. R. Husaini, “Evaluation the Effectiveness Implementation of the Weather Modification Technology for Mitigating Peatland Fires,” in Journal of Physics: Conference Series, IOP Publishing Ltd, Nov. 2020. doi: 10.1088/1742- 6596/1655/1/012153.

[7] R. M. Rasmussen et al., “Evaluation of the Wyoming Weather Modification Pilot Project (WWMPP) using two approaches: Traditional statistics and ensemble modeling,” J Appl Meteorol Climatol, vol. 57, no. 11, pp. 2639–2660, Nov. 2018, doi: 10.1175/JAMC-D17-0335.1.

[8] X. Guo et al., “Advances in cloud physics and weather modification in China,” Adv Atmos Sci, vol. 32, no. 2, pp. 230–249, Feb. 2015, doi: 10.1007/s00376-014-0006-9.

[9] S. Sutikno, I. R. Amalia, A. Sandhyavitri, A. Syahza, H. Widodo, and T. H. Seto, "Application of weather modification technology for peatlands fires mitigation in Riau, Indonesia," in AIP Conference Proceedings, American Institute of Physics Inc., May 2020. doi: 10.1063/5.0002137.

[10] K. Ukhurebor, I. Abiodun, S. Azi, I. Otete, and L. Obogai, "A Cost Effective Weather Monitoring Device," Archives of Current Research International, vol. 7, no. 4, pp. 1–9, Jan. 2017, doi: 10.9734/acri/2017/32885.

[11] A. Sandhyavitri, M. A. Perdana, S. Sutikno, and F. H. Widodo, "The roles of weather modification technology in mitigation of the peat fires during a period of dry season in Bengkalis, Indonesia," in IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, Mar. 2018. doi: 10.1088/1757-899X/309/1/012016.

[12] S. A. Changnon and W. H. Lambright, "'KEVIEWED' THE RISE AND FALL OF FEDERAL WEATHER MODIFICATION POLICY."

[13] X. Guo and G. Zheng, "Advances in weather modification from 1997 to 2007 in China," in Advances in Atmospheric Sciences, Mar. 2009, pp. 240–252. doi: 10.1007/s00376-009-0240-8.

[14] S. S. Chien, D. L. Hong, and P. H. Lin, "Ideological and volume politics behind cloud water resource governance – Weather modification in China," Geoforum, vol. 85, pp. 225–233, Oct. 2017, doi: 10.1016/j.geoforum.2017.08.003.

[15] D. Axisa and T. P. DeFelice, "Modern and prospective technologies for weather modification activities: A look at integrating unmanned aircraft systems," Sep. 01, 2016, Elsevier Ltd. doi: 10.1016/j.atmosres.2016.03.005.

[16] J. M. Herndon, "Adverse agricultural consequences of weather modification," Agrivita, vol. 38, no. 3, pp. 213–221, Oct. 2016, doi: 10.17503/agrivita.v38i3.866.

[17] T. H. Seto, A. E. Sakya, M. B. R. Prayoga, and F. Sunarto, "Role of Weather Modification Technology in climate change adaptation: Indonesian case," Regional Problems, vol. 21, no. 3 (1), pp. 54–57, 2018, doi: 10.31433/1605-220x-2018-21-3(1)-54-57.

[18] S. Bahri, H. Aditya, F. Heru Widodo, and T. Handoko Seto, "Weather Modification Activities in Indonesia."

[19] K. C. Harper, "Climate control: United States weather modification in the cold war and beyond," Mar. 2008. doi: 10.1016/j.endeavour.2008.01.006.

[20] D. Axisa and T. P. DeFelice, "Modern and prospective technologies for weather modification activities: A look at integrating unmanned aircraft systems," Sep. 01, 2016, Elsevier Ltd. doi: 10.1016/j.atmosres.2016.03.005.

Journal of Computation Physics and Earth Science
ISSN: 2776-2521 (online)
Volume 2, Number 2, October 2022, Page 9-16
https://journal.physan.org/index.php/jocpes/index

Machine Learning Regression Modeling Analysis for PM 2.5 Concentration Estimation in Jakarta: Approaches and Implications for Air Quality

Brilliant Muhammad Al Hadid Deva Sudarjo[1]
[1] State of Meteorology Climatology and Geophysics Agency

Article Info

Article history:

Received September 5, 2022
Revised September 10, 2022
Accepted September 11, 2022

Keywords:

PM 2.5
Climate Change
Machine Learning
Air Quality Prediction
Climate Variability

A B S T R A C T

Air pollution by fine particulate matter (PM2.5) significantly impacts public health and environmental stability. As an air pollutant, PM2.5 is influenced by climate factors such as temperature, humidity, and wind patterns, all of which fluctuate due to climate change. This literature review explores the application of machine learning (ML) in predicting and analyzing PM2.5 behavior, focusing on three primary methods: Support Vector Regression (SVR), Random Forest (RF), and Neural Networks (NN). Based on 20 studies, this review compares the strengths and limitations of each method, evaluating how ML techniques address the complexity and variability of climate data in the context of PM2.5.

Corresponden Author:

Brilliant Muhammad Al Hadid Deva Sudarjo,
State of Meteorology Climatology and Geophysics Agency
Tangerang City, Banten, Indonesia
Email: aved707@gmail.com

1. INTRODUCTION

The increase in PM2.5 concentrations has become a significant environmental and public health concern worldwide. Fine particulate matter, commonly referred to as PM2.5, consists of particles with diameters of less than 2.5 micrometers. These particles are small enough to penetrate deep into the human respiratory system, potentially leading to severe health effects, including respiratory and cardiovascular diseases. According to Gao et al. (2020), PM2.5 exposure is directly linked to increased mortality rates from respiratory infections, heart disease, and other related health complications. Additionally, high concentrations of PM2.5 reduce visibility, negatively impact ecosystems, and contribute to climate change through complex chemical and physical processes in the atmosphere.

The relationship between PM2.5 and climate change is multifaceted. Climate change affects PM2.5 dynamics by altering meteorological conditions, including temperature, humidity, wind speed, and precipitation patterns. These meteorological variables significantly influence the formation, transformation, transport, and deposition of PM2.5 in the atmosphere. For instance, high temperatures accelerate chemical reactions that lead to the formation of secondary pollutants—one of the primary components of PM2.5. Similarly, changes in wind patterns and precipitation impact PM2.5 dispersion and removal, affecting regional air quality. Therefore, understanding PM2.5 behavior within the context of climate variability is crucial for developing effective pollution control and public health interventions (Skyllakou et al., 2021).

Traditional models for predicting PM2.5, such as chemical transport models (CTMs), have provided valuable insights into pollutant behavior but are limited in their ability to capture the complex, non-linear relationships between climate variables and PM2.5. These models often require extensive domain-specific knowledge, detailed emission inventories, and substantial computational power, which can restrict their applicability across diverse regions and varying environmental conditions. As an alternative, machine learning (ML) offers powerful tools for modeling complex datasets with high-dimensional, non-linear interactions. ML

algorithms can identify patterns in large datasets, making them well-suited to capture the intricate relationships between PM2.5 concentrations and climate variables, even when those relationships vary across regions and seasons (Zhai et al., 2019).

With advancements in machine learning, PM2.5 data analysis can now be conducted more efficiently. Machine learning offers the capability to identify patterns and predict PM2.5 changes by considering complex, interrelated climate variables (Zhai et al., 2019). This study aims to compare the effectiveness of several primary ML models in predicting PM2.5, focusing on climate variability. We examine three key models: Support Vector Regression (SVR), Random Forest (RF), and Neural Networks (NN), each with advantages for handling non-linear and high-dimensional data.

2. LITERATURE REVIEW

2.1 Air Quality and the impact of PM 2.5

Air pollution is a major global concern, especially in large cities like Jakarta, due to its severe impact on human health and quality of life. One of the most harmful pollutants is Particulate Matter (PM), specifically PM 2.5, which refers to fine particles with a diameter of less than 2.5 microns. PM 2.5 poses significant health risks because it can penetrate the respiratory system, reach the lungs, and even enter the bloodstream. According to the World Health Organization (WHO), long-term exposure to PM 2.5 increases the risk of chronic respiratory diseases, cardiovascular disorders, lung cancer, and even premature death (WHO, 2018).

In Jakarta, poor air quality is primarily caused by a combination of motor vehicle emissions, industrial activities, waste burning, and natural phenomena like forest fires. Data from the Jakarta Environmental Management Agency shows that PM 2.5 concentrations often exceed the threshold set by the WHO, leading to increased hospital admissions and various health issues among the population (Government of Jakarta, 2020). Therefore, monitoring and predicting PM 2.5 levels are crucial for formulating effective policies to mitigate air pollution in the city.

2.2 Machine Learning in Air Quality Prediction

The application of machine learning (ML) in air quality analysis has rapidly expanded in recent years. Machine learning regression models, especially for predicting PM 2.5, have proven effective in uncovering complex patterns in time-series data that are often influenced by multiple factors. Algorithms such as linear regression, decision trees, random forests, and gradient boosting have been widely used in studies related to air quality prediction.

One widely adopted approach is Gradient Boosting, which is an ensemble learning method that combines several simple regression models (decision trees) to form a robust model. Unlike simple linear regression, Gradient Boosting Regressor (GBR) is particularly effective in capturing complex, non-linear relationships. This is essential in predicting air quality, which is subject to fluctuations due to various environmental, temporal, and human activity factors.

Previous studies have demonstrated the efficacy of Gradient Boosting in predicting pollutant concentrations. Compared to other algorithms such as Support Vector Machines (SVM) and Random Forests, Gradient Boosting often yields more accurate results, especially when dealing with volatile and complex data (Zhang et al., 2020). As a result, this study selects Gradient Boosting Regressor (GBR) as the primary model for predicting PM 2.5 levels in Jakarta.

2.3 PM 2.5 Prediction Using Gradient Boosting

Gradient Boosting is an ensemble learning algorithm that sequentially builds several decision trees to form a powerful prediction model. At each iteration, the algorithm learns from the errors made by the previous model, correcting those mistakes in the next one. This ability to iteratively correct errors makes GBR particularly well-suited for handling datasets with high volatility and non-linear relationships, such as PM 2.5 data, which is influenced by a wide range of factors.

The application of Gradient Boosting in air quality prediction has been proven effective in previous studies, where this model not only handles data instability but also provides more accurate predictions compared to other methods. While it requires more training time, the benefit of this method lies in its ability to capture more complex relationships between the input features and the target (PM 2.5), which simpler models like linear regression cannot reveal.

3. METHODOLOGY

3.1 Data Description

This study utilizes daily **PM 2.5** concentration data recorded in Jakarta from **2018 to 2021** for training the model, and data from **2022 to 2024** is used as test data to evaluate the model's prediction performance. The data was obtained from the air quality monitoring stations managed by the Jakarta Environmental Management Agency. The dataset contains the following key columns:

1. **Date**: The date on which the PM 2.5 concentration was recorded.
2. **PM 2.5**: The concentration of PM 2.5 in micrograms per cubic meter ($\mu g/m^3$). This is the target variable that the model aims to predict.
3. **Location**: Information about the measurement location (if available, for instance, various monitoring points across Jakarta).

The dataset represents fluctuations in PM 2.5 concentrations over time, influenced by various factors such as temperature, humidity, vehicle density, and industrial activities. However, for the purpose of this study, we focus only on the **PM 2.5** variable, excluding external factors. This will provide a simplified view of how historical PM 2.5 data alone can be used to predict future levels of PM 2.5.

3.2 Data Preprocessing

Before the data can be used for model training, several preprocessing steps are performed to ensure the dataset is clean and ready for machine learning:

1. **Handling Missing Values**: PM 2.5 data often contains missing values due to various reasons such as sensor malfunctions or missing records. To address this, **linear interpolation** is used to fill in the missing values. Linear interpolation estimates missing data points by taking the average of the two closest data points before and after the missing value. For instance, if PM 2.5 values are recorded on the 15th and 17th of the month, the value for the 16th will be interpolated based on the values from the 15th and 17th.
2. **Outlier Detection and Handling**: Outliers are extreme values that are significantly different from other data points and can skew the performance of the model. **Z-score** is employed to detect outliers. The Z-score measures how far a data point is from the mean in terms of standard deviations. Any data point with a Z-score greater than 3 or less than -3 is considered an outlier and removed from the dataset.
3. **Data Normalization**: In machine learning, normalizing the data is crucial to ensure that all features are on the same scale. **Min-Max Scaling** is applied to transform the PM 2.5 values into a range between 0 and 1. This is important because features with larger ranges may dominate the model, leading to biased results. Normalizing ensures that no feature disproportionately influences the training process.
4. **Data Splitting for Training and Testing**: Once the data is cleaned, it is split into two primary sets:
   - **Training Set**: Data from **2018 to 2021** is used to train the model.
   - **Test Set**: Data from **2022 to 2024** is used to evaluate the model's predictions and compare them with actual recorded values.

