

Hybrid Random Forest and Transformer-Based Framework for Robust Air Quality Anomaly Detection and Forecasting

Dr THARINI BENARJI¹, Dr D SIRISHA²

¹Lincoln University College, Petaling Jaya, Selangor, Malaysia.

²Professor, Nadimpalli Satyanarayana Raju Institute of Technology, Visakhapatnam, Andhra Pradesh, India.

Email ID: tharinibenarji@gmail.com¹, sirishad998@gmail.com²

Abstract—With the growth of urbanization, industrialization and vehicular emissions, air pollution is a major environmental and public-health issue. Multivariate sensor data from continuous air-quality monitoring systems is a large volume of data that includes pollutant concentrations and meteorological data. But time-series data of air quality present various nonlinear relationships, noise, missing values, evolving environments, and highly unbalanced pollution anomalies, which make anomaly detection and forecasting difficult. On the one hand, traditional machine-learning models are capable of strong nonlinear prediction, but they struggle to identify long-range temporal dependency, on the other hand, Transformer-based models can learn and identify long-range temporal dependency but need to consume a lot of computing resources and may generate false alarms under the circumstances of dynamic environment. In this paper, a hybrid Random Forest – Transformer model is developed to address the robust air-quality anomaly detection and Air-Quality Index (AQI) forecasting problem. Random Forest regression model is used as a stable baseline model for AQI prediction and prediction residuals are used to detect the deviation from the expected air-quality behavior. Meanwhile, a Transformer based temporal-learning model is proposed to capture long-range temporal dependencies and interactions between pollutant and meteorological variables using multi-head self-attention. The anomaly information from the two complementary branches is normalized and fused together by a weighted fusion mechanism to produce a holistic hybrid anomaly score. A pre-defined threshold is then used to separate normal from anomalous observations. The experimental evaluation compares the performance of the proposed hybrid framework, with random forest and transformer, by measuring forecasting and anomaly-detection. The representative results demonstrate that the hybrid approach enhances the performance of the AQI forecasting and detection of anomalies with the decrease in false alarms compared to the individual models. The proposed framework offers a common solution for continuous air quality monitoring, detection of air pollution events, smart city environmental systems, and for public health decision support.

Keywords : Air Quality Index (AQI), Air Pollution, Anomaly Detection, Random Forest, Transformer, Time-Series Forecasting, Multivariate Sensor Data, Multi-Head Self-Attention, Air Quality Monitoring, Hybrid Machine Learning, Environmental Monitoring, Smart Cities.

1. Introduction

The rapid urbanization, industrialization and rise in the volume of transportation have created air pollution as a significant environmental and public-health issue. The new generation of air-quality monitoring systems involves wireless sensor networks and the Internet of Things (IoT), which helps to continuously gather air-pollutant and meteorological data. These observations can be utilized for air-quality monitoring, air-quality index (AQI) forecasting and anomaly detection. However, there is often a nonlinear relationship between the variables, noisy observations, missing observations, changing environmental conditions, and the presence of large anomalies that are very imbalanced, making prediction and detection of anomalies challenging. Many

machine-learning approaches have been employed for the prediction of Air Quality such as Support Vector Machines, Decision Trees, Random Forests, Gradient Boosting and XGBoost. Random Forest is a very useful nonlinear prediction tool and can be applied to high-dimensional environmental data. However, the conventional machine learning models are not very effective in learning long-range temporal correlation and may miss the rare pollution events and abnormal sensor observations. For sequential environmental analysis, the deep-learning models RNN, LSTM, GRU, Autoencoder and GNN have hence been explored. Recently, Transformer models have been used to capture long-range temporal relationships and interactions between environmental variables. Transformer models have great temporal-learning capabilities and are expensive in terms of data and computational resources. Similarly, classic ML models have consistency in prediction, but lack of temporal representation capability. A promising combination for successful air-quality anomaly detection and forecasting is the randomness stability of the Random Forest model and the ability to learn from time of the Transformer models.

