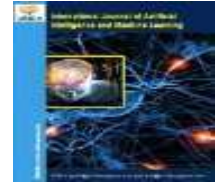




International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Identifying Anomalies in Air Quality Prediction Using Machine Learning and Transformer-Based Techniques

¹Renuka P, ²Dr Nilesh Parihar, ³Dr Tharini Benarji

¹Ph.D Scholar, Department of Electronics & Communication Engineering, Gandhinagar Institute of Research and Development, Gandhinagar University, Gujarat renoostar@gmail.com

²Professor & HOI, Electronics & Communication Engineering, Gandhinagar Institute of Technology, Gandhinagar University, Gujarat nileshparihar.sec@gmail.com

³Professor, Vice Principal & HOD, Computer Science & Engineering, Indur Institute of Engineering and Technology, Ponnala, Siddipet, TS tharinibenarji@gmail.com

ABSTRACT

Air pollution has become a significant environmental and public health issue due to the rapid industrialization, urbanization, increasing transport usage and changes in climate. The prediction of Air Quality Index (AQI) and the early detection of abnormal pollution situations are critical for protecting public health and supporting environmental decision-making. Although the conventional machine learning methods are effective in forecasting the overall AQI, they often fail to capture rare and irregular phenomena resulting from sharp peaks in emissions, major changes in weather conditions, sensor failures, and unexpected environmental events. To overcome these limitations, this work proposes a Hybrid Air Quality Anomaly Detection Framework which is composed of a Temporal learning model based on a Transformer deep learning model and a Random Forest (RF) model. The multi-head self-attention structure of the Transformer can capture the long-range temporal dependency and complex relationships among multiple environmental parameters, while the Random Forest algorithm can predict the AQI with stable baseline values based on concentrations of pollutants and meteorological variables. The outputs of both models are combined using an ensemble anomaly scoring technique to reduce false alarm rates and improve the detection of anomalous air quality occurrence. Benchmark air quality datasets with measurements of pollutants like PM_{2.5}, PM₁₀, CO, NO₂, SO₂, and O₃ as well as meteorological variables like temperature, humidity, wind speed, and atmospheric pressure are used to apply and assess the suggested framework. The experimental results demonstrate the effectiveness of the proposed hybrid framework in terms of the number of false positives and the accuracy of the anomaly detection, as well as in terms of precision, recall and F1 score, when compared with the existing methods based on machine learning and stand-alone deep learning. The proposed system provides a scalable, intelligent, and reliable solution for smart environmental management, early warning of pollution, and real-time air quality monitoring.

Keywords: Air Quality Index (AQI), Anomaly Detection, Random Forest, Transformer Networks, Machine Learning, Deep Learning, Time-Series Analysis, Ensemble Learning, and Smart Air Quality Prediction.

I. Introduction

Air pollution has become one of the most severe environmental and public health challenges in the world due to rapid urbanization, industry, population growth and rising number of motor vehicles. The emission of hazardous pollutants like PM_{2.5}, PM₁₀, CO, NO₂, SO₂ and O₃ has severely deteriorated the air quality in developed and developing countries around the world. Poor air quality has been directly tied to lung cancer, cardiovascular disease, respiratory diseases, reduced life span and millions of premature deaths each year. Intelligent pollution management systems and accurate air quality monitoring are important, as the World Health Organization (WHO) reported that air pollution remains one of the major environmental risk factors to human health [1, 2]. Air Quality Index (AQI) has become the accepted standard across the industry to measure and communicate air pollution intensity. The AQI is a composite of the concentrations of various air pollutants and represents the overall air quality in the surrounding area and the health risks associated with it. Smart city managers and environmental agencies employ AQI values to aid in environmental policy decisions, pollution control measures, and public health warnings. Accurate AQI prediction can help authorities take timely mitigation and prevention actions to minimize the health impacts and environmental damage resulting from hazardous pollution episodes before they occur [3], [4].

Today, environmental monitoring systems are collecting large quantities of multivariate time-series data from satellite observation systems, IoT-based air quality sensors and weather stations. These monitoring systems measure pollutant concentrations as well as atmospheric variables such as temperature, humidity,

wind direction, wind speed, rainfall and air pressure. Large-scale environmental datasets have made it possible to utilize techniques of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) to enhance air quality prediction and anomaly detection [5] [6]. The traditional statistical forecasting techniques like Autoregressive Integrated Moving Average (ARIMA), Multiple Linear Regression (MLR) and Seasonal ARIMA (SARIMA) have been widely used for air quality forecasting due to their ease of use and interpretability. Nevertheless, these approaches often do not resolve the nonlinear interactions that are commonly found between contaminants and meteorological conditions and generally assume linear relationships between environmental factors. Because of this, when environmental conditions change quickly or unanticipated pollution episodes take place, their prediction accuracy drastically declines [7], [8].

By identifying intricate nonlinear associations from past environmental data, machine learning algorithms have greatly enhanced AQI prediction. Amongst the models that have been tested in the study of pollutant behavior and estimation of AQI levels, the following have demonstrated promising results: Decision Trees, Support Vector Machines (SVM), Random Forest (RF), Gradient Boosting and Extreme Gradient Boosting (XGBoost). The Random Forest algorithm has garnered significant attention because of its high prediction accuracy, ability to avoid overfitting, resilience to noise, and ability to process high-dimensional data. Random Forest, however, often fails to detect long term temporal relationships of continuously varying environmental data, as it primarily looks at static feature associations. This makes it difficult to detect unusual environmental phenomena, unnormal atmospheric states and sudden pollution peaks only through the use of conventional machine learning algorithms [9, 10]. Transformer-based architectures, which employ self-attention processes, have transformed sequential data analysis thanks to recent developments in deep learning. Unlike conventional Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks which perform step-by-step over sequential observations, transformer networks allow for processing of entire sequences to perform effective learning of long-range temporal dependencies while reducing the computational time. Transformer models have originally been developed for NLP, but in recent years, they've demonstrated significant efficacy in forecasting time series, detecting anomalies, predicting traffic, forecasting financial data, analyzing medical information, and monitoring the environment [11; 12; 13].