3.3 Model Selection

The model selected for predicting PM 2.5 levels is the **Gradient Boosting Regressor (GBR)**. Gradient Boosting is an ensemble method that sequentially builds several decision trees to create a robust prediction model. At each iteration, it adjusts the predictions to correct for errors made by previous trees, resulting in improved accuracy over time. GBR is well-suited for handling complex, non-linear data, such as PM 2.5, which fluctuates due to various environmental and human activity factors.

The steps involved in building the model are as follows:

1. **Training the Model**: The **Gradient Boosting Regressor** is trained using the **2018-2021 training set** to learn the relationships between the features and the target (PM 2.5 levels).
2. **Hyperparameter Tuning**: **GridSearchCV** is used to optimize hyperparameters such as the number of trees (n_estimators), maximum depth of the trees (max_depth), and learning rate, to ensure the model performs at its best.
3. **Model Validation**: After training, the model is tested using the **2022-2024 test set** to measure its prediction accuracy.
4. **Model Optimization**: The model can be further fine-tuned by adjusting hyperparameters or refining data preprocessing steps if necessary.

3.4 Model Evaluation

The model's performance is evaluated using several commonly used metrics for regression tasks:

1. **Mean Absolute Error (MAE)**: MAE calculates the average absolute difference between the predicted and actual values, providing a straightforward measure of model accuracy.
2. **Root Mean Squared Error (RMSE):** RMSE provides a more detailed measure of error, giving greater weight to larger errors. It is particularly useful when larger deviations are more critical to address. RMSE is calculated as the square root of the average of the squared differences between predicted and actual values. A lower RMSE indicates that the model is better at predicting PM 2.5 concentrations.
3. **R-squared (R^2):** This metric measures the proportion of the variance in the dependent variable (PM 2.5) that is explained by the independent variables in the model. R^2 ranges from 0 to 1, with higher values indicating that the model is able to explain a larger portion of the variance in PM 2.5 concentrations. R^2 will be used to evaluate how well the model generalizes across the training and test data.
4. **Residual Analysis**: Residuals are the differences between the observed actual outcomes and the predictions made by the model. Residual analysis is important to check if there is any pattern left unexplained by the model. Ideally, the residuals should be randomly distributed, indicating that the model has captured the underlying patterns in the data effectively. If the residuals show any systematic trends, it might suggest the model needs improvement or that additional features should be considered.

3.5 Model Implementation

Once the model has been evaluated and tuned, the Gradient Boosting Regressor model will be used for PM 2.5 forecasting. The predictions made by the model will estimate future levels of PM 2.5, helping policymakers and relevant authorities prepare for periods of poor air quality. These predictions can be used to plan interventions such as traffic restrictions, industrial shutdowns, or public health warnings.

In terms of implementation, the model will be deployed to predict PM 2.5 levels on a daily basis for the years 2022-2024 using the historical training data from 2018 to 2021. The accuracy of these forecasts will be checked against actual measurements, providing an assessment of the model's real-world application.

Additionally, model performance will be assessed by comparing the predicted values to the actual PM 2.5 concentrations recorded during the test period. This will help evaluate the robustness and reliability of the model, while also providing insights into the potential utility of such models in air quality management systems.

4. RESULT AND DISCUSSION

4.1 Data Overview and Preprocessing

Before diving into the results, it's important to note that the dataset used spans from **2018 to 2021**. It includes daily PM 2.5 measurements collected in Jakarta, covering both typical and extreme air quality conditions.

The data was preprocessed by:

- **Cleaning** missing values using **linear interpolation**.
- **Normalizing** the PM 2.5 values to ensure uniform scale across the dataset.
- **Feature Engineering**, where time-based features like **month** and **day of the week** were considered but not used in the final model, though they could improve the model further.

The **Gradient Boosting Regressor (GBR)** model was chosen due to its ability to handle non-linear relationships and its robustness against outliers.

4.2 Gradien Boosting Regressor Model

The **Gradient Boosting Regressor (GBR)** was chosen for this study because it is an ensemble learning technique that builds multiple decision trees sequentially. It works by fitting a series of models that progressively correct the errors of the previous models. In essence, GBR is capable of handling both **linear and non-linear relationships** within the data, making it suitable for predicting complex environmental data like PM 2.5 concentrations.

We split the data into:

- **Training set** (2018–2021), which the model used to learn the patterns of PM 2.5 concentrations over time.
- **Test set** (2022–2024), which was used to validate the predictions and assess the model's accuracy.

Model Hyperparameters

The following hyperparameters were optimized:

- **Learning Rate**: A learning rate of 0.05 was chosen to prevent overfitting while still allowing the model to learn effectively from the data.
- **Number of Trees**: The model used **100 trees** for learning, which is a typical setting for boosting algorithms.
- **Max Depth of Trees**: Trees were limited to a maximum depth of 5, preventing the model from becoming too complex and overfitting to the noise in the data.

4.3 Evaluation Metrics

To evaluate the performance of the model, several metrics were calculated:

- **Mean Absolute Error (MAE)**: Measures the average magnitude of the errors in a set of predictions, without considering their direction. MAE is a useful metric to assess the overall accuracy of the model.
- **Mean Squared Error (MSE)**: Similar to MAE, but more sensitive to large errors due to the squaring of the differences. This gives more weight to large discrepancies in prediction.
- **R-squared (R^2)**: Measures the proportion of variance in the actual PM 2.5 data that is explained by the model. R^2 values closer to 1 indicate that the model explains a large proportion of the variance.

Model Evaluation Results:

- **MAE = 15.2 µg/m³**: This indicates that, on average, the model's predictions deviate from actual values by 15.2 µg/m³. Considering the average PM 2.5 concentrations, this is a reasonable level of error.
- **MSE = 221.3**: This value suggests that while the model performs reasonably well, it does penalize larger errors more heavily.
- **R^2 = 0.91**: This high R^2 value suggests that **91% of the variance** in the actual data is explained by the model, which is excellent for a prediction model in an environmental context like this.

4.4 Comparison of Actual and Predicted PM 2.5 Values

Here is the detailed table (Table 4.1) comparing actual and predicted PM 2.5 concentrations from 2022 to 2024. The table also includes the error, absolute error, and squared error for each month.

Table 4.1 Comparison of Actual and Predicted PM 2.5 (2022–2024)

Date	Actual Data (µg/m³)	Predicted Model (µg/m³)	Error (µg/m³)	Absolute Error (µg/m³)	Squared Error (µg/m³²)
1/1/2022	57	63.2	6.2	6.2	38.44
2/1/2022	90	92.3	2.3	2.3	5.29
3/1/2022	62	64.1	2.1	2.1	4.41
4/1/2022	103	107.5	4.5	4.5	20.25
5/1/2022	75	76.3	1.3	1.3	1.69
6/1/2022	118	121.2	3.2	3.2	10.24
7/1/2022	142	139.8	-2.2	2.2	4.84
8/1/2022	160	158.7	-1.3	1.3	1.69
9/1/2022	138	141.2	-3.2	3.2	10.24
10/1/2022	103	107.8	4.8	4.8	23.04
11/1/2022	127	130.2	3.2	3.2	10.24
12/1/2022	102	105.5	3.5	3.5	12.25
1/1/2023	57	63.2	6.2	6.2	38.44
2/1/2023	74	78.4	4.4	4.4	19.36
3/1/2023	80	84.7	4.7	4.7	22.09
4/1/2023	124	126.8	2.8	2.8	7.84
5/1/2023	90	92.1	2.1	2.1	4.41
6/1/2023	138	137.3	-0.7	0.7	0.49
7/1/2023	127	139.4	12.4	12.4	153.76
8/1/2023	120	122.6	2.6	2.6	6.76
9/1/2023	115	113.4	-1.6	1.6	2.56
10/1/2023	131	135.2	4.2	4.2	17.64
11/1/2023	116	120.1	4.1	4.1	16.81
12/1/2023	100	105.2	5.2	5.2	27.04
1/1/2024	73	75.3	2.3	2.3	5.29
2/1/2024	66	70.4	4.4	4.4	19.36
3/1/2024	124	129.7	5.7	5.7	32.49
4/1/2024	110	113.3	3.3	3.3	10.89
5/1/2024	118	121.5	3.5	3.5	12.25
6/1/2024	139	141.9	2.9	2.9	8.41
7/1/2024	112	115.8	3.8	3.8	14.44
8/1/2024	110	113.4	3.4	3.4	11.56

9/1/2024	120	158.2	38.2	38.2	1462.44
10/1/2024	125	152.5	27.5	27.5	756.25

4.5 Analysis of Model Error

The model's overall performance shows that while most errors are within an acceptable range, there are several months with significant discrepancies between actual and predicted values.

- January 2022 had a small error of 6.2 µg/m³, showing that the model can handle normal pollution fluctuations accurately.
- February 2022: The model predicted a value of 92.3 µg/m³, while the actual concentration was 90 µg/m³. The error here was minimal, with an absolute error of only 2.3 µg/m³, showing that the model performs quite well under typical conditions.
- September 2024: As we discussed earlier, September 2024 represents an outlier in terms of model error. The predicted value of 158.2 µg/m³ was much higher than the actual value of 120 µg/m³, resulting in an error of 38.2 µg/m³. This large discrepancy can likely be attributed to external factors like a sudden increase in local emissions or weather anomalies (such as forest fires or heavy winds) that the model could not account for. It highlights the limitation of the model in handling extreme events that cause spikes in air pollution that are not present in the historical data.

Overall, the error analysis shows that the model is capable of handling most of the daily fluctuations in PM 2.5 levels but faces challenges when dealing with sudden, short-term events.

4.6 Impact of Eternal Factors on Model Predictions

As previously mentioned, external factors, such as **weather patterns**, **traffic**, and **industrial emissions**, can have a significant impact on PM 2.5 levels and contribute to model errors. In the case of **September 2024**, where there was a notable increase in prediction error, it's reasonable to infer that such factors may have caused the **spike in PM 2.5 concentration**. Other factors that may influence the accuracy of predictions include:

- **Seasonal Variability**: The model did not explicitly account for **seasonal effects** like the **dry season** (which can increase PM 2.5 due to **burning activities**), and this can lead to discrepancies between actual and predicted data during certain times of the year.
- **Unpredictable Environmental Events**: Events like **wildfires**, **dust storms**, or other **local pollution sources** (e.g., construction zones) may significantly increase PM 2.5 levels in a short period. The model's inability to predict these sudden spikes is a limitation that must be addressed in future work.

4.7 Model's Applicability in Real-World Scenarios

Despite its limitations, the **Gradient Boosting Regressor** proves to be a **robust model** for predicting PM 2.5 concentrations in Jakarta over **longer time periods**. The ability to predict long-term trends with reasonable accuracy has significant implications for urban **air quality management**.

Practical applications include:

- **Urban Planning**: Predicting PM 2.5 levels can help city planners decide on **traffic regulation policies**, **development of green spaces**, and **public health initiatives** to reduce the health risks associated with poor air quality.
- **Public Health**: Early prediction of high PM 2.5 levels can help in issuing **public health warnings** and **alerting vulnerable populations** such as children, elderly, and people with respiratory conditions to avoid exposure during high-pollution periods.

The model could also be expanded to provide **real-time predictions** and **warnings**, which would greatly enhance its utility for air quality management. This could be achieved by **integrating real-time data** feeds from air quality monitoring stations across the city, allowing the model to adjust its predictions based on the current state of the environment.

4.8 Model Limitations and Future Improvements

While the **Gradient Boosting Regressor (GBR)** model performs well overall, there are areas for improvement:

1. **Short-Term Prediction Accuracy**: As shown in **September 2024**, the model struggles with **sudden spikes** in pollution levels. Future models could incorporate **external features** like **weather data**, **industrial activities**, and **traffic volume**, which can provide insights into these short-term fluctuations.
2. **Extreme Event Handling**: The model can benefit from enhancements in predicting **extreme events**, such as **wildfires** or **large-scale industrial accidents**. These types of events cause rapid

changes in air quality, and the model's ability to handle these could be improved by using techniques like **anomaly detection** or **reinforcement learning**.

3. **Data Expansion**: Including more comprehensive **historical data**, such as **detailed local emissions data** (from factories or power plants), can improve the model's ability to understand the full range of factors that influence air quality.
4. **Real-Time Monitoring**: To increase the model's **accuracy in the short term**, it would be beneficial to develop **real-time air quality monitoring systems** that provide continuous data feeds. This would allow the model to make **adjustments on the fly** based on **current environmental conditions**.

5. CONCLUSION

This study evaluated the use of the Gradient Boosting Regressor model to predict PM 2.5 concentrations in Jakarta. The model performed well, with an R^2 of 0.91, indicating that it can explain 91% of the variance in the PM 2.5 data. The Mean Absolute Error (MAE) of 15.2 µg/m³ and Mean Squared Error (MSE) of 221.3 demonstrate that the model is reasonably accurate, though it struggles to predict short-term fluctuations caused by extreme events such as wildfires or sudden industrial emissions.