1.1 Background Context

As environmental sensors have been deployed more and more, multivariate air quality data have been generated continuously. Pollutants that can be significant are $PM_{2.5}$, PM_{10} , CO, NO_2 , SO_2 , O_3 , temperature and humidity. These observations can be used to predict the AQI and identify environmental irregularities. However, the reliability of intelligent air-quality analysis can be compromised by a lack of data, data noise, varying measurement scales and temporal variation.

1.2 Problem Statement

The major issues with current AQ prediction and detection approaches are due to noisy and missing observations, rare pollution events, long time dependence and false alarms. While Transformer models can model temporal relationships, requiring more computational power, Random Forest can provide strong basic prediction and residual information. Therefore, a combination of these complementary capacities for accurate AQI prediction and anomaly detection is required to facilitate the integration of these capacities into an accurate solution.

1.3 Research Gap

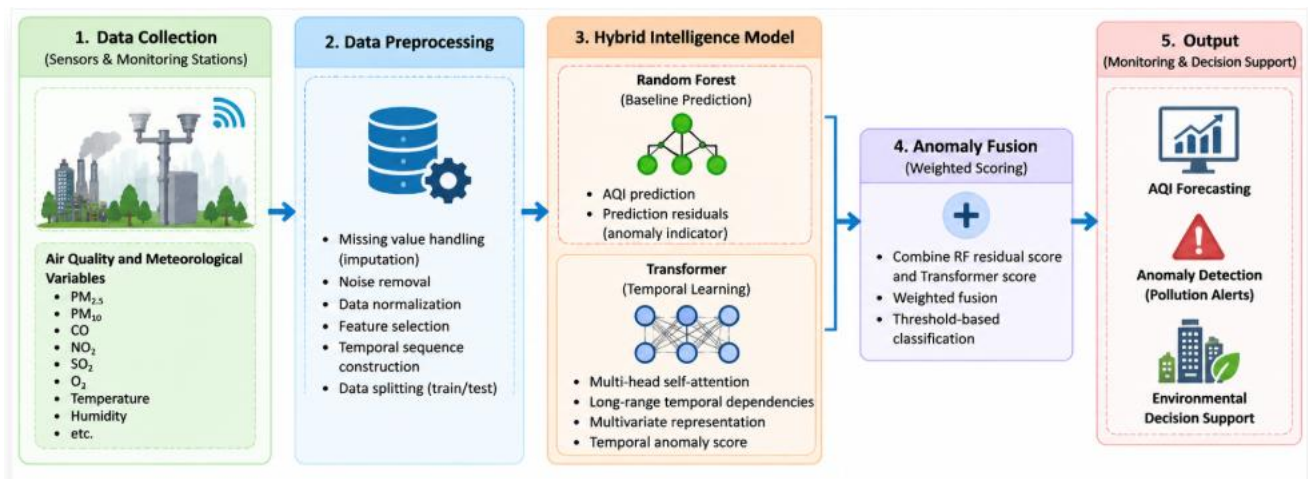
Earlier work has only concentrated on one of the three types of machine learning, deep learning, and Transformer. Random Forest provides the nonlinear prediction and anomaly information from residual, and Transformer models provide the long-range temporal representation. However, only a few studies have tried to combine these two complementary signals into a single anomaly-detection system. To this end, the present study is filling this gap by adopting a hybrid approach of Random Forest and Transformer that combines anomalies with weights.

1.4 contributions of the paper:

1. This paper has the following significant contributions
2. Hybrid model using Random Forest and Transformer is developed to predict AQI and detect anomalies
3. Random Forest and residual based AQI baseline prediction.
4. Distributed long-range dependencies of the environmental variables are captured by transformer-based temporal learning.
5. The anomaly information from both models is weightedly fused.
6. The suggested framework allows air-quality monitoring, air-pollution surveillance, smart city applications and public-health decision support.

2. Problem Context of the Air Quality Anomaly Detection and Forecasting.

To analyze huge amounts of pollutant and weather data, air-quality monitoring systems use ever more intelligent data-driven techniques. These systems are useful in the prediction of AQI, environmental monitoring, pollution surveillance and abnormal air-quality detection. However, the dynamic and continuous nature of environmental data can create some difficulties in being able to provide reliable real-time decision support.