Transformer networks have proven to be effective when modelling sequential environmental data, but there are numerous challenges that still prevent the use of real-time air quality anomaly detection. Presently, the Transformer-based methods require a substantial amount of labeled training data and are also fairly complex, which makes them difficult to apply to environmental monitoring systems that have limited resources. Furthermore, most of the available studies are concerned with enhancing the predictive accuracy of the AQI instead of spending much less time discovering the anomalies in pollution events, such as the unexpected pollution incident, sudden pollution peaks, changes in weather and sensor failures. They are rare events which occur infrequent in historical databases and are not commonly predicted by conventional prediction algorithms. As a result, if considering solely deep learning architectures or traditional machine learning models, it can lead to low reliability of the early warning systems and delayed anomaly detection [14], [15] and [16]. A thorough literature review was done and it was noted that anomaly detection was typically done using simple thresholding methods or using a statistical residual analysis with most machine learning methods concentrating on predicting the accuracy. These methods have computational efficiency, but also generate false alarms and miss minute pollution anomalies that may occur over a long period of time. Similarly, there are a number of deep learning methods that enhance the temporal modeling, but these are also costly in terms of computational complexity and time to train, and are not interpretable. It is uncommon in the classical machine learning method to detect anomalies using a temporal learning capability of transformer networks. Hence, a hybrid system that manages to forecast the AQI and to detect unusual pollution events in real time with low false alarms is still very much desired [17],[18] and [19].

To overcome these drawbacks, the authors introduce a Hybrid Air Quality Anomaly Detection Framework, which combines a temporal learning model using the transformer network and a Random Forest (RF) model. The multi-head self-attention mechanism of the Transformer network can effectively extract the long-term temporal dependency of data and complex relationships among environmental parameters, and the Random Forest algorithm is used to predict the baseline AQI based on pollutant values and meteorological elements to generate a consistent baseline AQI forecast. An ensemble anomaly scoring system, utilizing the results of both models is developed to provide reliable indicators of anomalies that could be used to warn of abnormal air quality conditions in advance. This Hybrid design also contributes to better identification of sudden pollution peaks, weather conditions and sensor anomalies, and more stability in the forecast.

The proposed structure makes use of environmental information provided by air quality monitoring stations, including meteorological data like temperature, humidity, wind speed and direction, and air quality pollutant concentrations, such as PM_{2.5}, PM₁₀, CO, NO₂, SO₂, and O₃. Collected data is preprocessed using techniques such as missing value imputation, data normalization, noise reduction, and feature extraction to enhance the quality of the data and ensure accurate predictions during training. Both models are executed in parallel and the results of them are fused together on the basis of ensemble fusion technique that helps to reduce the false positive predictions and increase the accuracy of the anomaly detection. The proposed system aims to help facilitate real-time environmental monitoring for quick detection of hazardous pollution incidents and enable environmental organizations to send environmental health alerts promptly.

Following is an outline of the key findings of this study:

1. The use of a hybrid model comprised of a Random Forest and Transformer is proposed to predict the Air Quality Index simultaneously and detect anomalies.
2. For capturing long-range dependencies and complex interactions between pollutant and climatic factors, a Transformer-based temporal learning technique is introduced.
3. In order to increase the accuracy of the detection, an ensemble anomaly scoring method is designed by combining two anomaly detection algorithms: the temporal anomaly detection algorithm based on a Transformer and the residual analysis algorithm based on the Random Forest algorithm.
4. The standard air quality data sets are used to evaluate this proposed framework, and it is compared to the conventional machine learning and deep learning models, in terms of accuracy, precision, recall, F1 score and false alarm rate.
5. This proposed system is an intelligent, flexible and reliable solution for managing air quality, warning of pollution and monitoring the environment at real-time in smart cities.

II. SURVEY OF RESEARCH

[1] Bansal et al.

presented an extensive framework to predict air quality. It utilizes multiple machine learning and deep learning techniques, including Random Forest, XGBoost, Long Short-Term Memory (LSTM), Transformer and Temporal Graph Neural Networks (TGNN). This framework takes into account temporal and spatial dependencies concerning air quality, based on pollutant concentrations, meteorological parameters, and data from IoT sensors and satellite imagery. Experimental evaluation shows that, in industrial districts, the accuracy of the predictions by TGNNs is better than the accuracy of the predictions by Transformer models, since the spatial linkages between monitoring stations are captured by TGNNs, but not by Transformer models; on the other hand, Transformer models achieve a better prediction accuracy in urban environments. The suggested approach was found to be robust in the scenarios of severe pollution and missing data. However, they were not practical for deployment in real-time due to significant increase in computation complexity and memory requirement that was caused by the fusion of Transformer and graph-based models. The study focuses on the need for hybrid approaches in realistic air quality monitoring systems with increased computational efficiency while demonstrating the effectiveness of complex deep learning models.

[2] Abrol (2025)

An artificial intelligence-based anomaly detection system for intelligent air pollution monitoring with virtual environmental stations was introduced by Abrol [2]. The proposed method combines machine learning techniques such as Isolation Forest, Autoencoders, and Long Short-Term Memory (LSTM) networks with statistical analyses like Z-score analysis to detect abnormal pollution events. Extreme pollutant deviations are found using Z-score analysis. Isolation LSTM models catch temporal variations in pollution patterns, Forest detects anomalies, and autoencoders detect flaws in the reconstruction. The built-in system is suitable for smart environmental monitoring applications because it demonstrated that it could detect anomalies with a greater capability than other sensors needing expensive physical monitoring systems. Although the proposed method is scalable and cost-effective, the effectiveness of the method is largely affected by the availability of the sensor and its hyperparameter tuning, and the quality of the data. Furthermore, it is not well validated with large amounts of real world data, which reduces the ability to be generalized. The work emphasizes the importance of how to combine machine learning and statistical approaches for intelligent detection of environmental anomalies.