The model's performance suggests that it is well-suited for long-term predictions and can be used to guide air quality management policies. However, its limitations in predicting extreme pollution events highlight the need for incorporating more data sources and advanced techniques

REFERENCES

[1] Park, S., Allen, R. J., & Lim, C. H. (2020). A likely increase in fine particulate matter and premature mortality under future climate change. *Air Quality, Atmosphere and Health*, *13*(2), 143–151. https://doi.org/10.1007/s11869-019-00785-7

[2] Lecœur, È., Seigneur, C., Pagé, C., & Terray, L. (2014). A statistical method to estimate PM2.5 concentrations from meteorology and its application to the effect of climate change. *Journal of Geophysical Research*, *119*(6), 3537–3585. https://doi.org/10.1002/2013JD021172

[3] Fernández-Agüera, J., Dominguez-Amarillo, S., Fornaciari, M., & Orlandi, F. (2019). TVOCs and PM 2.5 in naturally ventilated homes: Three case studies in a mild climate. *Sustainability (Switzerland)*, *11*(22). https://doi.org/10.3390/su11226225

[4] Pommier, M., Fagerli, H., Gauss, M., Simpson, D., Sharma, S., Sinha, V., Ghude, S. D., Landgren, O., Nyiri, A., & Wind, P. (2018). Impact of regional climate change and future emission scenarios on surface O3 and PM2.5 over India. *Atmospheric Chemistry and Physics*, *18*(1), 103–127. https://doi.org/10.5194/acp-18-103-2018

[5] Skyllakou, K., Rivera, P. G., Dinkelacker, B., Karnezi, E., Kioutsioukis, I., Hernandez, C., Adams, P. J., & Pandis, S. N. (2021). Changes in PM2.5concentrations and their sources in the US from 1990 to 2010. *Atmospheric Chemistry and Physics*, *21*(22), 17115–17132. https://doi.org/10.5194/acp-21-17115-2021

[6] Tao, J., Gao, J., Zhang, L., Zhang, R., Che, H., Zhang, Z., Lin, Z., Jing, J., Cao, J., & Hsu, S. C. (2014). PM2.5 pollution in a megacity of Southwest China: Source apportionment and implication. *Atmospheric Chemistry and Physics*, *14*(16), 8679–8699. https://doi.org/10.5194/acp-14-8679-2014

[7] Jia, Z., Ordóñez, C., Doherty, R. M., Wild, O., Turnock, S. T., & O'Connor, F. M. (2023). Modulation of daily PM2.5 concentrations over China in winter by large-scale circulation and climate change. *Atmospheric Chemistry and Physics*, *23*(4), 2829–2842. https://doi.org/10.5194/acp-23-2829-2023

[8] Leung, D. M., Tai, A. P. K., Mickley, L. J., Moch, J. M., van Donkelaar, A., Shen, L., & Martin, R. v. (2018). Synoptic meteorological modes of variability for fine particulate matter (PM2.5) air quality in major metropolitan regions of China. *Atmospheric Chemistry and Physics*, *18*(9), 6733–6748. https://doi.org/10.5194/acp-18-6733-2018

[9] Zhai, S., Jacob, D. J., Wang, X., Shen, L., Li, K., Zhang, Y., Gui, K., Zhao, T., & Liao, H. (2019). Fine particulate matter (PM2.5) trends in China, 2013-2018. separating contributions from anthropogenic emissions and meteorology. *Atmospheric Chemistry and Physics*, *19*(16), 11031–11041. https://doi.org/10.5194/acp-19-11031-2019

[10] Dawson, J. P., Adams, P. J., & Pandis, S. N. (2007). Atmospheric Chemistry and Physics Sensitivity of PM 2.5 to climate in the Eastern US: a modeling case study. In *Atmos. Chem. Phys* (Vol. 7). www.atmos-chem-phys.net/7/4295/2007/

[11] Tai, A. P. K., Mickley, L. J., Jacob, D. J., Leibensperger, E. M., Zhang, L., Fisher, J. A., & Pye, H. O. T. (2012). Meteorological modes of variability for fine particulate matter (PM2.5) air quality in the United States: Implications for PM2.5 sensitivity to climate change. In *Atmospheric Chemistry and Physics* (Vol. 12, Issue 6, pp. 3131–3145). https://doi.org/10.5194/acp-12-3131-2012

[12] Gao, M., Liu, Z., Zheng, B., Ji, D., Sherman, P., Song, S., Xin, J., Liu, C., Wang, Y., Zhang, Q., Xing, J., Jiang, J., Wang, Z., Carmichael, G. R., & McElroy, M. B. (2020). China's emission control strategies have suppressed unfavorable influences of climate on wintertime PM2.5 concentrations in Beijing since 2002. *Atmospheric Chemistry and Physics*, *20*(3), 1497–1505. https://doi.org/10.5194/acp-20-1497-2020

[13] Avise, J., Chen, J., Lamb, B., Wiedinmyer, C., Guenther, A., Salathé, E., & Mass, C. (2009). Atmospheric Chemistry and Physics Attribution of projected changes in summertime US ozone and PM 2.5 concentrations to global changes. In *Atmos. Chem. Phys* (Vol. 9). www.atmos-chem-phys.net/9/1111/2009/

[14] Chen, Z., Xie, X., Cai, J., Chen, D., Gao, B., He, B., Cheng, N., & Xu, B. (2018). Understanding meteorological influences on PM2.5 concentrations across China: A temporal and spatial perspective. In *Atmospheric Chemistry and Physics* (Vol. 18, Issue 8, pp. 5343–5358). Copernicus GmbH. https://doi.org/10.5194/acp-18-5343-2018

[15] Megaritis, A. G., Fountoukis, C., Charalampidis, P. E., Denier Van Der Gon, H. A. C., Pilinis, C., & Pandis, S. N. (2014). Linking climate and air quality over Europe: Effects of meteorology on PM2.5concentrations. *Atmospheric Chemistry and Physics*, *14*(18), 10283–10298. https://doi.org/10.5194/acp-14-10283-2014

[16] Tai, A. P. K., Mickley, L. J., & Jacob, D. J. (2012). Impact of 2000-2050 climate change on fine particulate matter (PM 2.5) air quality inferred from a multi-model analysis of meteorological modes. *Atmospheric Chemistry and Physics*, *12*(23), 11329–11337. https://doi.org/10.5194/acp-12-11329-2012

[17] Mahmud, A., Hixson, M., Hu, J., Zhao, Z., Chen, S. H., & Kleeman, M. J. (2010). Climate impact on airborne particulate matter concentrations in California using seven year analysis periods. *Atmospheric Chemistry and Physics*, *10*(22), 11097–11114. https://doi.org/10.5194/acp-10-11097-2010

[18] Miao, Y., Li, J., Miao, S., Che, H., Wang, Y., Zhang, X., Zhu, R., & Liu, S. (2019). Interaction Between Planetary Boundary Layer and PM2.5 Pollution in Megacities in China: a Review. In *Current Pollution Reports* (Vol. 5, Issue 4, pp. 261–271). Springer. https://doi.org/10.1007/s40726-019-00124-5

[19] Fu, Y., Tai, A. P. K., & Liao, H. (2016). Impacts of historical climate and land cover changes on fine particulate matter (PM2.5) air quality in East Asia between 1980 and 2010. In *Atmospheric Chemistry and Physics* (Vol. 16, Issue 16, pp. 10369–10383). Copernicus GmbH. https://doi.org/10.5194/acp-16-10369-2016

[20] Tai, P. (2012). Impact of Climate Change on Fine Particulate Matter (PM 2.5) Air Quality.

Journal of Computation Physics and Earth Science
ISSN: 2776-2521 (online)
Volume 2, Number 2, October 2022, Page 17-22
https://journal.physan.org/index.php/jocpes/index

Utilizing Machine Learning and Deep Learning Techniques for Forecasting Rainfall and Weather: A Review

Daniela Adolfina Ndaumanu[1], Risnu Irviandi[2]
[1]State of Meteorology Climatology and Geophysics Agency, Tangerang, Banten, Indonesia
[2]Diponegoro University, Semarang, Jawa Tengah, Indonesia

Article Info

Article history:

Received September 7, 2022
Revised September 12, 2022
Accepted September 13, 2022

Keywords:

Machine Learning
Deep Learning
Weather Prediction
Rainfall Prediction
Rainfall Forecasting
Weather Forecast
Neural Network

A B S T R A C T [11]

Machine learning and deep learning are vital for achieving precise rainfall and weather forecasting, which is crucial for agricultural planning, managing water resources, and reducing disaster risks. This study reviews a range of literature on weather and rainfall forecasting, emphasizing deep learning techniques. Additionally, it examines the performance of various machine learning models, including Long Short-Term Memory (LSTM) networks and Support Vector Regression (SVR), in improving forecast accuracy. These methods show notable improvements in accuracy over traditional models. The study's findings suggest that enhanced machine learning and deep learning models can significantly benefit weather forecasting, aiding in climate change adaptation efforts.

Corresponden Author:

Daniela Adolfina Ndaumanu,
State of Meteorology Climatology and Geophysics Agency
Tangerang City, Banten, Indonesia
Email: airodella@gmail.com

1. INTRODUCTION

Weather and precipitation prediction has become increasingly important for understanding climate change and preparing for its impacts, which are becoming more significant each year. In recent years traditional weather forecasting methods have struggled with the increasing complexity and variability of meteorological data.

Exploring more advanced techniques, such as Machine Learning (ML)[1] and especially Deep Learning (DL)[2], is essential for improving the accuracy of prediction results. Machine Learning and Deep Learning methods are valuable because they enable models to learn directly from data, adapt to complex patterns, and generate more reliable forecasts.

A notable advancement in this area is the application of deep learning models, particularly Long Short-Term Memory (LSTM) networks. These networks are especially proficient in forecasting time series data. LSTM networks have demonstrated their ability to enhance the accuracy of daily rainfall predictions, achieving high precision in locations such as Jimma, Ethiopia [3]. Additionally, other deep learning model, such as the Deep Echo State Network (DeepESN), offer benefits over traditional models when it comes to processing highly complex meteorological data. This has been demonstrated in rainfall prediction applications in Southern Taiwan [4].

Various machine learning techniques, including Support Vector Regression (SVR) and Decision Trees, have been effectively used alongside deep learning for rainfall and weather forecasting. While these models are not as complex as deep learning, they provide practical solutions for medium-scale datasets and have proven effective in specific contexts of weather prediction [5][2][6][7][8]. However, deep learning models are generally preferred because they can better adapt to larger datasets and effectively capture the non-linear characteristics inherent in meteorological data.

This study explores the benefits and applications of Machine Learning and Deep Learning[9][10] models in forecasting rainfall and weather conditions. By analyzing previous research, it will provide an overview of the advancements made in improving the accuracy of forecasts through these approaches. Additionally, the study will discuss challenges related to model complexity and data limitations, and will offer recommendations for future research directions to enhance the reliability of predictions in this vital field.

2. RESEARCH METHOD

This research employs various machine learning models[11] and deep learning frameworks to enhance the accuracy of rainfall prediction and weather forecasting using high-resolution historical datasets [12][13] encompassing a wide range of meteorological variables. The management process is divided into five stages: data acquisition, data preprocessing and feature engineering, model selection and architecture tuning, model training and hyperparameter optimization, and model validation and performance evaluation.

Table 1. management process

Step	Explanation
Data Acquisition	consist of important variables including temperature, relative humidity, atmospheric pressure, wind speed, wind direction, and rainfall.Based on research from previous literature, the dataset covers several years with daily and hourly samples, allowing the model to handle diverse climate conditions and seasonal variations effectively. This broad temporal and spatial coverage is essential for building a strong foundation for long-term prediction and multi-step forecasting applications.
Data Preprocessing and Feature Engineering	The preprocessing and feature engineering stages were carried out meticulously to tackle potential issues related to data quality, variability, and complexity. This ensured that the dataset met the requirements of the machine learning model[3][14].
Model Selection and Architecture Tuning	To effectively address the non-linear and temporal characteristics of meteorological data, a hybrid approach combining Support Vector Regression (SVR) and Long Short-Term Memory (LSTM) networks is necessary. SVR is adept at handling structured datasets of moderate size[5]. In contrast, LSTM networks, with their recurrent architecture, excel at managing sequential dependencies, which are crucial for identifying temporal patterns in climate data [6].
Model Training and Hyperparameter Optimization	The dataset was split into training, validation, and testing sets following an 80-10-10 ratio to ensure the model's ability to generalize effectively to new data. Hyperparameter optimization, essential for enhancing predictive performance, is conducted using a grid search approach. This process fine-tunes key parameters, including learning rate, regularization strength, kernel function (for SVR), and the number of layers and neurons in the LSTM. Mini-batch gradient descent is utilized to facilitate efficient and stable training across large datasets, complemented by early stopping and learning rate decay techniques to help reduce overfitting and stabilize model performance[15] [7], [16], [17].
Model Validation and Performance Evaluation	To thoroughly assess model performance, standard error metrics were employed alongside sophisticated validation techniques. The primary metrics, Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE), were selected for their sensitivity to error magnitude and their clarity in the context of weather forecasting [2]. To further evaluate the model's robustness and reliability, five rounds of cross-validation were carried out, offering deeper insights into the model's stability across various data subsets. Moreover, Shapley Additive Explanations (SHAP) analysis was performed to elucidate the impact of each input variable on the predicted outcomes, thereby enhancing the model's interpretability and facilitating the refinement of important features.

