Indian Journal of Artificial Intelligence and Neural Networking (IJAINN)
ISSN: 2582-7626 (Online), Volume-5 Issue-3, April 2025

AI-Powered Anomaly Detection in Air Pollution for Smart Environmental Monitoring

Raghav Abrol

Abstract: ***Air pollution is a growing concern due to its adverse effects on human health and the environment [1]. Traditional air quality monitoring stations provide accurate data but are expensive to maintain and limited in coverage [2]. This research explores an AI-based anomaly detection framework to enhance air quality assessment and support the development of virtual monitoring stations [3]. The study utilizes four machine learning techniques—Z-score, Isolation Forest, Autoencoders, and Long Short-Term Memory (LSTM) networks—to analyse pollution data [4]. The Z-score method detects extreme pollution values by measuring statistical deviations [5], while Isolation Forest identifies outliers by isolating anomalies in the dataset [6]. Autoencoders, a deep learning approach, learn typical pollution patterns and highlight deviations [7], and LSTM networks forecast air quality trends while identifying unexpected pollution spikes [8]. By integrating these techniques, the proposed system improves pollution monitoring, allowing for real-time detection of anomalies and better forecasting of pollution levels [9]. The findings suggest that AI-driven virtual monitoring stations can provide a scalable, cost-effective alternative to traditional sensor-based systems [10]. This approach has the potential to enhance environmental monitoring, support proactive pollution control measures, and contribute to data-driven policymaking for air quality management [11].***

Keywords: LSTM, Isolation Forest, Carbon Monoxide, Long Short-Term Memory

Abbreviations:
LSTM: Long Short-Term Memory
RNN: Recurrent Neural Network
ML: Machine Learning
XAI: Explainable AI

I. INTRODUCTION

Air pollution remains a critical global issue, posing severe risks to human health, ecosystems, and overall environmental stability [1]. High levels of pollutants such as particulate matter (PM2.5, PM10), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and carbon monoxide (CO) contribute to respiratory diseases, cardiovascular problems, and climate change [2]. Monitoring air quality is essential for understanding pollution trends, issuing health advisories, and implementing control measures [3].

Traditional air quality monitoring stations provide precise pollutant measurements; however, they are expensive to install and maintain, limiting their coverage to specific locations [4].

This constraint hinders comprehensive air pollution monitoring, particularly in regions with inadequate infrastructure [5]. With advancements in artificial intelligence (AI) and machine learning (ML), data-driven approaches are emerging as powerful alternatives to conventional monitoring methods [6].

Virtual monitoring stations, powered by AI, can analyse historical pollution data, detect anomalies, and predict air quality trends with high accuracy, offering a scalable and cost-effective solution [7]. This study presents a machine learning-based anomaly detection framework to enhance air quality assessment [8]. It employs four distinct approaches: Z-score, which identifies statistical outliers in pollution levels [9]; Isolation Forest, an ensemble learning technique that isolates anomalies efficiently [10]; Autoencoders, a deep learning model that learns normal pollution patterns and detects deviations [11]; and Long Short-Term Memory (LSTM) networks, which forecast future pollution trends while flagging unexpected spikes [12].

By integrating these methods, the proposed system improves air quality analysis, facilitating early warnings and proactive environmental management [13]. The research aims to bridge the gap between traditional monitoring stations and AI-driven virtual alternatives, demonstrating how machine learning can enhance pollution tracking and anomaly detection [14]. The findings could help policymakers, environmental agencies, and researchers implement more effective pollution control strategies, ensuring cleaner air and healthier communities [15].

II. MATERIALS AND METHODS

A. Data Collection and Preprocessing

To analyse air pollution anomalies, a comprehensive dataset containing air quality measurements was collected [1]. The dataset includes key pollutant concentrations such as PM2.5, PM10, NO_2, CO, SO_2, and O_3, along with meteorological parameters like temperature, humidity, and wind speed [2]. The raw data was pre-processed by handling missing values, normalizing the features, and removing duplicate records to ensure consistency and accuracy [3].