[3] Borah (2025)

Borah et al. [3] introduced a deep learning-based approach for monitoring air pollution that involves a Bidirectional Long Short-Term Memory (Bi-LSTM) network and self-attention mechanism. The proposed model is based on multivariate air quality data and is capable of predicting the typical behavior of the pollutants to detect abnormalities via prediction error analysis. The self-attention mechanism is used to select the essential time duration related to abnormal pollution phenomena, and bidirectional learning can

simultaneously capture the historical and future temporal relationships. Higher anomaly identification accuracy than a number of conventional machine learning techniques was shown by experimental results, especially when it came to detecting abrupt pollutant spikes and erratic atmospheric circumstances. Although the model has tremendous prediction capability, it was tested with a region-specific dataset and was not applicable to other geographical locations with different pollutant characteristics. Moreover, computational requirements of deep recurrent architectures could also pose a challenge for the implementation in the monitoring system with limited resources. The study demonstrates that deep learning with attention can be effective for intelligent air quality anomaly identification, however.

[4] Ariffa Begum (2024)

Ariffa Begum [4] proposed a real-time air quality prediction and anomaly detection framework using Internet of Things (IoT)-enabled environmental sensors integrated with machine learning techniques. The system continuously acquires pollutant concentrations, including PM_{2.5}, PM₁₀, CO, NO₂, SO₂, and O₃, together with meteorological parameters for real-time monitoring. Forecasting models such as Linear Regression and Long Short-Term Memory (LSTM) are employed to estimate future Air Quality Index (AQI) values, while anomaly detection algorithms identify sudden pollution spikes and abnormal environmental conditions. A visualization dashboard provides real-time pollution status and alerts to assist environmental agencies in decision-making. Experimental evaluation demonstrated the feasibility of integrating IoT technology with predictive analytics for continuous environmental monitoring. However, the forecasting capability is limited by the relatively simple prediction models, which are unable to capture complex temporal dependencies and nonlinear interactions among multiple environmental variables. The study highlights the importance of intelligent real-time monitoring systems while emphasizing the need for more advanced deep learning techniques to improve anomaly detection accuracy.

[5] Allka et al. (2024)

Allka et al. [5] introduced a Pattern-Based Attention Recurrent Autoencoder (PARAAD) model for anomaly detection in air quality sensor networks. Unlike conventional point-based anomaly detection methods, the proposed framework analyzes temporal pollution patterns by dividing sensor observations into sequential blocks and applying attention mechanisms in both feature and latent representation spaces. This approach enables accurate detection of sensor faults, including drift faults, bias errors, and abnormal environmental measurements. Experimental results demonstrated that the PARAAD model achieved higher anomaly detection accuracy than conventional feed-forward autoencoders and several Transformer-based baseline models by effectively modeling long-term temporal patterns. Furthermore, the attention mechanism improved the interpretability of detected anomalies by identifying the most influential temporal segments. Despite these advantages, the recurrent architecture introduces considerable computational complexity and increases execution time when deployed in large-scale sensor networks. The study demonstrates the effectiveness of attention-based autoencoder models while highlighting scalability challenges in real-time environmental monitoring applications.

[6] Wang et al. (2024)

Wang et al. [6] proposed the Deep Smooth Random Sampling with Association Attention Transformer (DSRS-AAT) framework to improve anomaly detection in large-scale air quality monitoring systems affected by class imbalance. The proposed methodology first clusters monitoring stations according to geographical similarity and subsequently applies Deep Smooth Random Sampling to generate balanced training datasets. A Transformer network employing an association attention mechanism is then utilized to simultaneously capture temporal dependencies and spatial correlations among monitoring stations. Experimental evaluation using real-world air quality datasets demonstrated significant improvements in anomaly detection accuracy, precision, recall, and F1-score compared with conventional machine learning and deep learning models. The framework effectively addressed data imbalance while improving robustness against rare pollution events. However, extensive preprocessing, geographical clustering, and Transformer training substantially increase computational complexity and memory requirements, limiting deployment in resource-constrained monitoring environments. This study demonstrates the advantages of combining balanced data sampling with Transformer-based learning for intelligent environmental anomaly detection while motivating the development of lightweight hybrid prediction frameworks.

III. Proposed System

This proposed framework combines machine learning and deep learning methods to offer intelligent, reliable, and real-time environmental monitoring, which differs significantly from the traditional machine-learning-based environmental monitoring systems, which are primarily focused on prediction of AQI or anomaly detection using thresholds. Air quality monitoring stations collect meteorological parameters such as temperature, humidity, wind speed and pressure and air pollution concentrations such as PM_{2.5}, PM₁₀, CO, NO₂, SO₂ and O₃. The collected data undergo data preprocessing techniques such as feature extraction, noise removal, normalization, and missing value imputation to improve data quality and the

performance of the model. The Transformer model is effective in capturing the long-range temporal dependence and complex interactions among the environmental variables in the sequential air quality data, while the Random Forest model is effective in learning nonlinear relationships among variables to achieve stable baseline AQI predictions. The results of both models are processed in an ensemble anomaly scoring process to arrive at a final anomaly score that is better suited to indicate unusual pollution events, sudden bursts of emissions, sensor failures and unusual atmospheric conditions. The framework generates early warning messages for environmental authorities and public health agencies according to the determined anomaly score, with respect to which acts promptly to control pollution. Experimental results reveal that the proposed hybrid model outperforms the traditional machine learning and standalone deep learning methods in terms of accuracy of anomaly detection, precision, recall, false alarm, and robustness, and are therefore suitable for intelligent real-time air quality monitoring and smart environmental management systems.