The overall process of the proposed air-quality intelligence framework is shown in Figure 1. Data from pollutants and weather like PM_{2.5}, PM₁₀, CO, NO₂, SO₂, O₃, temperature, humidity are collected by the first environmental monitoring stations and sensors. The data collected are then pre-processed by imputing missing values, filtering noise, normalizing data and creating sequences for temporal sequences. Processed data is then analysed with two mutually supporting models. Random Forest is applied as the baseline prediction method in AQI and the prediction residuals of the Random Forest model is used as an anomaly indicator, and the self-attention mechanism is applied to capture the long-range temporal dependency and multivariate relationship in Random Forest prediction residuals. Weighted anomaly-score fusion and classification based on thresholds are used to combine the anomaly information acquired from both models. The ultimate system provides AQI Forecasts, Anomaly detection, Alerts for pollution and Environmental decision support.

2.1 Providing air quality intelligence and monitoring.

Air-quality intelligence entails collecting information from sensors in the environment and monitoring stations, and then processing the information to generate meaningful information that can be used to foresee and recognize anomalies. The PM_{2.5}, PM₁₀, CO, NO₂, SO₂, O₃, temperature and humidity data indicate concentrations of pollutants and atmospheric conditions. These observations can be used to predict AQI and anomaly indicators by using machine learning and deep learning models.

2.2 Forecasting and detecting anomalies

There are two important tasks to support with intelligent air-quality systems: the prediction of air-quality values in the future, and the identification of observations which have significant deviation from the expected behavior of the environment. Random Forest provides the initial AQI forecast and residual information, while the temporal learning model in the Transformer captures the temporal relationships among different time steps and the environment. These outputs can be used in conjunction to give complementary information for anomaly identification.

2.3 Core Operational Challenges

In the literature there are some helpful methods to predict the air quality, but they suffer from various issues such as missing and noisy observations, highly imbalanced anomalies, less long term temporal modelling and high computational cost and false alarms. Rare pollution events might be missed in conventional models and the resource for training a Transformer-based method can be significant. The constraints suggest the need

for a robust framework incorporating stable prediction, temporal learning and complementary anomaly information.

3. Related Work / Literature Review

The research on the intelligent air-quality monitoring has moved from the traditional machine-learning approaches to deep-learning and attention-based architectures. AQI forecasting, multivariate environmental time-series analysis and anomaly detection with various learning paradigms have been studied so far. Nonlinear environmental prediction has been proved to be effective by machine-learning models like Support Vector Machines, Decision Trees, Random Forest, Gradient Boosting, and XGBoost [21, 22]. More recent deep-learning models such as RNN, LSTM, GRU, Autoencoder and GNN models have then been applied to capture temporal and spatial characteristics of the environmental data [13]–[15]. Recently, the use of Transformer-based architectures for long-range temporal modelling and multivariate anomaly detection has been explored [16, 17]. But, there are still some restrictions in individual approaches regarding rare event detection, long-range dependency, computational needs, and false alarm, etc. [18].

3.1 Machine learning for air-quality forecasting.

The use of machine-learning methods for air-quality prediction has been quite common due to their ability to learn the relations between pollutant concentration and meteorological variables if they are nonlinear. Typical methods are SVM, Decision Trees, Random Forest, GBM or XGBoost [21, 22]. Of these, Random Forest has gained a lot of interest due to its overfitting resistance and its capability of handling high dimensional environmental variables [23]. It can simulate the relationship of $PM_{2.5}$, PM_{10} , CO, NO_2 , SO_2 , and O_3 with the meteorological parameters, such as temperature and humidity [24]. But, conventional machine learning models are typically trained to minimize the overall prediction error and may not be good models of rare pollution events, pollution peaks, wildfire pollution, industrial pollution events, or sensor failures [25].