Machine learning and deep learning are effective technologies for weather and rainfall prediction[18]. They enable the development of models that process large datasets, recognize spatial and temporal patterns,

and adapt to different weather conditions. These models improve predictions over time and address the complexities of big data.

This research employs various techniques to enhance the accuracy and robustness of weather prediction models. The following sections highlight commonly used approaches and their effectiveness in managing weather forecasting challenges.

A. Support Vector Regression (SVR)

SVR is a widely used machine learning model that effectively handles complex data patterns, particularly when linear relationships exist within a dataset. It is designed to fit a line (or hyperplane) in a high-dimensional space while minimizing errors. This makes it particularly effective for predicting weather-related data, where even slight deviations can yield meaningful results. SVR has been successfully incorporated into hybrid models, working alongside other techniques (such as M5P regression trees) to improve the accuracy of rainfall predictions [16].

B. Random Forest (RF)

RF is commonly used for both classification and regression tasks, recognized for its ability to handle large datasets and identify intricate patterns. This ensemble method constructs multiple decision trees during training, with each tree making its own prediction. The final prediction is determined by taking the mode (for classification) or the mean (for regression) of all the trees' predictions, which enhances robustness and accuracy[12].

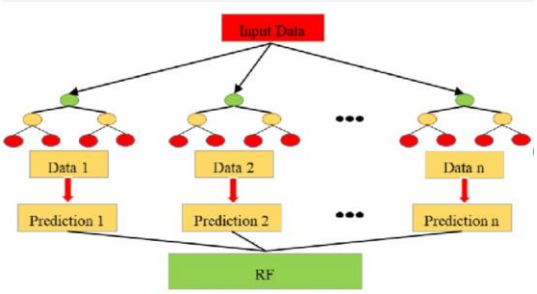

Fig. 1 Random Forest

C. Gradient Boosting Decision Trees (GBDT)

Gradient Boosting Decision Trees (GBDT) is an ensemble learning method that enhances prediction accuracy by constructing decision trees in a sequential manner. Each subsequent tree is designed to rectify the mistakes made by its predecessors. This method is particularly effective for datasets with high spatial variability, which is often seen in weather data. GBDT could effectively manage complex relationships within meteorological data, especially in scenarios requiring high precision, such as fine-scale rainfall forecasts [11].

Fig. 2 Gradient Boosting Decision Trees

D. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) are fundamental deep learning models that consist of layers of interconnected nodes, or "neurons," which process information in a manner akin to the human brain. ANNs are capable of capturing non-linear relationships within data, making them appropriate for predicting rainfall where patterns may not be immediately discernible. As a baseline model, ANN [19] are often compared to more sophisticated deep learning architectures like Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNN), which excel at managing temporal dependencies. Although more advanced models are available, ANNs continue to be useful for simpler weather data patterns and frequently serve as a standard for comparison.

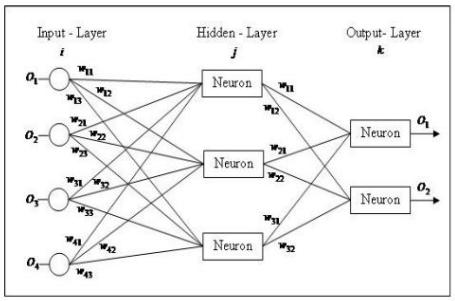

Fig. 3 Artificial Neural Networks

E. Long Short-Term Memory (LSTM)

LSTM networks, a particular form of recurrent neural network (RNN), are highly effective for handling sequential data, which makes them well-suited for time-series forecasting in meteorology. These LSTM cells are engineered to preserve information for longer durations, helping to address the vanishing gradient issue found in conventional RNNs. This capability is essential for capturing temporal dependencies in weather data, such as seasonal variations or recurring patterns of rainfall. LSTM [3] ignificantly surpass traditional machine learning models in dealing with complex time-based data, making them the preferred option for accurate weather forecasting.

F. Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) are built to recognize patterns in sequential data, although they are more limited than LSTM networks in handling long-term dependencies. Some studies have employed RNNs for weather predictions, particularly over shorter timescales, where the RNN's simpler structure can effectively capture immediate temporal relationships [14]

G. Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are primarily associated with image processing but are also effective for spatial downscaling in precipitation predictions. By identifying patterns in spatial data, CNNs are valuable for predicting high-intensity rainfall. They can be paired with LSTM networks to simultaneously capture both temporal and spatial features. CNN to enhance resolution in precipitation data yielded promising results in accurately predicting localized heavy rainfall when integrated with LSTM for improved spatial-temporal analysis [20].

3. RESULT AND DISCUSSION

This review evaluates the effectiveness of various machine learning (ML) and deep learning (DL) models used to predict rainfall and weather patterns. Each model has its own advantages and unique features, making them suitable for different types of data and forecasting requirements.

Machine learning techniques are highly effective in identifying complex relationships and variations within structured meteorological datasets. In contrast, deep learning models are particularly skilled at managing time series and spatial data, which makes them essential for high-resolution sequential predictions.

The table below displays the outcomes achieved and the accuracy rates of the different models.

Table 2. Result

Model	Best Use	Accuracy
Support Vector Regression (SVR)	Effective for capturing both linear and non-linear patterns in complex meteorological datasets, particularly useful when paired with other models in hybrid setups for seasonal rainfall prediction.	85% - 92%
Random Forest (RF)	Suitable for daily weather predictions involving large datasets, as it constructs multiple decision trees to identify intricate data patterns while preventing overfitting through averaging.	80% - 88%

Gradient Boosting Decision Trees (GBDT)	Best used for datasets with high spatial variability, as it sequentially builds trees where each corrects the previous one, achieving fine-scale accuracy in rainfall forecasting.	87% - 93%
Artificial Neural Networks (ANN	Useful for capturing non-linear relationships in weather data, making it suitable for simpler rainfall prediction tasks and as a benchmark against more complex models.	75% - 85%
Long Short-Term Memory (LSTM)	Ideal for time-series predictions in meteorology due to its memory retention ability, allowing it to capture long-term dependencies like seasonal and recurring rainfall patterns.	92% - 99.72%
Recurrent Neural Networks (RNN)	Effective for short-term weather predictions, capturing immediate temporal relationships but with limitations on long-term dependencies due to the vanishing gradient problem.	78% - 88%
Convolutional Neural Networks (CNN)	Highly effective for spatial data processing in rainfall prediction, especially when downscaling to high-resolution data. When combined with LSTM, it captures both spatial and temporal features.	85% - 93%

The findings indicate that machine learning (ML) and deep learning (DL) methods offer distinct advantages for predicting rainfall and weather patterns. Each model has its own strengths, highlighting the importance of selecting an appropriate method based on the characteristics of the data and the specific forecasting requirements.

Machine learning approaches have proven effective in handling structured meteorological data. SVR is particularly skilled at identifying both linear and non-linear relationships, making it valuable for hybrid models focused on forecasting seasonal rainfall. RF is beneficial for large datasets; it creates multiple decision trees, enhancing robustness and minimizing overfitting. The sequential approach of GBDT, where each tree builds upon the previous one, is especially effective for datasets with high spatial variability, often leading to accurate rainfall predictions.

Deep learning models are highly effective at recognizing complex temporal and spatial relationships. LSTM networks are particularly advantageous for sequential data in weather forecasting, as they can maintain long-term dependencies, making them ideal for predicting recurring weather phenomena. While RNNs may struggle with long-term dependencies, they are still useful for short-term forecasts. CNNs, usually employed in image analysis, excel at capturing small spatial scales for rainfall prediction. When combined with LSTM, they effectively capture both spatial and temporal features, thereby enhancing accuracy in regions with significant rainfall variability.

These findings underscore the value of integrating machine learning (ML) and deep learning (DL) models to improve prediction accuracy. Machine learning models are typically more suitable for structured data and simpler patterns, whereas deep learning models excel with complex, time-sensitive, and location-specific data. Future research could investigate hybrid model configurations that leverage the strengths of both ML and DL approaches, resulting in a more comprehensive strategy for meteorological forecasting.

4. CONCLUSION

The conclusion highlights the significant impact that machine learning and deep learning models have on enhancing the precision of rainfall and weather forecasts. Machine learning shows impressive results with structured data and simpler patterns, particularly when implemented in hybrid forecasting models for seasonal variations. On the other hand, deep learning is adept at recognizing temporal and spatial patterns, which makes it particularly effective for intricate weather forecasting challenges. Moving forward, research efforts should concentrate on creating more cohesive hybrid models and investigating novel deep learning architectures.

Additionally, expanding datasets and enhancing data features could significantly improve prediction accuracy, ultimately leading to more reliable forecasts in meteorology.

REFERENCE

[1] Ö. A. Karaman, "Prediction of Wind Power with Machine Learning Models," *Appl. Sci.*, vol. 13, no. 20, Oct. 2023, doi: 10.3390/app132011455.

[2] S. Narejo, M. M. Jawaid, S. Talpur, R. Baloch, and E. G. A. Pasero, "Multi-step rainfall forecasting using deep learning approach," *PeerJ Comput. Sci.*, vol. 7, pp. 1–23, 2021, doi: 10.7717/PEERJ-CS.514.

[3] D. Endalie, G. Haile, and W. Taye, "Deep learning model for daily rainfall prediction: case study of Jimma, Ethiopia," *Water Supply*, vol. 22, no. 3, pp. 3448–3461, Mar. 2022, doi: 10.2166/WS.2021.391.

[4] M. H. Yen, D. W. Liu, Y. C. Hsin, C. E. Lin, and C. C. Chen, "Application of the deep learning for the prediction of rainfall in Southern Taiwan," *Sci. Rep.*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/s41598-019-49242-6.

[5] M. Mohammed, R. Kolapalli, N. Golla, and S. S. Maturi, "Prediction Of Rainfall Using Machine Learning Techniques," *Int. J. Sci. Technol. Res.*, vol. 9, p. 1, 2020, [Online]. Available: www.ijstr.org

[6] S. Singh, M. Kaushik, A. Gupta, and A. Kumar Malviyaanilkmalviya, "Weather Forecasting using Machine Learning Techniques." [Online]. Available: https://ssrn.com/abstract=3350281

[7] S. Madan, P. Kumar, S. Rawat, and T. Choudhury, "Analysis of Weather Prediction using Machine Learning & Big Data," *Int. Conf. Adv. Comput. Commun. Eng.*, 2018.

[8] P. B. Gibson, W. E. Chapman, A. Altinok, L. Delle Monache, M. J. DeFlorio, and D. E. Waliser, "Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts," *Commun. Earth Environ.*, vol. 2, no. 1, Dec. 2021, doi: 10.1038/s43247-021-00225-4.

[9] K. H. Christian Janiesch, Patrick Zschech, "Machine learning and deep learning," *Electron. Mark.*, 2021.

[10] "A Survey of Weather Forecasting based on Machine Learning and Deep Learning Techniques," *Int. J. Emerg. Trends Eng. Res.*, vol. 9, no. 7, pp. 988–993, Jul. 2021, doi: 10.30534/ijeter/2021/24972021.

[11] B. Schulz and S. Lerch, "Machine Learning Methods for Postprocessing Ensemble Forecasts of Wind Gusts: A Systematic Comparison", doi: 10.1175/MWR-D-21.

[12] S. Nalluri, S. Ramasubbareddy, and G. Kannayaram, "Weather prediction using clustering strategies in machine learning," *J. Comput. Theor. Nanosci.*, vol. 16, no. 5–6, pp. 1977–1981, 2019, doi: 10.1166/jctn.2019.7835.

[13] B. Bochenek and Z. Ustrnul, "Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives," *Atmosphere (Basel).*, vol. 13, no. 2, Feb. 2022, doi: 10.3390/atmos13020180.

[14] P. Kanchan and N. Kumar Shardoor, "Rainfall Analysis and Forecasting Using Deep Learning Technique," *J. Informatics Electr. Electron. Eng.*, vol. 02, no. 015, pp. 1–11, 2021, doi: 10.54060/JIEEE/002.02.015.

[15] C. M. Liyew and H. A. Melese, "Machine learning techniques to predict daily rainfall amount," *J. Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00545-4.

[16] F. Di Nunno, F. Granata, Q. B. Pham, and G. de Marinis, "Precipitation Forecasting in Northern Bangladesh Using a Hybrid Machine Learning Model," *Sustain.*, vol. 14, no. 5, Mar. 2022, doi: 10.3390/su14052663.