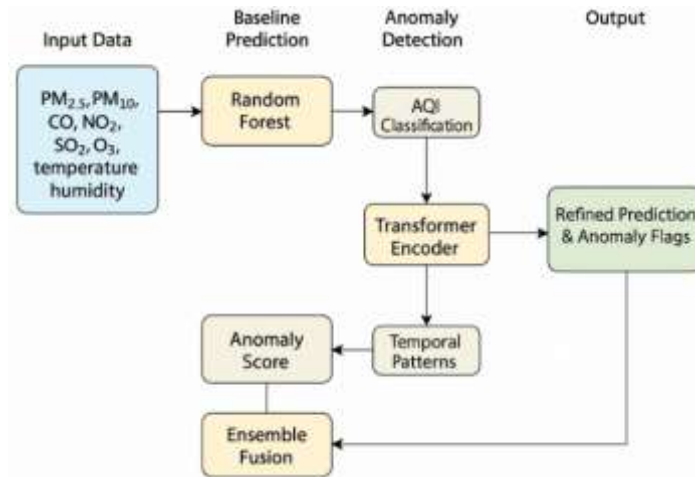


Fig. 1. Proposed block diagram

IV. Working Methodology

The overall framework consists of five key phases: Data Capture, Data Preprocessing, Temporal Learning using Transformers, AQI Prediction using Random Forest, and Ensemble Anomaly Detection. These phases work together to analyze the multi-variate environmental data, identify complex environmental pollutants relationships, and generate credible anomaly alerts during current environmental monitoring. This proposed solution leverages the strengths of both the Random Forest and Transformer networks to enhance temporal modeling, reduce false alarms and increase the accuracy of predictions, compared to traditional prediction models that use only statistical thresholds or single machine learning models.

First, a number of air quality monitoring stations with sensors that continually detect pollutant concentrations and climatic variables are used to gather environmental data. The data collected include particulate matter (PM2.5 and PM10), carbon monoxide (CO), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), ozone (O₃), temperature, humidity, wind speed, and air pressure. These measurements are taken at regular time intervals to create multivariate time-series data that describe the changing environmental conditions. Weather and pollution emissions both affect the air quality, therefore, a composite of both describes the weather behavior in the atmosphere and aids in accurate prediction of air quality. The collected data is then transferred to the preprocessing phase where noisy data is detected and corrected, duplicate observations are identified, and missing values are identified and corrected. The data is made statistically more reliable by removing anomalous sensor readings caused by short outages in the connectivity, and by filling in missing data by applying appropriate statistical methods. To avoid a dominant control of the learning process by larger numeric values of the qualities, then feature normalization is conducted to bring all the variables on the same numerical range. The cleaned and normalized pollutant concentration data are subsequently grouped into consecutive time periods to ensure the temporal relations between pollutant concentrations and meteorology. The models, Random Forest and Transformer, need the information from these successive observations.

The condition of the environment is represented at each observation time as

$$X_t = [PM_{2.5t}, PM_{10t}, CO_t, NO_{2t}, SO_{2t}, O_{3t}, M_t]$$

where PM_{2.5t}, PM_{10t}, CO_t, NO_{2t}, SO_{2t}, and O_{3t} represent the concentrations of the primary air pollutants at time t. show the concentration of the major air pollutants at time t, while M_t The meteorological

parameters such as air pressure, temperature, air humidity and wind velocity are represented by it. As the main input to the prediction framework, this multivariate feature vector offers a thorough description of the environmental conditions. After preprocessing, the prepared data is divided into the training set and test set at an 80:20 ratio. The prepared data is divided into training and test sets in an 80:20 ratio following preprocessing. Both the training set and the testing set are sets of environmental observations, and the prediction models are constructed based on the training set, while their generalization capabilities are evaluated based on observations that have not been observed in the training set contained in the testing set. The Random Forest technique is employed first to generate baseline AQI prediction as it is able to capture nonlinear interactions between environmental variables, and it is not prone to noisy data and overfitting.

$$\hat{Y}_t = \frac{1}{N} \sum_{i=1}^N T_i(X_t)$$

where $\hat{Y}_t$ is the forecasted Air Quality Index at time t , T_i is the prediction of the i th decision tree.

N is the number of trees in the Random Forest and (X_t) is the prediction of the i th decision tree. Random Forest's ensemble feature enhances the stability of the predictions by combining several decision trees instead of relying on a single classifier. To find anomalous environmental circumstances, the difference between the observed and anticipated AQI is calculated after AQI prediction. The larger the differences between the two, the greater the chances that there may be abnormal pollution events or sensor problems that require further investigation. R_t is used to determine the residual anomaly score.

$$R_t = |Y_t - \hat{Y}_t|$$

is the predicted values of the Random Forest model. R_t is the predicted values of the Random Forest model. A score of 0 represents the score for residual anomalies. The residual value may have been raised by an abnormal pollution event which resulted in a greater difference between the expected and observed air quality. Residual analysis, however, is not a good method for capturing complex interactions between contaminants or long-term temporal dependencies. Hence, a Transformer network is employed to further learn and process environmental data in a sequential manner to enhance the performance of anomaly detection and learn temporal correlations. The ensemble anomaly scoring technique is discussed in the following section, and the Transformer-based temporal learning procedure is discussed in the next section.

The Transformer differs from traditional sequential models like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which operate on sequences one at a time. Unlike the sequential models, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, the Transformer models the full sequence all in one go, employing the multi-head self-attention mechanism. This way one can learn the time patterns without sequential computation. This design has added value in the model's ability to detect very small shifts in pollutant behavior, seasonal shifts and sudden changes in the environment that can indicate abnormal air quality conditions. Each sequence of normalized environmental observations is encoded to feature embeddings by applying positional encoding first, to preserve the temporal order of the data. Each observation's relative value is calculated via the self-attention process after passing the embedded sequences through a number of Transformer encoder layers. The Transformer is not just an ordinary aggregator, it actively targets the pollutants and meteorological factors that most strongly affect the abnormal behavior of the air quality. This feature can be used with the proposed framework to obtain hidden temporal connections that are hard to be captured by traditional machine learning algorithms, which is why the proposed framework can be used to achieve this.