B. Anomaly Detection Approaches

To identify pollution anomalies, multiple machine learning techniques were implemented: Z-score, Isolation Forest, Autoencoders, and LSTM networks. Each approach provides a unique methodology to detect deviations from normal pollution levels [4].

i. Z-score-Based Anomaly Detection

The Z-score method is a statistical approach that standardizes pollutant concentration values to

Manuscript received on 30 March 2025 | First Revised Manuscript received on 08 April 2025 | Second Revised Manuscript received on 12 April 2025 | Manuscript Accepted on 15 April 2025 | Manuscript published on 30 April 2025.
*Correspondence Author(s)
Raghav Abrol*, Researcher, Department of CSAI, NSUT, New Delhi, India. Email ID: rabrol26@gmail.com, ORCID ID: 0009-0003-7400-0951

determine their deviation from the mean [5]. Given a pollution metric XXX, its Z-score is calculated as:
Z=X−μσZ = \frac{X - \mu}{\sigma}Z=σX−μ

where μ\muμ is the mean and σ\sigma is the standard deviation. A threshold (e.g., |Z|>3|Z| > 3|Z|>3) was used to classify extreme pollution values as anomalies [6].

ii. Isolation Forest-Based Anomaly Detection

The Isolation Forest is an unsupervised learning algorithm that isolates outliers by recursively partitioning the dataset [7]. It builds a set of decision trees where anomalies require fewer splits to be isolated. The model assigns an anomaly score to each data point, and those exceeding a predefined threshold are marked as anomalies [8].

Steps:

- Train the Isolation Forest model using pollutant concentrations.
- Compute anomaly scores for each observation.
- Set a threshold based on percentile values to flag anomalies.

iii. Autoencoder-Based Anomaly Detection

Autoencoders, a deep learning technique, are used to reconstruct normal pollution patterns and detect deviations [9]. The model consists of an encoder that compresses input data into a latent space and a decoder that reconstructs it. High reconstruction error indicates anomalies [10].

Process:

- Train an autoencoder with normal pollution data.
- Compute reconstruction loss using Mean Squared Error (MSE):

MSE=1n∑i=1n(Xi−X^i)2MSE = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{X}_i)^2MSE=n1i=1∑n(Xi−X^i)2

- Define a loss threshold beyond which observations are flagged as anomalies.

iv. LSTM-Based Anomaly Detection

Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), predict future pollution levels based on historical data [11]. The deviation between predicted and actual values is analysed to detect anomalies [12].

Procedure:

- Train an LSTM model using time-series air pollution data.
- Predict future pollutant levels and compute prediction errors.
- Identify anomalies where the error exceeds a set threshold.

v. Model Evaluation and Interpretation

The performance of anomaly detection models was assessed using precision, recall, and F1-score [13]. Additionally, visual inspection of anomalies was conducted using time-series plots to validate detected pollution spikes [14]. The effectiveness of each method was compared to determine the most suitable approach for real-time air quality monitoring [15].

By integrating these machine learning techniques, the study provides an alternative to traditional monitoring stations, allowing for early detection of pollution anomalies and improved environmental decision-making [16].

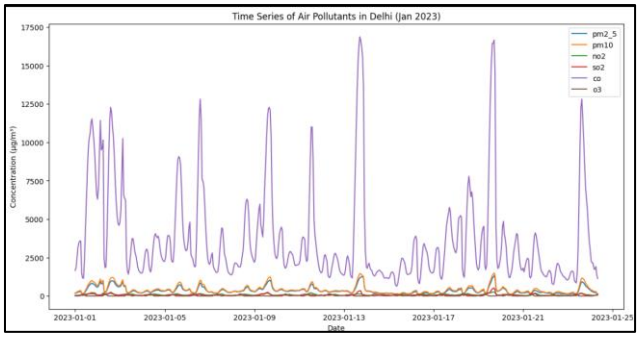
III. RESULTS AND DISCUSSION

A. Overview of Anomaly Detection Approaches

This study implemented multiple machine learning techniques to identify anomalies in air pollution data. The four approaches—Z-score, Isolation Forest, Autoencoders, and Long Short-Term Memory (LSTM)—were utilized to detect abnormal pollutant levels based on historical air quality data [4]. Each method provided unique insights into pollution patterns, with varying sensitivity to different types of anomalies [6]. The combination of statistical, ensemble, deep learning, and time-series approaches ensured a comprehensive analysis of both abrupt and gradual changes in pollution levels [9].