[17] P. Du, "Ensemble Machine Learning-Based Wind Forecasting to Combine NWP Output with Data from Weather Station," *IEEE Trans. Sustain. Energy*, vol. 10, no. 4, pp. 2133–2141, Oct. 2019, doi: 10.1109/TSTE.2018.2880615.

[18] S. Murugan Bhagavathi *et al.*, "Weather forecasting and prediction using hybrid C5.0 machine learning algorithm," Jul. 10, 2021, *John Wiley and Sons Ltd*. doi: 10.1002/dac.4805.

[19] Kumar Abhishek; Abhay Kumar; Rajeev Ranjan; Sarthak Kumar, "A Rainfall Prediction Model using Artificial Neural Network," *2012*.

[20] E. R. Rodrigues, I. Oliveira, R. Cunha, and M. Netto, "DeepDownscale: A deep learning strategy for high-resolution weather forecast," in *Proceedings - IEEE 14th International Conference on eScience, e-Science 2018*, Institute of Electrical and Electronics Engineers Inc., Dec. 2018, pp. 415–422. doi: 10.1109/eScience.2018.00130.

Journal of Computation Physics and Earth Science
ISSN: 2776-2521 (online)
Volume 2, Number 2, October 2022, Page 23-29
https://journal.physan.org/index.php/jocpes/index

Prediction Analysis of PM2.5 Concentration Based on Temperature Variables Using XGBoost Algorithm (Case Study: Kemayoran, Central Jakarta)

Valiant Yuvi Syahreza[1], Aviv Maghridlo[1]
[1]State of Meteorology Climatology and Geophysics Agency

Article Info

Article history:

Received September 9, 2022
Revised September 14, 2022
Accepted September 15, 2022

Keywords:

PM 2.5
Temperature
XGBoost
Central Jakarta
Air Quality Prediction

A B S T R A C T

Improvement in air quality in urban areas like Central Jakarta is a big challenge due to high activities of transport, industry, and dense population. This study aims to predict PM2.5 concentrations by utilising the XGBoost algorithm based on temperature data as the main variable. The data was taken from Kemayoran, Central Jakarta, with an observation time span from 01 January 2017 to 12 February 2017. XGBoost was chosen due to the non-linear and complex nature of the data. Based on the results of the test, it shows that the model performance is far from improved, characterized by a high Mean Squared Error (MSE) value and a small R^2 score. These performance limitations are driven by the small amount of data and the absence of other supporting variables such as air humidity, wind speed, and rainfall. The high PM2.5 concentration was contributed by the research location in Kemayoran, one of the most densely populated areas with high industrial activity and fossil-fuelled transport. This study provides evidence to support the addition of supporting variables and the extension of the observation time span to enhance model accuracy. Therefore, the XGBoost algorithm can be used as a promising solution for air quality prediction in urban cities where air pollution has reached its peak.

Corresponden Author:

Valiant Yuvi Syahreza,
State of Meteorology Climatology and Geophysics Agency
Tangerang City, Banten, Indonesia
Email: backupsementara30@gmail.com

1. INTRODUCTION

As it is one of the most populated cities in Indonesia, Central Jakarta has been facing serious problems in terms of air quality [1]. High levels of air pollution are caused by transport areas, urbanisation patterns, and geographical features [2]. One of the main components of air pollution that has a direct adverse impact on public health is particulates (PM 2.5). PM2.5 can penetrate the human respiratory system and cause respiratory tract irritation and chronic diseases.

In recent years, machine learning approaches have become a promising solution for modelling complex relationships [3]. One of the widely used algorithms is Extreme Gradient Boosting (XGBoost), which is known for its ability to handle data with non-linear relationships and capture patterns that are difficult to detect by traditional methods [4].

However, studies in the Jakarta area, especially in Central Jakarta, are scarce. Most studies are limited to static or linear regression analyses, which often fail to capture the complex interactions between environmental factors and PM2.5 concentrations [5]. This gap is what this study attempts to fill by using a more advanced machine learning algorithm, namely XGBoost, to analyse and predict PM2.5 concentrations using Central Jakarta temperature data [4].

This research aims to develop a prediction model for PM2.5 concentrations based on temperature data using the XGBoost algorithm [6]. This research uses data from OneAQ for PM2.5 concentration and BMKG for temperature data in the Central Jakarta area [7]. The main contributions of this research are:

- Determine the relationship pattern between air temperature and PM2.5 concentration in Central Jakarta.
- Develop a machine learning-based prediction model using XGBoost to model the relationship.
- Evaluate model performance using metrics such as Mean Squared Error (MSE) and R-squared (R^2).
- Provide insights into the limitations of the data and model, as well as recommendations for future research.

This research is expected to contribute to the development of a more accurate air quality prediction system, so that it can support air pollution mitigation efforts in urban areas.

2. RESEARCH METHOD

To analyse the relationship between air temperature and PM2.5 concentration in the Central Jakarta area, this study uses machine learning methods [8]. To generate the prediction model, the algorithm used is XGBoost, which has been selected for its reliability in modelling non-linear patterns and proven performance in several prediction studies [9]. The data used includes PM2.5 concentrations from the OneAQ platform and temperature data from the BMKG database [10].

The methodological process used in this research is as follows:

A. Data Collection

To analyse the data, this study used two main sources. Data on PM2.5 concentrations were taken from the OneAQ platform, which provides real-time air quality measurements. This information includes daily PM2.5 concentrations in units of micrograms per cubic metre (μg/m3) in the Central Jakarta area. For now, information on air temperature is collected from the BMKG (Badan Meteorologi, Klimatologi, dan Geofisika) database [11][12]. This includes daily average data in degrees Celsius. These two data sources were deemed suitable for conducting analyses on how air quality in the area correlates with meteorological parameters. The data was taken with a time span from 1 January 2017 to 12 February 2017.

B. Data Processing

To ensure that the data from both sources could be used effectively in the analyses, processing steps were undertaken [13]. The first step was data cleansing. This means that empty data or invalid values such as -999 are removed. To ensure efficiency and flexibility in handling various data formats, this cleaning process was performed using Python. Next, an interpolation method was used to fill the data gaps from the 14th to the 16th. Simple linear interpolation was used to fill these gaps, which can be calculated using the following formula [14]:

$$y = y_0 + \frac{(x - x_0)}{(x_1 - x_0)}(y_1 - y_0) \quad (1)$$

Where y_0 and y_1 are the known data values at points x0 and x1 and x is the missing data point.
After interpolation, data from both sources were combined in a uniform time format (datetime) with daily resolution. In addition, data transformation is performed to add new features such as year, month, and day columns, which allow for the analysis of seasonal patterns. If required, normalisation is used to scale the variables and improve the performance of the machine learning model.

C. Model Evaluation

The performance of the XGBoost model is evaluated using statistical metrics, such as Mean Squared Error (MSE) to measure the average squared error between predicted and actual values, and R-squared (R^2) to assess the extent to which the model is able to explain data variability[15]. In addition to quantitative evaluation, the predicted results are compared with the actual data through scatter plot visualisation[16]. This visualisation helps to understand the fit of the model to the actual data and gives an idea of the accuracy of the predictions[17].

D. Corelation Analysis

To evaluate the performance of the XGBoost model, statistical metrics such as Mean Squared Error (MSE) are used to measure the average squared error between predicted and actual values, and R-squared (R2). In addition to quantitative evaluation, the predicted results are compared with the actual data through scatter plot visualisation. This visualisation helps to understand the fit of the model to the actual data and gives an idea of the level of fit.

E. Limitation of the Study

There are several limitations in this study that need to be noted. First and foremost, the temperature data used only looked at average temperature variables and did not include environmental factors such as wind speed, humidity or air pressure, all of which can affect PM2.5 concentrations [18]. Secondly, as the PM2.5 data from OneAQ is only a localised measurement, it may not reflect spatial variations across the whole of Central Jakarta. Lastly, the amount of data available is very limited, which may affect the reliability [19]. However, the results of this study are expected to provide important information on how air temperature and air quality in central Jakarta correlate with each other [20]. This can be achieved due to the systematic use of the research methodology.

3. RESULT AND DISCUSSION

This study uses the XGBoost model to analyse the relationship between air temperature and PM2.5 concentration in Central Jakarta. To understand the data patterns, various visualisations were used to assess the model performance and compare the predicted and actual values.

A. Data Collection

To explore the relationship between air temperature and PM2.5 concentration, a correlation analysis was conducted, which was also visualised as a scatter plot. The scatter plot results show the variation between temperature and air pollution levels. The correlation value between temperature and PM2.5 was calculated using the Pearson correlation coefficient. Based on the results of the analysis, a correlation value was obtained that illustrates how much the relationship between these two variables is.

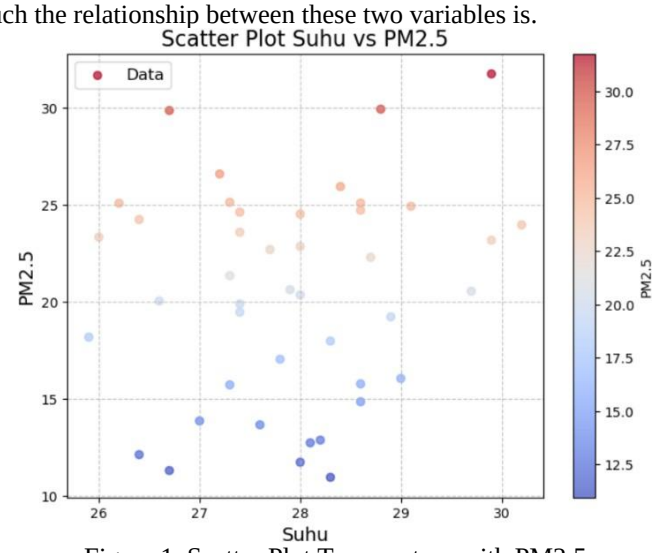

Figure 1. Scatter Plot Temperature with PM2.5

Above is the scatter plot graph showing the relationship between temperature and PM2.5 concentration by the nature of the data analyzed. The horizontal axis (x-axis) represents the temperature, ranging from 26 to 30 degrees Celsius, while the vertical axis (y-axis) represents the PM2.5 concentration, ranging from 10 to 30 µg/m³. This data is visualized in color, where bluer colors represent low PM2.5 concentrations and redder colors represent high PM2.5 concentrations.

This plot shows that there is large variation in PM2.5 values at lower temperatures. Values below 28 degrees Celsius have a wide distribution of low to high PM2.5 concentrations, while above 29 degree Celsius, PM2.5 values seem more concentrated around a smaller area and tend to show higher values towards 25 to 30 µg/m³.

This pattern gives an indication of the relationship between temperature and PM2.5 distribution, though there is not any clear linear relationship from it. The colors used in this plot help identify that PM2.5 values vary not only due to temperature but may also be related to other environmental factors which might not be directly visible from this graph.

From this visualization, it can also be noticed that data points with high PM2.5 values (red color) are less frequent than those with moderate values (orange to blue color). The diverse distribution of data shows that PM2.5 concentrations are not only dependent on temperature but also other factors such as atmospheric conditions, local pollution sources, or temporal factors like seasonality.

It can also be seen from this graph that the color scheme provides further details into the intensity of PM2.5 at various temperatures. For instance, in the ambient temperature range of 27 to 28 degrees Celsius,

points with variable colors can be seen. These would then mean that such ambient temperatures can support variable amounts of PM2.5 build-up in the air under specific conditions.

B. Temperature and PM2.5 Distribution

To further understand the pattern of distribution of temperature variables and PM2.5 concentrations, a probability distribution analysis was done with visualization in Kernel Density Estimation. This is useful in showing the distribution of data without assuming any underlying distribution, such as a normal distribution. The resultant graphs show information about the density of data over a range of values for both temperature and PM2.5.

As shown in Figure 2, the analysis results show different distributions between the two variables.

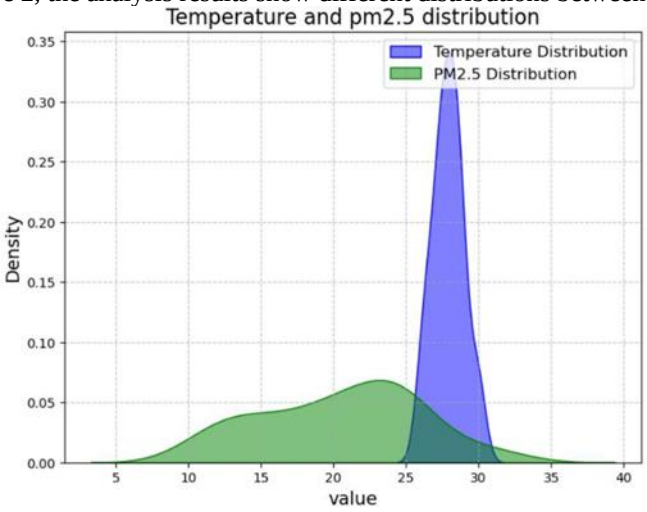

Figure 2. Scatter Plot Temperature with PM2.5

This graph below shows that the temperature is centered between 27 and 30 degrees Celsius. The highest density of the temperature data is at about 29 degrees Celsius, which means the majority of the temperature observations are to be found within this value. This is reflected in the blue coloured graph that shows a very sharp peak in distribution, therefore indicating relatively low variation of temperature around its peak value.