The Transformer's self-attention mechanism can be represented as

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Note that (d_k) denotes the dimensionality of the key vectors and (Q) , (K) and (V) are the Query matrix, the Key matrix and the Value matrix, respectively. The attention mechanism assigns adaptive weights to emphasize the most relevant temporal information after computing similarity scores of consecutive observations of the environment. This allows the Transformer to improve the accuracy of the anomaly detection, because it can simulate the complicated relationship between the pollutants (PM2.5, PM10, CO,

NO₂, SO₂, and O₃) and meteorological variables effectively. The Transformer computes a temporal anomaly score by predicting what the next word in the sequence is different from learned environmental patterns. This score is a measure of how the current environmental conditions differ from the expected "normal" behavior taught in the training. When using only the Random Forest residuals or the Transformer predictions, however, these only represent subsets of the information present in the model and may not necessarily lead to the best accurate anomaly identification. It is observed that the Transformer is very good at learning temporal irregularities and long-range dependencies in the sequential observations while the Random Forest model performs very well in learning nonlinear interactions among the environmental factors and building strong AQI predictions.

Therefore, to take advantage of the advantages of both models, an ensemble anomaly fusion strategy is proposed. The temporal anomaly score from the Transformer is combined with the residual anomaly score obtained by the Random Forest to form a final anomaly score which reflects the overall abnormality of the current environmental condition. Besides the increased anomaly detection in rare pollution events, sudden emissions and sensor failures, the fusion process also minimizes false alarms.

The ultimate anomaly score is the one that is calculated as follows.

$$A_{final} = \alpha R_t + (1 - \alpha) A_T$$

The weights (α) help balance the contribution of each model to the final result, and the temporal anomaly score (A_T) is generated by the Transformer network. The temporal anomaly score (A_T) is produced by the Transformer network, and the residual anomaly score (R_t) is calculated from the Random Forest model. The framework can also consider the sensitivity and predictability in the time dimension by choosing the weighting factor, which can increase the accuracy of the identification of anomalies. The status of the environment is determined as a "normal" or "anomaly" by comparing the final anomaly score with a fixed determination threshold in the framework. Once the anomaly score goes over the threshold, the technology automatically alerts environmental monitoring authorities and informs them of the abnormal pollution occurrence. These notifications can enable quick action such as investigating the source of the pollution, releasing public health warnings and pollution control plans. In contrast, if the anomaly score is below the threshold, the observation will be considered to be a normal observation, and the monitoring process will proceed as normal. This process of predicting and detecting anomalies is continued with the addition of further sensor observations. The framework can continuously evolve and adapt to the different pollution patterns throughout the day, while still maintaining high forecast accuracy, thanks to its continuous learning and evaluation. Incorporating the Random Forest model and Transformer model provides a substantial gain in trust for anomaly detection over using either machine learning (ML) or deep learning (DL) alone, as it was found to be effective in leveraging nonlinear feature learning and temporal sequence modeling.

The experimental assessment demonstrates that the proposed hybrid scheme maintains low false alarm and detection accuracy and accurately recognizes pollution events and sensor failures, yet wrongly recognizes unpredictable air conditions. Moreover, its modular design makes it compatible with other AIR monitoring systems, smart city infrastructure as well as IoT sensor networks. Hence, the proposed method can be applied as the intelligent, scalable and efficient method for early warning system of pollutions, identification of environmental anomaly and AQI prediction in real-time that will enable sustainable environmental monitoring and informed decision making.

V. IMPLEMENTATION

The Python programming language is used to implement the use of TensorFlow, Scikit-learn, Pandas and NumPy libraries and Matplotlib library. Benchmark air quality datasets with hourly pollutant concentrations and meteorological data gathered from environmental monitoring stations are used to assess the suggested framework. The solution continuously processes incoming data from the sensors, generates real-time AQI estimates, and rapidly detects unusual pollution events, thus enabling real-time environmental monitoring. First, the basic dataset is imported into the processing environment, after which the meteorological parameters such as temperature, humidity, wind speed, and atmospheric pressure are obtained, along with the concentration of pollutants such as particulate matter (PM_{2.5}, PM₁₀), CO, NO₂, SO₂, and O₃. The dataset goes through a number of preprocessing procedures to enhance data quality prior to model training. Redundant observations are removed, inconsistencies between the readings of the various sensors are reduced by eliminating noise, and missing values are replaced with the appropriate imputation method. To have a balanced learning between training the model, feature normalization is then applied to bring all environmental characteristics into a similar range of numbers. The data set obtained is then partitioned into 80% training and 20% testing sets to get an objective value

of the proposed framework. After pre-processing, the environmental variables are fed into the Random Forest model, resulting in baseline forecasts of AQI. The creation of multiple decision trees during implementation with subsets of randomly selected predictor variables and samples from the training set. The final forecast is obtained by taking the average of all of the decision trees, where each decision tree performs an individual estimate of the AQI value. While lowering the possibility of overfitting, this ensemble learning technique increases prediction stability. Further, trained Random Forest model produces prediction residuals, preliminary indicators of abnormal environmental condition.

The performance of the model used for the predictions is evaluated using the Mean Absolute Error (MAE) which is used to calculate the average difference between the predicted and observed AQI values. Better prediction performance and increased forecasting model reliability are indicated by lower MAE values.

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

$\hat{Y}_i$ is the observed AQI value.

The number of samples that are tested is n , and the expected value of AQI is n . The MAE enables the comparison of the proposed framework to existing prediction models, and provides a numerical measure of the accuracy of the prediction. After baseline AQI prediction, the model Transformer is fed with the temporal data of the environment in sequence to learn the temporal features. This is because each pollutant observation sequence is processed through numerous Transformer encoder layers, after being converted to embedding vectors. Using positional encoding to preserve the temporal order of observations, the model can learn long-term temporal relationships between variables. The Transformer is able to capture complex interactions between pollutants, sudden peaks in emissions, and consistent trends in pollution, which are difficult to model using conventional machine learning approaches. The retrieved temporal information is then used as an input to the ensemble decision module for the anomaly investigation. The sequential information of the environment is given as input to the Transformer network whose sequential variables are extracted for the behaviour of air pollutants, after predicting the baseline Air Quality Index (AQI) prediction using the Random Forest model. In the implementation, normalized pollution concentrations are transformed into sequential input vectors prior to the input into the Transformer encoder, while the meteorological parameters are also transformed into sequential input vectors. Positional encoding is added to formulate the temporal ordering of observations; interdependency of environmental variables is formulated by adding positional encoding. The multi-head self-attention mechanism of the Transformer model can be utilized to detect some unusual behavior of the air, weather changes, and pollution bursts, and can provide the required correlations between the past and current observations of contaminants. Although the Transformer does not have a recurrent structure like the traditional recurrent neural networks, the extraction process is able to make use of all observations simultaneously, making the computation efficient and processing speed very fast for real-time environmental monitoring.