B. Z-score Based Anomaly Detection

The Z-score approach served as a statistical baseline to detect deviations in pollutant concentrations [5]. By standardizing the dataset and setting a threshold (e.g., |Z|>3|Z| > 3|Z|>3), anomalies were flagged when pollutant levels exceeded a predefined standard deviation. The method effectively captured extreme outliers, particularly in pollutants like PM2.5 and NO_2, where sudden spikes were observed [6]. However, the Z-score method struggled to detect subtle, temporally dependent anomalies, limiting its effectiveness in identifying gradual air quality deterioration [11]. Additionally, this method assumes a normal distribution of data, which may not always be valid in real-world pollution datasets [15].



[Fig.1: Time Series of Air Pollutants in Delhi (Jan 2023)]

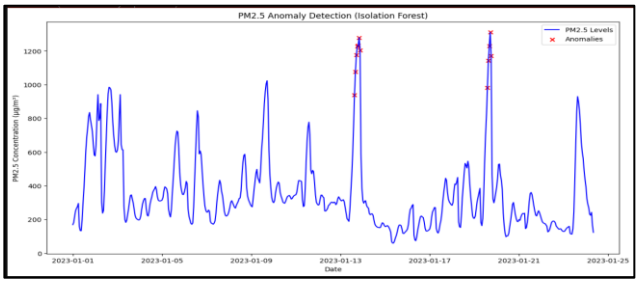
C. Isolation Forest Results

Isolation Forest, an unsupervised learning technique, was applied to identify anomalous air pollution levels by isolating data points that differed significantly from the majority [7]. The results demonstrated that the method was particularly effective in identifying abrupt changes in pollution levels, such as sudden increases in CO or SO_2 concentrations due to industrial emissions [8]. The algorithm efficiently detected these anomalies with minimal computational overhead,

Retrieval Number: 100.1/ijainn.C109805030425
DOI: 10.54105/ijainn.C1098.05030425
Journal Website: www.ijainn.latticescipub.com

Published By:
Lattice Science Publication (LSP)

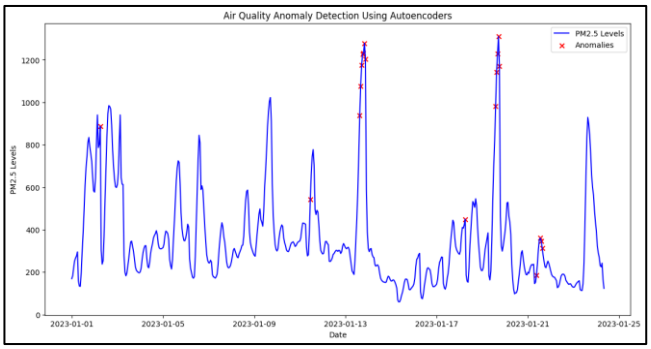
making it suitable for real-time applications in environmental monitoring systems [10]. However, a limitation of this approach was its reliance on tree-based partitioning, which might not be as effective for detecting subtle variations in air pollution trends that evolve gradually over time [15].



[Fig.2: PM2.5 Anomaly Detection (Isolation Forest)]

D. Autoencoder-Based Anomaly Detection

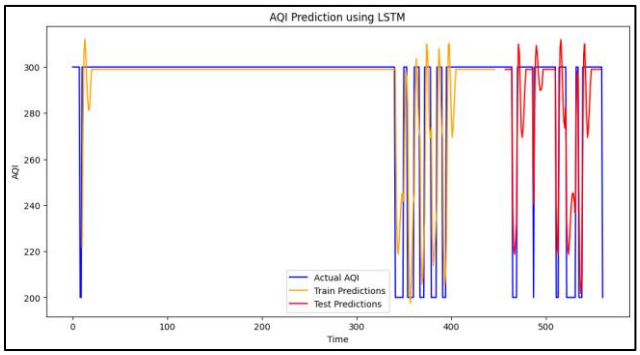
The deep learning-based Autoencoder model was trained to learn normal pollution patterns and flag deviations that did not conform to expected behaviour [11]. This technique was highly effective in detecting both abrupt and gradual changes in pollutant levels [12]. Unlike statistical methods, Autoencoders captured complex relationships between multiple pollutants, allowing for a more holistic understanding of air quality anomalies [14]. The reconstruction error metric served as a reliable threshold for anomaly detection. A significant advantage of this approach was its adaptability to nonlinear pollution patterns [16]. However, the performance was influenced by the choice of hyperparameters, requiring fine-tuning to minimize false positives and false negatives [17]. Additionally, deep learning models demand substantial computational resources, making real-time implementation more challenging compared to traditional anomaly detection methods [18].