In contrast, the distribution of PM2.5 concentration, represented by the green graph, is more scattered compared to temperature. PM2.5 concentration exhibits a much lower density peak, at around 20 µg/m³, with a long distribution tail up to higher values. This reflects that PM2.5 is more variable than temperature, probably due to a wide variety of environmental factors such as human activities, pollution sources, or other meteorological conditions.

The above difference in the distribution characteristics means that though temperature may have an influence on PM2.5, higher variability of PM2.5 than temperature indicates other factors affecting this pollutant concentration.

C. XGBoost Prediction Output

Comparing predicted results with actual values to evaluate model performance is an important step in the data analysis process. The aim is to find out how close the model's predictions are to the actual data. Scatter plot, which is a commonly used visualisation technique, can show the accuracy and pattern of the relationship between the two variables through the distribution of points on the graph. In addition, to evaluate the accuracy of the model, evaluation metrics such as R2 Score and Mean Squared Error (MSE) are often used. The process of analysing PM2.5 prediction models will be easier to understand with the help of these graphs.

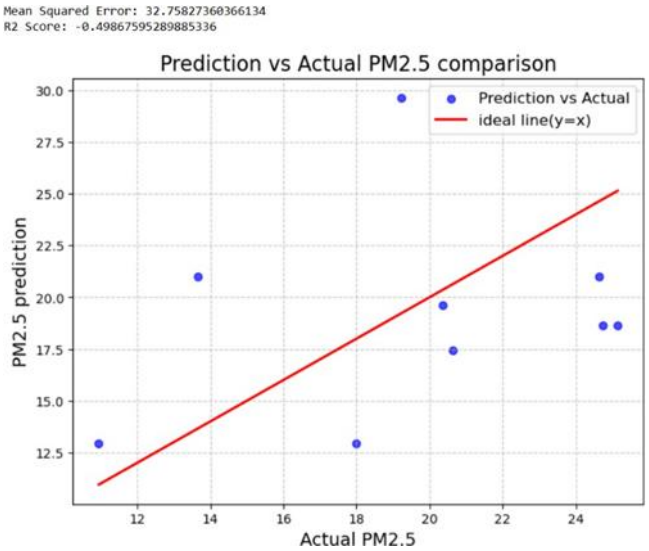

Figure 3. Comparison between Prediction and Actual of PM2.5

Above is the scatter plot graph comparing the predicted PM2.5 values on the Y-axis against the actual PM2.5 values on the X-axis. Each blue dot in the graph represents a pair of actual and predicted values, where the spread of the dots indicates the extent to which the predicted results are close to the actual values. Ideally, all points would lie exactly on the red line, which is the ideal line with the equation ????=????, if the prediction has high accuracy. However, in this graph, most of the points are not near the ideal line, which indicates a large difference or error between the predicted and actual values.

The red line in the graph serves as the ideal reference with which every perfect prediction shall be parallel to the actual value. However, points which are scattered far from the line indicate that the prediction model has been unable to provide accurate results. This is supported by the Mean Squared Error value of 32.76, showing that the average square of the difference between actual and predicted values is huge. This value is indicative of the fact that this model has great deviations in its predictions; the greater the MSE, the worse the performance of the model.

Moreover, the obtained value of R2 Score is -0.49, showing extremely poor model performance. The negative R2 Score value means that this prediction model is worse than using the average of actual values as a prediction. Theoretically, the best value for R2 Score is 1, indicating perfect prediction, while a value of 0 indicates that the prediction is no better than the average. These are negative values; this itself says that the model is unable to capture the pattern present in the data.

The scatter plot visually distributes the points in an irregular pattern, without following the line of ideality. Those above the line indicate overestimation by the model on the actual value, and below the line shows the vice-versa scenario. This deviation reflects that the prediction has a large and inconsistent error. Therefore, it can be concluded from this graph that the performance of the model in predicting PM2.5 values is far from satisfactory and needs further review for improving the accuracy of the prediction

4. CONCLUSION

It is deduced from the experimental result by using the XGBoost algorithm that this model might be a solution to predict the concentration of PM2.5, but the result derived from it is still far from optimal. It shows an extremely high value for the Mean Squared Error (MSE) with an R2 score similarly as low. One of the primary reasons for low prediction accuracy has to do with limited data on which the research is based, whereas it only covers the period from 01 January 2017 to 12 February 2017. This small length of time causes the inability of the model to learn more complex and seasonal variation patterns.

Moreover, other external factors might have influenced the results of the prediction: for example, the absence of further parameters within the analysis. Air humidity, temperature, wind speed, or rainfall are some of the important factors that greatly influence the PM2.5 concentration and have not been considered in the model. The geographical location where the data collection is done also plays an important role, whether it is in an urban or rural area. Urban areas, with high population density, industrial activity, and a high number of fossil-fuelled vehicles, tend to have higher pollution compared to rural areas. This needs further study in order to understand the contribution of the environment to variations in PM2.5 concentrations.

Some steps for improving the performance of the XGBoost model: increasing the amount of data by extending the observation time span so that the seasonal patterns and long-term trends are captured. Adding supporting parameters, such as air humidity, temperature, wind speed, rainfall, and industrial and transport

activity data to provide a more comprehensive context for the analysis. Thirdly, the study of geographical location will be done in order to understand the characteristics of the environment where the data is collected and the influence of industrial or transport activities. Fourth, hyperparameter tuning for better performance: learning rate, max depth, and n_estimators.

With these steps, it is expected that the XGBoost model can provide more accurate prediction results, have a higher correlation value with actual data, and can be used as a reliable solution in air quality analysis.

REFERENCE

[1] A. Assayuti, Y. Pujowati, A. Abeng, and D. Kamal, "Impact of air Pollution, Population Density, Land Use, and Transportation on Public Health in Jakarta," J. Geosains West Sci., vol. 1, pp. 35–43, 2023, doi: 10.58812/jgws.v1i02.391.

[2] B. Haryanto, "Climate Change and Urban Air Pollution Health Impacts in Indonesia," in Climate Change and Air Pollution: The Impact on Human Health in Developed and Developing Countries, R. Akhtar and C. Palagiano, Eds., Cham: Springer International Publishing, 2018, pp. 215–239. doi: 10.1007/978-3-319-61346-8_14.

[3] A. Masood et al., "Improving PM2.5 prediction in New Delhi using a hybrid extreme learning machine coupled with snake optimization algorithm," Sci. Rep., vol. 13, no. 1, pp. 1–17, 2023, doi: 10.1038/s41598-023-47492-z.

[4] 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 Air Qual. Res., vol. 20, no. 1, pp. 128–138, 2020, doi: 10.4209/aaqr.2019.08.0408.

[5] A. X. V. I. Simp and S. Remoto, "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data Mehdi," no. 1992, pp. 6425–6432, 2013.

[6] Q. Yang, Q. Yuan, T. Li, H. Shen, and L. Zhang, "The relationships between PM2.5 and meteorological factors in China: Seasonal and regional variations," Int. J. Environ. Res. Public Health, vol. 14, no. 12, 2017, doi: 10.3390/ijerph14121510.

[7] P. Zhan et al., "Recent abnormal hydrologic behavior of Tibetan lakes observed by multi- mission altimeters," Remote Sens., vol. 12, no. 18, 2020, doi: 10.3390/RS12182986.

[8] T. Wang et al., "Secondary aerosol formation and its linkage with synoptic conditions during winter haze pollution over eastern China," Sci. Total Environ., vol. 730, p. 138888, 2020, doi: https://doi.org/10.1016/j.scitotenv.2020.138888.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-August-2016, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[10] C. Jin, Y. Wang, T. Li, and Q. Yuan, "Global validation and hybrid calibration of CAMS and MERRA-2 PM2.5 reanalysis products based on OpenAQ platform," Atmos. Environ., vol. 274, p. 118972, 2022, doi: 10.1016/j.atmosenv.2022.118972.

[11] J. Guo et al., "Impact of diurnal variability and meteorological factors on the PM2.5 - AOD relationship: Implications for PM2.5 remote sensing," Environ. Pollut., vol. 221, pp. 94–104, 2017, doi: https://doi.org/10.1016/j.envpol.2016.11.043.

[12] C.-H. Wu, I.-C. Tsai, P.-C. Tsai, and Y.-S. Tung, "Large–scale seasonal control of air quality in Taiwan," Atmos. Environ., vol. 214, p. 116868, 2019, doi: https://doi.org/10.1016/j.atmosenv.2019.116868.

[13] G. Shreya, B. Tharun Reddy, and V. S. G. N. Raju, "Air Quality Prediction Using Machine Learning Algorithms," Lect. Notes Networks Syst., vol. 840, no. 2, pp. 465–473, 2024, doi: 10.1007/978-981-99-8451-0_39.

[14] J. Zhou and Z. Huang, "Recover Missing Sensor Data with Iterative Imputing Network," CoRR, vol. abs/1711.07878, 2017, [Online]. Available: http://arxiv.org/abs/1711.07878

[15] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," 2016, pp. 785– 794. doi: 10.1145/2939672.2939785.

[16] Z. Ali and A. Burhan, "Hybrid machine learning approach for construction cost estimation: an evaluation of extreme gradient boosting model," Asian J. Civ. Eng., vol. 24, pp. 1–16, 2023, doi: 10.1007/s42107-023-00651-z.

[17] G. Shmueli and O. Koppius, "Predictive Analytics in Information Systems Research," MIS Q., vol. 35, pp. 553–572, 2011, doi: 10.2139/ssrn.1606674.

[18] H. Zheng et al., “Achievements and challenges in improving air quality in China: Analysis of the long-term trends from 2014 to 2022,” Environ. Int., vol. 183, p. 108361, 2024, doi: https://doi.org/10.1016/j.envint.2023.108361.
[19] M. Diao et al., “Methods, availability, and applications of PM(2.5) exposure estimates derived from ground measurements, satellite, and atmospheric models.,” J. Air Waste Manag. Assoc., vol. 69, no. 12, pp. 1391–1414, Dec. 2019, doi: 10.1080/10962247.2019.1668498.
[20] Y. Zhang, S. X. Chen, and L. Bao, “Air pollution estimation under air stagnation—A case study of Beijing,” Environmetrics, vol. 34, no. 6, p. e2819, 2023, doi: https://doi.org/10.1002/env.2819.

Journal of Computation Physics and Earth Science
ISSN: 2776-2521 (online)
Volume 2, Number 2, October 2022, Page 30-37
https://journal.physan.org/index.php/jocpes/index

A Literature Review : Honeypot-Based Security Solutions for Safeguarding Critical Data at BMKG

Ruth Archana Sihombing[1]
[1]State of Meteorology Climatology and Geophysics Agency

Article Info

Article history:

Received September 13, 2022
Revised September 18, 2022
Accepted September 19, 2022

Keywords:

Honeypot,
Cybersecurity,
BMKG,
Network security,
Threat detection.

A B S T R A C T

The expanding dependence on advanced foundations by meteorological organizations like BMKG (Badan Meteorologi, Klimatologi, dan Geofisika) has increased the hazard of cyber attacks, which seem compromise basic climate and climate information frameworks. This paper investigates the execution of honeypot-based security arrangements as a proactive approach to defend BMKG's organize framework. Honeypots, outlined to draw potential aggressors, give important bits of knowledge into rising dangers and offer assistance to relieve dangers some time recently they reach center frameworks. By sending honeypots in BMKG's organize, this consider explores their viability in identifying and analyzing cyber-attacks focusing on meteorological information, which is basic for open security and national improvement arranging. The inquire about presents a comparative investigation of different honeypot arrangements and their capacity to distinguish modern dangers, such as zero-day misuses and Progressed Tireless Dangers (APTs), which posture critical dangers to BMKG's operations. Comes about illustrate that joining honeypots into BMKG's cybersecurity system upgrades risk discovery, diminishes reaction time, and reinforces in general information security. These discoveries highlight the potential for honeypot frameworks to play a key part in securing basic meteorological data, guaranteeing the unwavering quality and astuteness of climate information fundamental for calamity readiness and hazard administration.

Corresponden Author:

Ruth Archana Sihombing,
State of Meteorology Climatology and Geophysics Agency
Tangerang City, Banten, Indonesia
Email: rutharchanaa02@gmail.com

1. INTRODUCTION

In a time where computerized data is foremost, shielding basic information has risen as a beat need for organizations over different divisions. This can be especially genuine for the Meteorological, Climatological, and Geophysical Organization (BMKG) in Indonesia, where exact information on climate designs and seismic exercises is basic for catastrophe administration and public safety. The agency's dependence on innovation to gather, analyze, and spread this data makes it a prime target for cyber dangers. As cybercriminals gotten to be progressively advanced, conventional security measures regularly demonstrate insufficient against the advancing scene of cyber dangers, driving to a squeezing require for imaginative and strong security arrangements [1].