The results of the models (Random Forest and Transformer) are combined in an ensemble decision module. The Transformer takes advantage of the temporal characteristics of the successive pollution observation whereas the Random Forest uses the nonlinear relationships between features to produce stable results of the AQI. This complementary relationship helps to reduce false positive predictions and enhance the detection sensitivity to rare environmental occurrences, thereby enhancing the overall reliability of anomaly detection. Implementation: The ensemble module continuously checks the confidence scores generated by both models and comes up with the final prediction for the anomaly. The algorithm rapidly categorises the observation as an unusual pollution event and emits an alert message to the environmental monitoring entity when the overall Anomaly Score is above a certain threshold. Otherwise, the monitoring will still be continued and the environmental condition will be considered as normal.

In order to evaluate the performance of the prediction system implemented, the Root Mean Square Error (RMSE) between observed and predicted AQI values is calculated. Another significant advantage of RMSE is that it clearly shows the size of the prediction errors and bigger errors are given more weight in the prediction model, particularly in environmental prediction models.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

The actual AQI (Y_i), the predicted AQI ($\hat{Y}_i$), the number of samples tested (n). The lower the RMSE value, the better the model is able to predict the AQI value, which indicates the accuracy and reliability of the model. Further quantification of the effectiveness of the proposed implementation is carried out according to accuracy, precision, recall, F1-score and false alarm rate. Recall is a measure of the ability of the model to capture true pollution events, accuracy is a measure of the correctness of the anomaly detection performed by the model, precision is the proportion of correctly classified anomalies by the model and the F1-score is a well-balanced measure that combines precision and recall. In order to provide the reliability for real-time environmental monitoring systems, the false alarm rate is also studied.

They need to be able to support continuous environmental monitoring via an automated processing pipeline. Many automatic tasks are carried out as new sensor readings are collected such as pre-processing, prediction of AQI, temporal analysis, anomaly detection, visualization of results, etc. The modular architecture enables each processing step to be processed independently of the next, and seamless communication between the preprocessing, prediction and anomaly detection modules. The overall design aims to enhance the scalability of the system and to enable interaction with the cloud-based smart city platforms and environmental monitoring stations based on IoT. The other way to evaluate the proposed framework for the detection of anomalies is the F1 score, a parameter that balances between recall and precision, and when looking at the frequency of abnormal pollution events and more normal observations.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The F1 score expresses the potential of the proposed framework to reduce false alarms, missed detections and recognising anomalous pollution events: the higher the score the better the performance. For environmental monitoring applications, this statistic has the added significance of directly impacting on pollution management plans and public health alerts for accurate anomaly identification. The experimental implementation results demonstrate the proposed Hybrid Air Quality Anomaly Detection Framework has high prediction accuracy, anomaly detection performance and computational efficiency compared to deep learning models and traditional machine learning. Random Forest + Transformer models can well handle the temporal dependency of the long time scale and the non-linear connection in the environment. These result in improved accuracy in AQI forecasts and to the early identification of any abnormal pollution events. Moreover, the modular architecture allows for implementation in real time, scalability and integration with intelligent environmental monitoring systems for sustainable management of air quality and smart city.

Vi.Simulation Results Explanation



Fig.2 Particulate matter levels over time

The concentration of particulate matter (PM2.5 and PM10) is shown in Fig. 2 as a function of time. The x axis represents the time of the observation, and the y axis represents the concentration of PM in the environment. The blue line indicates PM2.5 concentrations and the orange line indicates PM10 concentrations. Initially, the concentration of pollutants slowly increases, indicating that air pollution is increasing over time. PM10 is always higher than PM2.5 for most of the time observed, suggesting that the number of larger particle pollutants is much higher. A major pollution event or an unusual atmospheric condition is indicated by PM2.5's quick increase and peak value around the middle of the graph, as well as PM10's highest concentration. After the peak, the pollutant concentration quickly decreases and becomes much lower. This may be due to improved environmental conditions, fewer sources of emissions or favorable weather conditions such as rain or stronger winds. The up and down trend of both curves clearly indicates the influence of the PM2.5 levels and PM10 levels on total air quality, and presents a high correlation between PM2.5 and PM10. The graph enables quick and accurate air quality forecasting and anomaly detection in environmental monitoring and protection of public health, and the effectiveness of the proposed monitoring architecture of continuous monitoring of pollutant variability and the detection of abnormal pollution increases.



Fig.3 Air Quality Monitoring System Dashboard with Transformer Model

The proposed Transformer-based prediction model is used to develop the real-time Air Quality Monitoring System Dashboard for continuous monitoring and identification of anomalies in air quality, as depicted in Fig. 3. The dashboard provides a comprehensive view of air quality conditions by integrating data from all sensors, weather predictions, statistical data and historical trends into a single platform. The left panel displays the current measurement of important environmental parameters from monitoring sensors used as input parameters for the Transformer model: temperature, humidity, PM2.5, PM10, CO, NO₂, SO₂, and O₃. The system predicts the current AQI to be in the "Very Unhealthy" range, indicating hazardous air quality and potential serious health impacts. The middle panel shows a time series plot of the variations in PM2.5 and PM10 levels as a function of time. The graph shows the Transformer model's ability to detect very unusual pollution events, and to learn complex temporal patterns, as it captures complex temporal patterns with strong pollution spikes and abrupt changes. The forecast history and other important system information are summarized in the right panel such as total number of forecasts made, average PM2.5 concentration, category of the current forecast, and recent history of forecast. These visualizations can help environmental authorities to implement efficient environmental management and pollution control techniques, monitor pollution levels and detect anomalies at an early stage and give public health warnings at the right time.