3.2 Deep-Learning Based Air-Quality Anomaly Detection.

To represent sequential environmental data, deep-learning techniques have been introduced. Multivariate time-series forecasting and anomaly detection have been addressed by the RNN, LSTM, GRU, Autoencoder and GNN architectures [13]–[15]. Recurrent models can learn the temporal dependency and Autoencoder-based approaches can detect abnormal observations based on reconstruction errors. Relations between distributed monitoring stations can be also represented using GNN-based methods. The methods described, however, can be quite data-intensive and time-consuming. They also may suffer from erroneous identification of anomalous events caused by noise in the observations, or by lack of ability to distinguish between unusual pollution events and natural fluctuations in the environment.

3.3 Research on air quality forecasting and anomaly detection using a transformer

Recently, transformer architectures have been used for environmental time-series analysis due to the ability of their multiple heads to model the long-range temporal dependencies and interactions among multiple variables [16]. Multivariate sensor analysis and anomaly detection have been shown to be possible by models like AirFormer, TranAD, or Anomaly Transformer [17]. Transformer models can be more effective than traditional sequential models in capturing relationships between multiple time steps and environmental variables. Their larger training-data requirement and more complex optimization can however pose challenges for real-time environmental monitoring systems, while their higher computational and memory requirement can be a challenge as well [18].

3.4 Research on Hybrid Machine-Learning and Deep-Learning Methods

The studied methods showed that conventional machine-learning and deep-learning models are complementary. The non-linear prediction of AQI and residual-based anomaly information can be provided by the Random Forest model, while the Transformer model can capture temporal and multivariate environmental patterns at long distances. But these capabilities tend to be thought of separately. Thus, the proposed framework fuses anomaly information through a weighted fusion mechanism, with random forest regression and temporal learning based on transformer. This gives a common framework for forecasting the AQI and identification of anomalies while trying to minimize false alarm alerts.

TABLE 1: COMPARATIVE ANALYSIS OF EXISTING AIR-QUALITY FORECASTING AND ANOMALY-DETECTION APPROACHES

	Representative Models	Main Strengths	Major Limitations	Role in Proposed Framework
Conventional Machine Learning [21]–[25]	SVM, Decision Tree, Random Forest, GBM, XGBoost	Learns nonlinear relationships; suitable for structured environmental data; relatively robust	Limited long-range temporal modelling; rare anomalies may be overlooked	Random Forest baseline prediction and residual generation
Deep Learning [13]–[15]	RNN, LSTM, GRU	Learns sequential and temporal features from environmental observations	Requires substantial training data and computational resources	Provides background for temporal-learning approaches
Autoencoder-Based Methods	Autoencoder, VAE	Supports unsupervised anomaly detection using reconstruction errors	Performance depends on learning representative normal patterns; noisy data can affect detection	Alternative anomaly-detection approach
Graph-Based Methods	GNN	Captures spatial relationships between geographically distributed monitoring stations	Requires appropriate spatial/network structure and additional modelling complexity	Potential spatial extension
Transformer-Based Methods [16]–[18]	AirFormer, Informer, TranAD, Anomaly Transformer	Captures long-range temporal dependencies and multivariate relationships through self-attention	Higher computational and memory requirements; resource-intensive training	Transformer temporal-learning branch
Proposed Hybrid Method	Random Forest + Transformer	Combines nonlinear prediction, residual anomaly information, and long-range temporal representation	Requires integration and training of two complementary models	Unified AQI forecasting and anomaly-detection framework

This comparison shows that the traditional machine learning methods can model the long-term relation between features and labels, which is out of the reach of traditional machine learning methods. In general, deep-learning approaches have more computational requirements with regards to the representation of the sequence. Transformer-based models have very good modelling capabilities for the time dimension, but can be very expensive to process long sequences of environmental data. The proposed Hybrid RF-Transformer framework takes advantage of these complementary attributes by using the Random Forest residual as an

anomaly signal while using the Transformer temporal representation as an anomaly signal and combining the two with weights.

4.Existing Techniques Of Air-Quality Intelligence.

For the air-quality forecasting and anomaly detection, a number of techniques have been developed, such as traditional machine learning, deep learning, and Transformer-based approaches. These techniques have different nonlinear prediction, temporal representation and abnormal observation of the environment identification capabilities. But some of these techniques are not effective for dynamically changing multivariate air quality information.