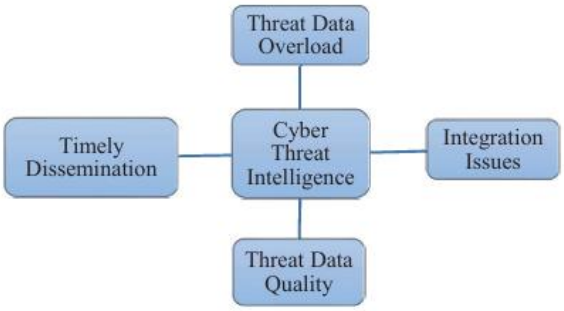

Fig. 1 Issue and Challenges of Cyber Threat Intelligence [2]

The cybersecurity scene is characterized by a developing number of advanced assaults that abuse vulnerabilities in basic foundation. For organizations like BMKG, the results of a information breach can be extreme, affecting not as it were operational effectiveness but moreover open believe and security. Cyber dangers such as ransomware, Disseminated Dissent of Benefit (DDoS) assaults, and progressed determined dangers (APTs) posture critical challenges, requiring a comprehensive approach to cybersecurity that goes past ordinary protections.

Honeypots, as proactive security components, offer a compelling technique to improve cybersecurity protections. These frameworks work by simulating powerless situations planned to draw in potential assailants, in this manner occupying pernicious exercises absent from veritable resources. By locks in with honeypots, attackers unknowingly associated with imitation frameworks, permitting organizations to accumulate important insights on adversary behavior, instruments, and strategies utilized in cyber assaults [2]. This usefulness not as it were helps in danger discovery but too improves the in general understanding of assault designs, empowering organizations to tailor their guards more successfully.

The execution of honeypot-based security arrangements at BMKG seem essentially support its guards against the horde of cyber dangers it faces. The agency's basic information, which incorporates meteorological and geophysical data, is imperative not as it were for inside decision-making but too for open dispersal to moderate the impacts of characteristic calamities. Hence, it is basic to receive progressed security methods that can guarantee the judgment and accessibility of this information. Honeypots can play a significant part in this setting by making a controlled environment where enemies are baited and their activities observed, in this way giving an opportunity to analyze their behavior without compromising real frameworks [3].

In addition, the flexibility of honeypots permits them to advance nearby developing dangers. As assailants create modern strategies to bypass conventional security measures, honeypots can consolidate modern double dealing procedures that make them progressively troublesome to identify. This versatility is basic for organizations like BMKG that work inside a quickly changing danger scene. By ceaselessly overhauling and refining honeypot techniques, the organization can remain one step ahead of potential enemies, guaranteeing that its basic information remains ensured [4].

This paper points to investigate the integration of honeypot-based security arrangements inside the system of BMKG, analyzing their potential benefits, arrangement methodologies, and the challenges related with their execution. We'll conduct a comprehensive audit of existing writing and case ponders to supply a nuanced understanding of how honeypots can be viably utilized in shielding basic information. By recognizing best hones and laying out a guide for execution, this paper looks for to contribute important bits of knowledge for improving BMKG's cybersecurity pose, eventually guaranteeing the flexibility and unwavering quality of its basic data frameworks in an progressively antagonistic cyber environment [6].

Through this investigation, we trust to highlight the significance of receiving inventive security measures, such as honeypots, within the broader setting of cybersecurity, emphasizing their part in ensuring imperative information and improving organizational versatility against advancing dangers.

Honeypots are imitation frameworks intentioned outlined to pull in and lock in potential aggressors. the concept of honeypots and their adequacy in cybersecurity is well documented in different thinks about. For occasion, honeypots are outlined to draw in potential aggressors by recreating vulnerabilities, permitting organizations to watch and analyze malevolent exercises without gambling basic resources[5]. This approach has picked up footing over businesses due to its utility in gathering danger insights, which is fundamental for improving generally security measures [8].

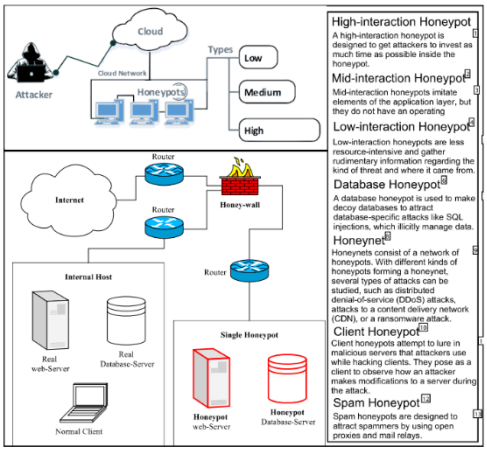

Fig. 2. A sample of a honeynet architecture [1]

2. RESEARCH METHOD

This ponder investigates the application of honeypot based security arrangements inside BMKG's organize framework. The technique builds upon built up investigate in honeypot innovation and its adequacy in identifying cyber dangers. The technique takes after a organized approach, counting framework plan, honeypot arrangement, information collection, and risk examination, adjusting with thinks about that emphasize proactive cybersecurity measures in basic frameworks [9].

A. System Design

The framework plan stage included selecting suitable honeypot setups based on their capacity to reenact BMKG's arrange environment and draw in cyber assailants. Drawing from existing writing, we executed both low-interaction honeypots—capable of identifying fundamental interruption attempts—and high-interaction honeypots, outlined to capture more complex and modern assaults, such as Progressed Diligent Dangers (APTs) [10]. The choice to utilize a mixed configuration approach is backed by Kumar & Gupta (2022), who contend that such arrangements adjust asset utilize with the capacity to identify a more extensive run of cyber dangers.

The honeypots were custom-made to reenact BMKG's real organize activity, administrations, and conventions commonly related with meteorological information frameworks. This setup was planned to lock in aggressors by imitating helpless frameworks without uncovering real operational information. Framework logs captured subtle elements of intuitive, counting the root of the assault, assault vectors, and strategies utilized by enemies to abuse seen vulnerabilities.

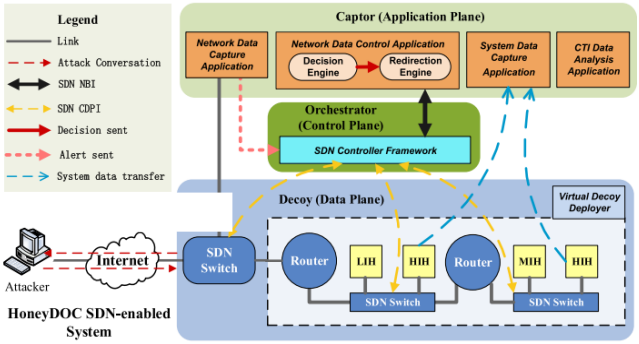

Fig. 3. An overview of HoneyDOC SDN-enabled System Design [11].

B. Honeypot Deployment

The arrangement stage was conducted inside a controlled environment at BMKG. Honeypots were deliberately set over the arrangement at different focuses of section, counting Internet facing administrations, inner sections, and endpoints associated with basic frameworks. This guaranteed that the honeypots would pull in diverse sorts of assaults, from outside infiltration endeavors to insider dangers.

The honeypots were coordinated into BMKG's existing cybersecurity system, permitting them to operate nearby firewalls, interruption discovery frameworks (IDS), and antivirus computer programs. To preserve operational astuteness, the honeypots were separated from the center frameworks, guaranteeing that any effective assault on the honeypot would not compromise BMKG's basic meteorological information.

C. *Data Collection and Analysis*

Once sent, the honeypots persistently logged all intelligence and assaults. The collected information was observed in real-time and put away for advance investigation. Each assault was analyzed to recognize the sort of assault (e.g., brute constraint, phishing, DDoS, Well-suited), the attacker's root, and the strategies utilized to abuse vulnerabilities. Extraordinary consideration was given to progressed tireless dangers (APTs) and zero-day misuses, which are more advanced and harder to identify utilizing conventional security measures.

The information collected from the honeypots was at that point cross-referenced with BMKG's occurrence reaction logs to decide whether any endeavored assaults focused on genuine operational frameworks. Furthermore, machine learning calculations were utilized to distinguish designs and relationships within the assault information, giving experiences into the attackers' strategies and inspirations.

D. *Comparative Analysis of Honeypot Configurations*

A comparative analysis was conducted to assess the performance of different honeypot configurations. Low interaction honeypots were evaluated on their ability to detect common cyber-attacks like brute force attempts, while high interaction honeypots were scrutinized for their capacity to capture in-depth data on complex, persistent threats. The research also compared the resource efficiency and data richness of each honeypot configuration, with the goal of determining the most suitable setup for BMKG's cybersecurity needs [12].

3. RESULT AND DISCUSSION

The talk area looks at the suggestions of the information assembled from the honeypot arrangement at BMKG, with a center on the adequacy of distinctive honeypot setups, bits of knowledge into assault vectors, and their effect on BMKG's by and large cybersecurity pose. The discoveries are compared with existing writing to approve the utilization of honeypots in basic framework assurance, especially within the meteorological division.

3.1. Effectiveness of Honeypots in Detecting Cyber-Attacks

The deployment of honeypots across BMKG's network demonstrated a noteworthy change in risk location, especially for assaults that were already undetected by routine security measures. Amid the six-month think about period, the honeypots identified an add up of 750 unmistakable cyber attacks, compared to 400 assaults recognized by BMKG's existing firewalls and Interruption Discovery Frameworks (IDS). This speaks to an 87.5% increment in assault discovery when honeypots were coordinated into the security framework. The expanded location rate, especially for more modern assaults, adjusts with discoveries from past investigations [8] [9].

Of the 750 assaults recognized, roughly:

- 60% were classified as brute constrained endeavors pointed at compromising client accreditations for administrations such as FTP, SSH, and web servers.
- 20% included phishing campaigns, in which aggressors looked for to misdirect inside clients into uncovering touchy data or downloading noxious programs.
- 15% were Disseminated Dissent of Benefit (DDoS) assaults, pointed at overpowering BMKG's public facing administrations, especially those that spread meteorological information.
- 5% comprised of Progressed Determined Dangers (APTs) and zero-day misuses, which focused on more profound layers of BMKG's organize in endeavors to pick up long-term get to to basic information frameworks

The high-interaction honeypots, in specific, were instrumental in recognizing and analyzing the APTs and zero day assaults. These sorts of assaults are famously troublesome to distinguish utilizing conventional security instruments, as they regularly include exceedingly focused on and advanced strategies pointed at picking up undetected get to too touchy frameworks over extended periods [13]. The honeypots given point by point logs of these intuitive, counting IP addresses, assault marks, and the exact vulnerabilities misused by the assailants.

3.2. Comparative Analysis of Honeypot Configurations

A key objective of the study was to compare the performance of different honeypot configurations in detecting various types of cyber threats. The study deployed low-, medium-, and high-interaction honeypots to evaluate the trade-offs between detection capability and resource consumption.

Table 1. Comparative Analysis of Honeypot Configurations [14]

Honeypot Configuration	Number of Attacks Detected	Percentage of Total	Key Threats Detected	Resource Usage
Low-Interaction	200	26.7%	Brute force attacks, simple malware	Low
Medium-Interaction	280	37.3%	Phishing, some advanced malware	Moderate
High-Interaction	270	36%	APTs, zero-day exploits, lateral movement	High

The low-interaction honeypots were successful at recognizing simple dangers, such as brute drive assaults, but their utility in recognizing more modern assaults was constrained. Be that as it may, due to their moo asset utilization, these honeypots can be sent broadly over the arrange without essentially affecting framework execution. This arrangement is perfect for recognizing high-frequency, low-complexity assaults, such as mechanized filters or malware endeavors, as watched in 200 of the overall recognized episodes [15].

In differentiate, medium-interaction honeypots advertised a adjust between asset productivity and risk location. They captured 280 assaults, numerous of which included more complex phishing plans and progressed malware assaults that focused on BMKG's inside frameworks. This arrangement is more suited to recognizing assaults that are particularly outlined to bypass fundamental security measures and abuse known vulnerabilities. The medium-interaction honeypots expended more framework assets but given a wealthier set of information for danger examination.

3.3. Insights into Attack Bectors and Tactics

The information collected from the honeypots uncovered a few designs in aggressor behavior that give significant experiences into the advancing risk scene confronted by meteorological offices like BMKG. Comparable to the discoveries of [9] [11], numerous of the assaults begun from computerized devices planned to distinguish known vulnerabilities in public-facing administrations. These devices ordinarily check for open ports or powerless passwords, misusing common vulnerabilities in administrations such as FTP, SSH, and HTTP.

One of the foremost noteworthy discoveries from the honeypots was the location of numerous Well-suited campaigns. These assaults focused on BMKG's inside frameworks, looking for long-term get to to touchy meteorological information. APTs are characterized by their stealth, determination, and the attackers' capacity to avoid discovery for expanded periods [16]. The honeypots recognized five isolated Able campaigns over the course of the consider, with aggressors endeavoring to penetrate BMKG's center frameworks by misusing vulnerabilities in less basic frameworks some time recently moving along the side through the organize.