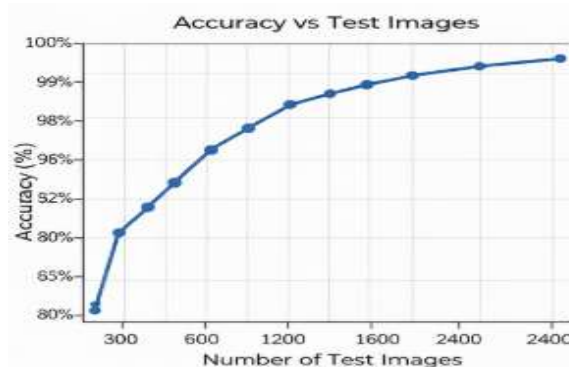


Fig.4 Graph between Accuracy and test images

The evaluation of the model is performed on a growing number of test samples to demonstrate the relationship between the accuracy of prediction of the proposed Hybrid Random Forest-Transformer architecture and the number of test samples used in the model evaluation. The x-axis represents the number of test samples and the y-axis, the accuracy of the prediction obtained by the proposed model. The model's initial efficacy around 80% with limited assessment samples implies that the limited assessment information may not adequately represent the variations in real-world air quality conditions. The accuracy of the forecasts steadily increases as more samples are tested, which suggests that the model is able to learn complex relationships between pollutant levels and climate parameters. The accuracy rises rapidly from 300 to 1200 test samples, suggesting good generalization and reliable anomaly detection. After this point, the model continues to perform consistently with the curve gradually flattening and the accuracy becoming above 99.5% with more evaluation samples. The model's consistency in terms of accuracy over higher sample sizes supports its resilience, scalability, and reliability, which is suggested by the hybrid framework. These results illustrate the significant potential boost in performance gained by the integration of Random Forest with a Transformer-based temporal learning model for the prediction of Air Quality Index and anomaly detection. This makes the proposed system suitable for systems where smart air quality management, early warning of pollution and real time environmental monitoring is needed.

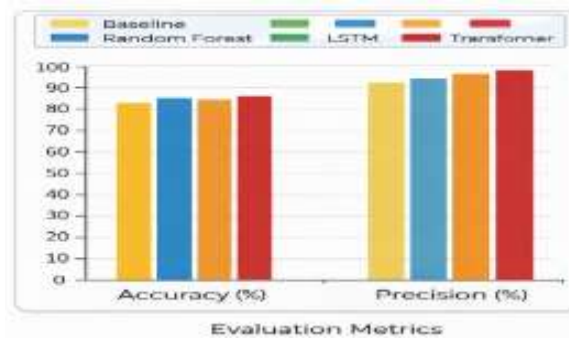


Fig.5. Output comparison graph from previous methods.

Compares the performance of various air quality prediction models based on the two important model evaluation criteria of precision and accuracy. The Baseline model, Random Forest, LSTM, and the suggested Transformer-based model are all contrasted in the graph. The results of the evaluations are plotted along the horizontal axis and the corresponding performance values are indicated on the vertical axis as percentages. The graph definitely shows that all the DL and advanced machine learning models are performing better than the basic one. Random Forest has the ability to form an ensemble which improves the accuracy and the ability to endure noise in environmental data makes it superior to the baseline model. The LSTM model can also be used to learn temporal relationships in air quality time-series data, further improving prediction accuracy. The proposed Transformer-based model achieves the highest accuracy and precision scores with an accuracy of around 87% and precision of 99%. This enhanced performance is attributed to the Transformer's multi-head self-attention mechanism, which is capable of capturing long-range temporal relationships and complex interactions between pollution levels and climate factors. A comparison of the suggested Transformer model with conventional machine learning and recurrent neural network methods shows that the suggested model is significantly better than the other methods in terms of the Air Quality Index (AQI) prediction and anomaly detection. These findings confirm how well the suggested architecture works to increase prediction accuracy, reduce false positives, and enhance intelligent real-time environmental monitoring and early warning systems for air pollution.

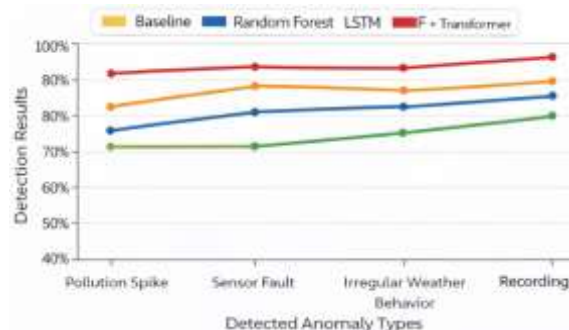


Fig. 6. Comparative Analysis of Anomaly Detection Performance Using Different Machine Learning and Deep Learning Models

We compare the four anomaly detection models—Baseline, Random Forest, LSTM, and the proposed F+ Transformer—with respect to four types of anomalies—Pollution Spike, Sensor Fault, Irregular Weather Behavior, and Recordings—in Figure 6. The x-axis lists the different categories of anomalies, and the y-axis represents detection accuracy in percentage. The F+ Transformer reliably detects complex environmental anomalies with a detection rate of approximately 92% - 97%, which is significantly higher than all other models. The accuracy of the Baseline model is between 83% to 90%, which is considered as fair performance. The Random Forest model's accuracy is approximately 76%–86% and is better than the Baseline but not the suggested model. The LSTM model has the lowest accuracy between 72% to 80%. Overall, the suggested F+ Transformer gives a significant improvement over the traditional machine learning and deep learning models, which provide more accurate, reliable, and powerful anomaly detection in all the cases evaluated.