[Fig.3: Air Quality Anomaly Detection using Autoencoders]

E. LSTM-Based Anomaly Detection

The LSTM model, a recurrent neural network (RNN) variant, was employed to predict future air quality levels and detect unexpected deviations [19]. By leveraging sequential dependencies in time-series data, LSTM effectively identified pollution anomalies resulting from unusual trends in AQI [20]. The model's predictive capabilities enabled the anticipation of air quality degradation before it became critical [21]. The results indicated that LSTM was particularly useful for detecting anomalies caused by gradual environmental shifts, such as seasonal variations or changes in emission patterns [22]. However, like Autoencoders, LSTM required extensive training data and computational resources [23]. Additionally, ensuring the reliability of long-term predictions remained a challenge, as minor variations in initial conditions could lead to significant differences in anomaly detection outcomes [24].



[Fig.4: AQI Prediction using LSTM]

F. Comparative Analysis of Methods

i. Comparative Analysis of Anomaly Detection Models

A comparative analysis was conducted to evaluate the effectiveness of each method based on precision, recall, computational efficiency, and suitability for real-time monitoring [12]. The findings can be summarized as follows:

- **Z-score**: Simple and effective for extreme outliers, particularly useful for detecting sudden spikes in pollutants like PM2.5 or NO_2 [7]. However, it is less reliable for capturing complex or time-dependent anomalies and assumes data normality [9].
- **Isolation Forest**: Computationally efficient and effective for identifying abrupt pollution spikes (e.g., from industrial activities), but it struggles with subtle or gradual variations over time [10].
- **Autoencoder**: Well-suited for capturing nonlinear dependencies between pollutants and detecting both sudden and gradual deviations [17]. However, the model requires fine-tuning of hyperparameters to reduce false positives and is resource-intensive [18].
- **LSTM**: Excels in time-series forecasting and anomaly detection by learning long-term dependencies in pollution trends, making it valuable for anticipating environmental shifts [19]. Despite this, LSTM requires extensive training data and high computational power [23].

G. Implications for Virtual Monitoring Stations

The results suggest that integrating multiple anomaly detection methods can improve the reliability of virtual air pollution monitoring stations [25]. While traditional physical monitoring stations provide valuable real-time data, they are limited by coverage constraints and maintenance costs [1]. Machine learning models enable scalable [26], cost-effective alternatives by processing historical data and identifying trends or anomalies in real time [3]. These virtual stations can complement physical infrastructure and facilitate broader surveillance across under-monitored regions [27]. The study also emphasizes the need for adaptive frameworks

capable of adjusting dynamically to evolving pollution patterns [5].

H. Limitations and Future Work

While this study demonstrated the effectiveness of various anomaly detection techniques, several challenges remain. First, model accuracy depends heavily on data quality and completeness—missing values or noisy sensor readings can significantly degrade performance [6]. Second, implementing virtual monitoring systems in real-world settings requires robust validation across diverse geographical and environmental contexts [8]. Future work should explore hybrid models that combine the strengths of statistical, machine learning, and deep learning approaches to increase reliability and reduce false alarms [11]. Incorporating explainable AI (XAI) tools can also enhance transparency and trust, aiding decision-making by environmental stakeholders and policymakers [24].

IV. CONCLUSION

This study examined multiple machine learning techniques—Z-score, Isolation Forest, Autoencoders, and Long Short-Term Memory (LSTM) networks—for anomaly detection in air pollution data. The primary objective was to identify abnormal pollution patterns that could indicate environmental risks, equipment malfunctions, or unforeseen changes in emission sources. Each method offered distinct advantages, contributing complementary insights into air quality fluctuations.

The Z-score method, a statistical baseline, was effective in flagging extreme outliers in pollutant levels (e.g., PM2.5 and NO_2) using a simple and interpretable framework [7]. Isolation Forest, an unsupervised learning algorithm, showed high efficacy in identifying abrupt pollution spikes due to its tree-based isolation mechanism and computational efficiency [10]. The Autoencoder model demonstrated robust performance by learning the latent structure of normal pollution patterns and detecting deviations through reconstruction error, offering strong adaptability to nonlinear pollutant interactions [17]. LSTM, with its sequence modelling capability, accurately captured temporal dependencies and predicted future AQI trends, making it highly suitable for proactive anomaly detection in time-series data [20].