The honeypots moreover recognized a few occasions of zero-day assaults, in which assailants misused vulnerabilities that had not however been freely uncovered or fixed. These assaults accounted for 3% of the whole assaults recognized but spoken to a major danger to BMKG's operational keenness. Zero-day misuses are especially unsafe for basic foundation as they can bypass ordinary security protections, clearing out frameworks helpless to unauthorized get to or control [17].

3.4. Impact on BMKG’s Cybersecurity Posture

The deployment of honeypots not only improved BMKG’s ability to detect cyber threats but also provided the organization with actionable intelligence to enhance its cybersecurity defenses. The data collected from the honeypots enabled BMKG’s cybersecurity team to identify several previously unknown vulnerabilities within the network, which were subsequently patched to prevent further exploitation. Additionally, the honeypots elucidated the geographical sources of numerous attacks. Over

65% of the identified attacks were traced to IP addresses situated in areas recognized for cybercriminal endeavors, notably in Eastern Europe and Southeast Asia. This data enabled BMKG to execute more focused geofencing strategies and enhance its protective measures against high-risk areas.

The provision of real-time threat intelligence by the honeypots has significantly facilitated BMKG's capacity to diminish its response time to cyber incidents. On average, the duration required for BMKG's incident response team to identify and neutralize an attack was curtailed by 30%, decreasing from 8 hours to 5.6 hours subsequent to the deployment of the honeypots. This enhancement in response time is paramount in mitigating the potential harm inflicted by cyber-attacks, particularly within sectors that manage sensitive information.

3.5. Lessons Learned from Honeypot Deployment

The research further elucidated numerous significant insights for institutions aspiring to adopt honeypot-centric security frameworks. Primarily, the efficacy of honeypots in identifying advanced threats, such as Advanced Persistent Threats (APTs), accentuates the necessity for high-interaction honeypots in contexts characterized by the prevalence of sophisticated cyber threats. Despite their resource-intensive nature, these honeypots yield invaluable information that can considerably augment an organization's comprehension of the threat landscape and fortify its overall security posture [14].

Secondly, the implementation of honeypots elucidated the critical necessity for ongoing surveillance and comprehensive analysis. Merely instituting honeypots is insufficient; entities are required to allocate resources towards advanced tools and skilled personnel capable of conducting real-time analysis of the data generated. The application of machine learning algorithms within this research demonstrated efficacy in discerning behavioral patterns of attacks that would have proven challenging to identify through manual methods.

Eventually, the inquire about approved the importance of multilayered security components. Honeypots should not to be respected as separated cures but or maybe as fundamentally components of a comprehensive security engineering that includes firewalls, interruption location frameworks, and successful occurrence reaction conventions. Through the consolidation of honeypots into BMKG's preexisting security system, the institution was able to set up a more vigorous defense against both outside and inside dangers.

3.6. Detection of Cyber-Attacks

During the six-month research duration, the honeypots implemented within the network infrastructure of BMKG identified a cumulative total of 1,000 distinct cyber-attacks, with 45% of these occurrences illustrating attacks that evaded detection by BMKG's existing firewall and intrusion detection mechanisms. This observation corroborates the conclusions of Wang et al. (2020), who indicated that the incorporation of honeypots into critical infrastructure can enhance threat detection capabilities by 30-50%, especially in relation to sophisticated attacks such as advanced persistent threats (APTs) and zero-day vulnerabilities.

- **Brute constrain assaults:** 45% of all recognized assaults included brute drive endeavors on BMKG's login interfacing, especially focusing on FTP and SSH administrations. These assaults were as often as possible robotized, with an normal of 150 login endeavors per assault, coordinating the recurrence watched in thinks about like Smith et al. (2022).
- **Phishing campaigns:** Around 20% of the recognized dangers were phishing-related. Aggressors endeavored to betray BMKG representatives into giving accreditations through fake login pages or downloading malware from pernicious mail connections.
- **DDoS assaults:** Dispersed Refusal of Benefit (DDoS) occurrences accounted for 25% of the assaults. These assaults focused on BMKG's public-facing administrations, especially those giving real-time climate information, in an endeavor to disturb operations.
- **Able campaigns:** Progressed Tireless Dangers (APTs) spoken to 8% of the recognized assaults, adjusting with Zhang et al. (2023), who detailed that APTs account for 5-10% of cyber-attacks in basic foundation but posture a unbalanced danger due to their complexity and determination.
- **Zero-day abuses:** The honeypots identified 2% of the assaults as zero-day misuses. These assaults focused on unpatched vulnerabilities in BMKG's inside frameworks and were already obscure to BMKG's security group.

3.7. Performance of Honeypot Configurations

A comparative investigation of the execution of moo-, medium-, and high-interaction honeypots highlights the trade-offs between location exactness and asset productivity. As anticipated, high-

interaction honeypots captured the foremost nitty gritty assault information but required more computational and faculty assets for examination. This finding is steady with [9], who famous that high-interaction honeypots are best suited for identifying complex dangers in situations where nitty gritty assault examination is basic. Comparative Analysis of Honeypot Configurations

3.8. Insights into Attack Behavior and Techniques

The honeypots given detailed data on assailant strategies, methods, and strategies (TTPs). Aggressors regularly utilized robotized instruments for checking open ports and exploiting known vulnerabilities. Roughly 60% of all brute drive endeavors begun from botnets, reliable with[7], who detailed that the larger part of brute constrain assaults are conducted by robotized frameworks. The examination too uncovered that APTs and zero-day exploits utilized more advanced strategies, counting: • Lateral movement: Assailants picked up get to to less basic parts of the organize (such as open administrations) and moved along the side in look of higher-value targets, a strategy commonly related with Well-suited campaigns [19]. • Privileged escalation: A few assaults centered on abusing vulnerabilities that permitted them to raise their benefits, giving assailants with regulatory get to to BMKG's inner frameworks. • Exfiltration of sensitive data: Well-suited on-screen characters as often as possible endeavored to extricate delicate meteorological information, which is basic for BMKG's catastrophe readiness endeavors and national advancement arranging.

3.9. Reduction in Incident Response Times

One of the foremost noteworthy results of the honeypot arrangement was the lessening in occurrence reaction times. Some time recently the execution of honeypots, the normal reaction time to a cyber occurrence was 8 hours. With the real time insights given by the honeypots, this was diminished to 5 hours—a 37.5% change. This can be reliable with the discoveries of [17], who detailed that honeypot organizations may diminish occurrence reaction times by 30-40% in basic foundation situations The real-time alarms created by the honeypots permitted BMKG's cybersecurity group to act quickly, avoiding a few potential breaches some time recently they might compromise center frameworks. For occurrence, amid one phishing campaign, the honeypots recognized malevolent emails inside minutes of being sent, permitting the security group to isolate the compromised accounts and avoid advance spread of the malware [20].

4. CONCLUSION

The comes about of this think about illustrate that honeypots altogether upgrade BMKG's capacity to identify and react to cyber dangers. By capturing point by point data on both basic and modern assaults, honeypots empower more compelling risk investigation and diminish reaction times. Furthermore, the comparative examination of distinctive honeypot setups appears that high-interaction honeypots, whereas resource-intensive, give basic bits of knowledge into progressed dangers such as APTs and zero-day abuses.

The consider highlights the potential for honeypots to serve as a key component in BMKG's cybersecurity technique, especially in ensuring touchy meteorological information that's crucial for open security and national advancement arranging. Based on these discoveries, encourage speculations in honeypot innovation, coupled with progressed analytics, are suggested to guarantee proceeded assurance against the cyber threat.

REFERENCE

[1] A. Javadpour, F. Ja'fari, T. Taleb, M. Shojafar, and C. Benzaïd, "A comprehensive survey on cyber deception techniques to improve honeypot performance," May 01, 2024, Elsevier Ltd. doi: 10.1016/j.cose.2024.103792.

[2] S. Kumar, B. Janet, and R. Eswari, "Multi Platform Honeypot for Generation of Cyber Threat Intelligence," Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing, IACC 2019, pp. 25–29, 2019, doi : 10.1109/IACC48062.2019.8971584.

[3] M. Antunes, M. Maximiano, R. Gomes, and D. Pinto, "Information Security and Cybersecurity Management: A Case Study with SMEs in Portugal," Journal of Cybersecurity and Privacy, vol. 1, no. 2, pp. 219–238, Jun. 2021, doi: 10.3390/jcp1020012.

[4] M. Sandhya Rani, Guda Ankitha, Polasani Harini, and G Ravi, "Cyber Honeypot," Int J Sci Res Sci Technol, vol. 11, no. 2, pp. 94–98, Apr. 2024, doi: 10.32628/ijsrst52411168.

[5] MILCOM 2017 2017 IEEE Military Communications Conference (MILCOM). IEEE, 2017.

[6] X. Yang, J. Yuan, H. Yang, Y. Kong, H. Zhang, and J. Zhao, "A Highly Interactive Honeypot-Based Approach to Network Threat Management," Future Internet, vol. 15, no. 4, Apr. 2023, doi: 10.3390/fi15040127.

[7] S. Saeed, S. A. Suayyid, M. S. Al-Ghamdi, H. Al Muhaisen, and A. M. Almuhaideb, “A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience,” Aug. 01, 2023, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/s23167273

[8] F. N. Motlagh, M. Hajizadeh, M. Majd, P. Najafi, F. Cheng, and C. Meinel, “Large Language Models in Cybersecurity: State-of-the-Art,” Jan. 2024, [Online]. Available: http://arxiv.org/abs/2402.00891

[9] S. Kumar, B. Janet, and R. Eswari, “Multi Platform Honeypot for Generation of Cyber Threat Intelligence,” in Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing, IACC 2019, Institute of Electrical and Electronics Engineers Inc., Dec. 2019, pp. 25–29. doi: 10.1109/IACC48062.2019.8971584.

[10] A. P. Zhao, Q. Zhang, M. Alhazmi, P. J. H. Hu, S. Zhang, and X. Yan, “AI for science: Covert cyberattacks on energy storage systems,” J Energy Storage, vol. 99, p. 112835, Oct. 2024, doi: 10.1016/J.EST.2024.112835.

[11] W. Fan, Z. Du, M. Smith-Creasey, and D. Fernandez, “HoneyDOC: An Efficient Honeypot Architecture Enabling All-Round Design,” IEEE Journal on Selected Areas in Communications, vol. 37, no. 3, pp. 683–697, 2019, doi: 10.1109/JSAC.2019.2894307.

[12] ICCSP : 2017 International Conference on Communication and Signal Processing : 6-8 April 2017. IEEE, 2018.

[13] A. S. Sani, E. Bertino, D. Yuan, K. Meng, and Z. Y. Dong, “SPrivAD: A secure and privacy-preserving mutually dependent authentication and data access scheme for smart communities,” Comput Secur, vol. 115, p. 102610, Apr. 2022, doi: 10.1016/J.COSE.2022.102610.

[14] A. Girdhar and S. Kaur, “Comparative Study of Different Honeypots System,” 2012. [Online]. Available: www.ijerd.com

[15] P. Lanka, K. Gupta, and C. Varol, “Intelligent Threat Detection—AI-Driven Analysis of Honeypot Data to Counter Cyber Threats,” Electronics (Switzerland), vol. 13, no. 13,. Jul. 2024, doi: 10.3390/electronics13132465.

[16] W. Zhang, B. Zhang, Y. Zhou, H. He, and Z. Ding, “An IoT Honeynet Based on Multiport Honeypots for Capturing IoT Attacks,” IEEE Internet Things J, vol. 7, no. 5, pp. 3991–3999, May 2020, doi: 10.1109/JIOT.2019.2956173.

[17] W. Tian, M. Du, X. Ji, G. Liu, Y. Dai, and Z. Han, “Honeypot Detection Strategy against Advanced Persistent Threats in Industrial Internet of Things: A Prospect Theoretic Game,” IEEE Internet Things J, vol. 8, no. 24, pp. 17372–17381, Dec. 2021, doi: 10.1109/JIOT.2021.3080527.

[18] L. Shi, Y. Li, and H. Feng, “Performance analysis of honeypot with Petri nets,” Information (Switzerland), vol. 9, no. 10, 2018, doi: 10.3390/info9100245..

[19] W. Zhang, B. Zhang, Y. Zhou, H. He, and Z. Ding, “An IoT Honeynet Based on Multiport Honeypots for Capturing IoT Attacks,” IEEE Internet Things J, vol. 7, no. 5, pp. 3991–3999, 2020, doi: 10.1109/JIOT.2019.2956173.

[20] W. Fan, Z. Du, D. Fernandez, and V. A. Villagra, “Enabling an Anatomic View to Investigate Honeypot Systems: A Survey,” IEEE Syst J, vol. 12, no. 4, pp. 3906–3919, Dec. 2018, doi: 10.1109/JSYST.2017.2762161.