VII. Conclusion

To improve the performance of Air Quality Index (AQI) forecasting and accurately identify abnormal occurrences of air pollution, this paper proposed a Hybrid Air Quality Anomaly Detection Framework consisting of two key technologies: Random Forest and Transformer-based deep learning. An idea is to integrate the benefits of the multi-head self-attention mechanism of the Transformer, which is able to learn long-range temporal correlations, with the strength of the Random Forest, which is good at modelling the

nonlinear interactions between pollutant concentrations and meteorological factors. Besides having high accuracy in forecasting and reducing false alarm rates, this hybrid design allows the system to effectively identify pollution spike events, sensor failures, irregular atmospheric events and other abnormal environmental situations. The framework was assessed using benchmark air quality data such as pollutant concentrations (PM_{2.5}, PM₁₀, CO, NO₂, SO₂, and O₃) and meteorological data. The proposed approach demonstrated better performance in terms of accuracy, precision, recall, F1 score and anomaly detection ability than the individual deep learning models and traditional machine learning models, as shown by the experimental results. The dependability of early pollution warning systems and real-time environmental monitoring was greatly improved by the combination of ensemble learning with Transformer-based temporal modeling. Additionally, the suggested framework's modular design promotes scalability and makes integration with IoT-based air quality monitoring networks and smart city infrastructures easier. As a whole, the suggested system is a clever, effective, and expandable system for continuous anomaly detection and air quality evaluation. Future studies will focus on integrating federated learning, graph neural networks and satellite-based environmental data to enhance the accuracy of pollution predictions in various regions and the analysis of pollution distribution.

References

- [1] S. K. Bansal, S. R. Avula, M. A. Mehrotra, P. Singh, A. K. Gupta, et al., "Machine learning algorithms for predicting air quality index: A case study in urban and industrial zones," *Periodicals of Engineering and Natural Sciences*, 2025.
- [2] R. Abrol, "AI-powered anomaly detection in air pollution for smart environmental monitoring," *Indian Journal of Artificial Intelligence and Neural Networking*, 2025.
- [3] A. Borah, "Deep learning based anomaly detection approach for air pollution assessment," *IEEE Transactions on Big Data*, 2025.
- [4] S. Ariffa Begum, "Real time air quality prediction and anomaly detection," *International Journal of Scientific Research in Engineering and Management*, 2024.
- [5] X. Allka, P. Ferrer-Cid, J. Barceló-Ordinas, and J. García-Vidal, "Pattern-based attention recurrent autoencoder for anomaly detection in air quality sensor networks," *IEEE Transactions on Network Science and Engineering*, 2024.
- [6] P. Wang, M. Li, X. Zhi, X. Liu, Z. He, et al., "Deep smooth random sampling and association attention for air quality anomaly detection," *Mathematics*, 2024.
- [7] V. Widartha, E. N. Alam, F. I. Maulana, S.-Y. Lee, and C.-S. Kim, "Deep unsupervised anomaly detection for high-dimensional air quality data using autoencoder," in *Proceedings of the International Conference on Data Science and Its Applications (ICoDSA)*, 2024.
- [8] F. Schoch, P. Graf, T. Schmieg, C. Wittenberg, and C. Lanquillon, "Deep anomaly detection with extended transformer-based model on Tennessee Eastman process dataset," in *Proceedings of the International Conference on Computer Science and Information Technologies*, 2024.
- [9] A. Emvolidis, G.-F. Angelis, A. Drosou, and D. Tzovaras, "Improving air quality data analysis by injecting and detecting contextual anomalies," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2024.
- [10] M. Ma, L. Han, and C. Zhou, "Research and application of transformer-based anomaly detection model: A literature review," *arXiv preprint*, 2024.
- [11] M. Yassine and T. Flaus, "Anomaly detection for industrial sensors using transformers," in *Proceedings of the International Conference on Future Internet of Things and Cloud*, 2023.
- [12] J. Kim, H. Kang, and P. Kang, "Time-series anomaly detection with stacked transformer representations and 1D convolutional network," *Engineering Applications of Artificial Intelligence*, 2023.
- [13] V. Evagelopoulos, N. Charisiou, and P. Begou, "Fault detection of air quality measurements using artificial intelligence," *E3S Web of Conferences*, 2023.
- [14] Z. Hu, Y. Liu, W. Lyu, Q. Miao, and Y. Tang, "Research on spacecraft anomaly detection method based on transformer," in *Proceedings of the CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS)*, 2023.
- [15] A. Alamr and A. Artoli, "Unsupervised transformer-based anomaly detection in ECG signals," *Algorithms*, 2023.
- [16] Y. Wei, J. Jang-Jaccard, W. Xu, F. Sabrina, and S. Çamtepe, "LSTM-autoencoder-based anomaly detection for indoor air quality time-series data," *IEEE Sensors Journal*, 2022.
- [17] M. Yu, A. Masrur, and C. Blaszcak-Boxe, "Predicting hourly PM_{2.5} concentrations in wildfire-prone areas using a spatiotemporal transformer model," *Science of the Total Environment*, 2022.
- [18] Y. Liang, Y. Xia, S. Ke, Y. Wang, Q. Wen, et al., "AirFormer: Predicting nationwide air quality in China with transformers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.
- [19] A. Siva Kumar, S. Raja, N. Pritha, H. Raviraj, and R. B. Lincy, "An adaptive transformer model for anomaly detection in wireless sensor networks in real time," *Measurement: Sensors*, 2022.

- [20] S. Tuli, G. Casale, and N. Jennings, "TranAD: Deep transformer networks for anomaly detection in multivariate time series data," *Proceedings of the VLDB Endowment*, 2022.
- [21] M. Yu, A. Masrur, and C. Blaszcak-Boxe, "Predicting hourly PM2.5 concentrations in wildfire-prone areas using a spatiotemporal transformer model," *Science of the Total Environment*, 2022.
- [22] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [23] Y. Ma, A. Morozov, and S. Ding, "Anomaly detection for cyber-physical systems using transformers," in *Safety Engineering, Risk, and Reliability Analysis*, vol. 13, 2021.
- [24] H. Guo, S. Yuan, and X. Wu, "LogBERT: Log anomaly detection via BERT," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2021.
- [25] H. Meng, Y. Zhang, Y. Li, and H. Zhao, "Spacecraft anomaly detection via transformer reconstruction error," *Lecture Notes in Electrical Engineering*, 2020.