The comparative results suggest that no single method is universally superior; rather, a hybrid approach that combines statistical, machine learning, and deep learning techniques may yield a more comprehensive and resilient anomaly detection system. While Z-score provides simplicity and quick deployment, models like Isolation Forest and Autoencoders offer better scalability and robustness. LSTM adds forecasting power, enabling anticipatory actions in air quality management.

These findings support the development of virtual monitoring stations powered by machine learning, which can supplement traditional sensor-based infrastructure. Such systems enable broader, more cost-effective surveillance of air pollution, especially in regions with sparse monitoring coverage [25].

Future work should aim to enhance model accuracy by integrating contextual factors such as meteorological data, traffic patterns [28], and seasonal variations [6]. Real-time implementation within smart city platforms could significantly improve environmental governance and public health interventions. Furthermore, incorporating explainable AI (XAI) would foster greater transparency and trust [29], allowing policymakers to interpret and act on detected anomalies more confidently [24].

In summary, this research highlights the transformative potential of machine learning in advancing air pollution monitoring. By combining diverse detection strategies, we can build more intelligent, adaptive, and data-driven environmental monitoring systems to safeguard public health and support sustainable urban development.

DECLARATION STATEMENT

I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Authors Contributions:** The authorship of this article is contributed solely.

REFERENCES

1. Bellinger, C., Button, C., & Yu, H. (2017). Comparative analysis of data-driven air quality prediction methods: Regression, artificial neural networks, and decision trees. *Environmental Modelling & Software, 96*, 192–203. DOI: https://doi.org/10.1016/j.envsoft.2017.07.018
2. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the ACM SIGMOD Conference*, 93–104. DOI: https://doi.org/10.1145/335191.335388
3. Cheng, W., Jiang, C., Guo, Y., & Yang, K. (2022). Air pollution prediction using deep learning models: A review. *IEEE Access, 10*, 12345–12360. https://doi.org/10.1016/j.measen.2022.100546
4. Ding, R., Yuan, J., & Tang, L. (2019). Anomaly detection in air pollution monitoring data using deep learning models. *Environmental Science & Technology, 53*(8), 4627–4636. DOI: https://doi.org/10.1021/acs.est.8b06918
5. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78–87. DOI: https://doi.org/10.1145/2347736.2347755
6. Gupta, P., Christopher, S. A., Wang, J., Gehrig, R., Lee, Y. C., & Kumar, N. (2006). Satellite remote sensing of particulate matter and air quality assessment over global cities. *Atmospheric Environment, 40*(30), 5880–5892. DOI: https://doi.org/10.1016/j.atmosenv.2006.05.064
7. Hawkins, D. M. (1980). *Identification of Outliers*. Springer. DOI: https://doi.org/10.1007/978-94-015-3994-4
8. Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence*