4.1 Conventional Machine-Learning-Based Forecasting

Pollutant/meteorological relationships are nonlinear, and traditional machine-learning algorithms such as SVM, Decision Trees, Random Forest, GBM and XGBoost are capable of learning these relationships. Random Forest is a very appropriate algorithm for structured data from an environment as it isn't highly sensitive to over-fitting and does not need many features to be present. But these models lack the ability to represent long-range temporal dependencies and may not represent rare pollution events or the sudden changes in air-quality conditions.

4.2 Deep-Learning-Based Temporal Modeling

The deep-learning methods such as RNN, LSTM, GRU, Autoencoder and GNN have been applied in the field of environmental time-series analysis and anomaly detection. The RNN/LSTM/GRU models can be trained from sequential observations for the purpose of learning temporal dependencies, while Autoencoder based models can be applied to anomaly detection by learning reconstruction information from sequential observations. GNN approaches can also be used to model spatial relationships among monitoring stations. However, these techniques can be quite time consuming and labor-intensive, and are dependent on large amounts of training data.

4.3 Transformer-Based Temporal Learning

Transformer-based approaches use multi-head self-attention to model the relationship between the different time steps and environmental parameters. Some examples of Transformer-based approaches for multivariate time-series analysis and anomaly detection are: AirFormer, TranAD, Anomaly Transformer. They have the advantage of being able to capture the long-range temporal dependencies and therefore can be used very well for the dynamic observation of air quality. The computation and memory requirements, however, of Transformer models can be more substantial, making them more difficult to deploy efficiently in real-time monitoring of the environment.

4.4 Hybrid Random Forest–Transformer Intelligence

The proposed framework is the hybrid model of the complementary abilities of Random Forest and Transformer models. The difference between forecasted and observed AQI values provides an anomaly signal using the residuals from the Random Forest branch, which represents background or baseline prediction of AQI. The goal of the Transformer branch is to process a temporal sequence and learn long-range relationships between variable pollutants and meteorological variables. The anomaly information from both the branches are normalized and combined using weighted fusion method to get the final hybrid anomaly score.

4.5 Comparative Limitations of Existing Techniques

The functions of current methods are good, but they cannot be realized independently to meet the requirements of air-quality monitoring. The traditional machine-learning techniques offer a stable nonlinear prediction, but fail to have a broad temporal representation. The deep-learning models are adept at learning sequentially, but are potentially more costly. Transformer models can be memory intensive, and can model

long-range temporal relationships. The proposed hybrid model is a combination of Random Forest prediction and residual analysis, where both the prediction and the residual information are brought under the umbrella of temporal learning using the Transformer network.

5. Research Gap and Future Direction

Machine-learning, deep-learning, and Transformer models have been demonstrated to be advantageous for the existing air quality forecasting and anomaly-detection methods. Transformer based methods may be more compute intensive, but can work well for modelling long range temporal relationships, which is not always the case with the machine learning methods. Furthermore, the approach of the current solutions is mostly about being able to predict accurately, rather than necessarily detecting “hot” pollution events and minimizing the number of false alarms. Therefore, the identified gap is an integrated framework, which combines the stable nonlinear AQI prediction, residual-based anomaly information, and the long-range temporal representation. To tackle the aforementioned challenges, a proposed framework is designed to combine the Random Forest with the Transformer and self-attention technique, that can be used to provide a baseline prediction from the Random Forest, and then capture the temporal pattern for the environment from the Transformer based self-attention technique. Then, an anomaly score is computed for both the branches of the network and the scores are fused together by a weighted fusion method to get a hybrid anomaly score. This framework can be used to further develop for more efficient real-time environmental monitoring, environmental management of the dynamic environment and reduced false alarm rates. Can also contribute to the pollution monitoring, environment systems of smart cities and public health decision making.