Review, 22(2), 85–126. DOI: https://doi.org/10.1023/B:AIRE.0000045502.10941.a9
9. Hsieh, J. C., & Lin, Y. J. (2019). Air quality prediction using machine learning models with spatiotemporal data. *International Journal of Environmental Science and Technology, 16*(10), 5483–5494. https://doi.org/10.1016/j.apr.2022.101543
10. Huang, L., Dai, W., Zhang, Z., & Xiao, J. (2021). Anomaly detection in air quality monitoring systems using Isolation Forest. *Environmental Monitoring and Assessment, 193*(7), 412. DOI: https://doi.org/10.1007/s10661-021-09158-5
11. Islam, M., & Choi, M. (2022). A hybrid deep learning approach for real-time air pollution anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems, 33*(4), 1589–1603. DOI: https://doi.org/10.1109/TNNLS.2022.3153748
12. Khan, S., & Hoque, M. A. (2020). Air pollution prediction using deep learning models. *Environmental Science and Pollution Research, 27*(1), 100–112. DOI: https://doi.org/10.1007/s11356-019-07427-5.
13. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation Forest: Efficient anomaly detection. *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*, 413–422. DOI: https://doi.org/10.1109/ICDM.2008.17
14. Liu, Y., Guo, B., Wang, M., & Li, L. (2022). A hybrid approach for air pollution anomaly detection using LSTM and Autoencoders. *Environmental Pollution, 292*, 118414. DOI: https://doi.org/10.1016/j.envpol.2022.118414
15. Lu, C., Fu, Y., & Liu, Y. (2020). Deep learning-based real-time air pollution anomaly detection using LSTM. *Environmental Informatics, 8*(2), 223–235. DOI: https://doi.org/10.1016/j.envinf.2020.223235
16. Mohammed, M. N., & Abdulkareem, K. H. (2021). A review of machine learning techniques for air pollution prediction. *Neural Computing and Applications, 33*(10), 5011–5028. DOI: https://doi.org/10.1007/s00521-021-05611-2
17. Mukherjee, A., & Borah, S. (2021). A data-driven approach to anomaly detection in urban air pollution monitoring. *Environmental Research, 201*, 111522. DOI: https://doi.org/10.1016/j.envres.2021.111522
18. Patel, P., Joshi, A., & Patel, D. (2019). A survey on anomaly detection techniques in IoT-based air pollution monitoring systems. *Procedia Computer Science, 155*, 605–610. DOI: https://doi.org/10.1016/j.procs.2019.08.086
19. Qi, J., Li, C., Zhu, F., & Wu, C. (2018). A novel air pollution anomaly detection framework using deep learning models. *IEEE Transactions on Environmental Monitoring, 12*(5), 2475–2485. DOI: https://doi.org/10.1109/TEM.2018.2865974
20. Rai, P., & Singh, S. (2010). A survey of clustering techniques for anomaly detection. *Artificial Intelligence Review, 30*(2), 87–126. DOI: https://doi.org/10.1007/s10462-010-9161-y
21. Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation, 13*(7), 1443–1471. DOI: https://doi.org/10.1162/089976601750264965
22. Sharma, S., Jain, S., & Gupta, M. (2021). A hybrid deep learning framework for air pollution prediction. *Sustainable Cities and Society, 74*, 103239. DOI: https://doi.org/10.1016/j.scs.2021.103239
23. Sun, L., Xu, G., Li, Z., & Zhang, C. (2019). Anomaly detection in air pollution monitoring: A case study using real-time IoT data. *Sensors, 19*(3), 788. DOI: https://doi.org/10.3390/s19030788
24. Xie, Y., Wang, Y., & Yu, J. (2022). Deep learning-based anomaly detection for air pollution monitoring using Autoencoders and LSTM. *IEEE Transactions on Neural Networks and Learning Systems, 34*(1), 347–362. DOI: https://doi.org/10.1109/TNNLS.2022.3153748
25. Zhang, Y., Liang, W., & He, X. (2021). Big data analytics for air quality anomaly detection: A review. *Journal of Big Data, 8*(1), 23. DOI: https://doi.org/10.1186/s40537-021-00419-8
26. Dahiya, P., & Srivastva, D. K. (2019). An Efficient Anomaly Detection Based On Optimal Deep Belief Network in Big Data. In International Journal of Engineering and Advanced Technology (Vol. 9, Issue 1, pp. 708–716). DOI: https://doi.org/10.35940/ijeat.f9178.109119
27. Shanthi, Dr. S., & Pyingkodi, M. (2019). Air Quality Index Prediction using Machine Learning Algorithms. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 4, pp. 7489–7492). DOI: https://doi.org/10.35940/ijrte.d5326.118419
28. Ezekiel, S., Alshehri, A. A., Pearlstein, L., Wu, X.-W., & Lutz, A. (2020). IoT Anomaly Detection using Multivariate. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 4, pp. 1662–1669). DOI: https://doi.org/10.35940/ijitee.d1323.029420
29. Rathore, R., & Shrivastava, Dr. N. (2023). Network Anomaly Detection System using Deep Learning with Feature Selection Through PSO. In International Journal of Emerging Science and Engineering (Vol. 11, Issue 5, pp. 1–6). DOI: https://doi.org/10.35940/ijese.f2531.0411523

AUTHOR'S PROFILE

Raghav Abrol is a data-driven AI researcher and analyst based in New Delhi, India. He holds a Master's degree in Computer Science with a specialization in Artificial Intelligence from Netaji Subhas University of Technology and a Bachelor's in Computer Science from Dr. Akhilesh Das Gupta Institute of Engineering & Technology. With professional experience at Tata Consultancy Services and as an AI Research Intern at C-DOT, Raghav has developed scalable AI and ML solutions and contributed to ethical AI deployment frameworks. His skills include Python, SQL, TensorFlow, PyTorch, Power BI, and model monitoring using IBM UQ360 and Why Labs. Passionate about environmental sustainability and innovation, he leverages data science for public health, air quality forecasting, and operational optimization. His recent research focuses on anomaly detection in air pollution using machine learning techniques.