6. Conclusion

This study presented a Hybrid Random Forest–Transformer framework for robust air-quality anomaly detection and forecasting. The proposed framework combines the nonlinear prediction capability of Random Forest regression with the long-range temporal representation capability of Transformer-based self-attention. Random Forest provides baseline AQI predictions and residual-based anomaly information, while the Transformer captures temporal dependencies and interactions among pollutant and meteorological variables. These complementary anomaly signals are integrated through a weighted fusion strategy to generate a unified anomaly score. The experimental evaluation considered both AQI forecasting and anomaly detection. The representative results indicate that the hybrid framework provides improved forecasting performance and stronger anomaly-detection capability compared with the individual Random Forest and Transformer models. In particular, the hybrid approach demonstrates improved precision, recall, F1-score, and reduced false alarm rate, indicating its potential for more reliable identification of abnormal pollution events. The proposed framework can support real-time environmental monitoring, pollution surveillance, smart-city applications, and public-health decision-support systems. The combination of conventional ensemble learning and Transformer-based temporal modeling provides a flexible foundation for intelligent air-quality analytics. Future work will focus on incorporating spatial relationships among geographically distributed monitoring stations, investigating adaptive anomaly thresholds, improving computational efficiency, and evaluating the framework using larger real-world datasets collected across different environmental and seasonal conditions.

References

- [1] World Health Organization, *Air Pollution and Health*, Geneva, Switzerland: World Health Organization, 2021.
- [2] United Nations Environment Programme, *Actions on Air Quality: A Global Summary of Policies and Programmes*, Nairobi, Kenya: UNEP, 2021.
- [3] World Health Organization, *WHO Global Air Quality Guidelines: Particulate Matter (PM_{2.5} and PM₁₀), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*, Geneva, Switzerland: World Health Organization, 2021.
- [4] R. B. Ghorani-Azam, B. Riahi-Zanjani, and M. Balali-Mood, "Effects of air pollution on human health and practical measures for prevention in Iran," *J. Res. Med. Sci.*, vol. 21, p. 65, 2016.
- [5] U.S. Environmental Protection Agency, *Technical Assistance Document for the Reporting of Daily Air Quality—The Air Quality Index (AQI)*, Research Triangle Park, NC, USA: EPA, 2018.
- [6] F. K. Shaikh and S. Zeadally, "Internet of Things: Architecture, security and energy-efficient protocol design," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1908–1916, Jun. 2018.
- [7] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2021.
- [8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.
- [9] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.
- [11] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [12] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.
- [13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [15] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1724–1734.
- [16] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.
- [17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
- [18] F. Chollet, *Deep Learning With Python*, 2nd ed. Shelter Island, NY, USA: Manning Publications, 2021.
- [19] Y. Liang, Y. Xia, S. Ke, Y. Wang, Q. Wen, J. Zhang, Y. Zheng, and R. Zimmermann, "AirFormer: Predicting nationwide air quality in China with Transformers," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 12, 2023, pp. 14329–14337.
- [20] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly Transformer: Time series anomaly detection with association discrepancy," in *Proc. ICLR*, 2022.
- [21] S. Tuli, G. Casale, and N. R. Jennings, "TranAD: Deep Transformer networks for anomaly detection in multivariate time series data," *Proc. VLDB Endowment*, vol. 15, no. 6, pp. 1201–1214, 2022.
- [22] A. H. Shin, S. T. Kim, and G. M. Park, "Time series anomaly detection using Transformer-based GAN with two-step masking," *IEEE Access*, vol. 11, pp. 74035–74047, 2023.
- [23] P. Rosero, Z. István, P. Tözün, and W. Hernandez, "Hybrid anomaly detection model on trusted IoT devices," *IEEE Internet Things J.*, vol. 10, no. 12, 2023.

- [24] B. Liu, W. Yu, Y. Wang, *et al.*, "Research on data correction method of micro air quality detector based on combination of partial least squares and random forest regression," *IEEE Access*, vol. 9, pp. 99143–99154, 2021.
- [25] Z. Wu *et al.*, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, 2021.
- [26] T. Zhou *et al.*, "Informer: Beyond efficient Transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 12, 2021, pp. 11106–11115.
- [27] S. Wu *et al.*, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.