Fault detection of air quality measurements using artificial intelligence

Vasilis Evagelopoulos[1,*], *Nikolaos* D. Charisiou[1], and *Paraskevi* Begou[2]

[1]Department of Chemical Engineering, University of Western Macedonia, 50100, Kozani, Greece
[2]Laboratory of Meteorology, Department of Physics, University of Ioannina, Ioannina, 45110, Greece

Abstract. In this work we use Artificial Intelligence (AI) for the detection of faults in air quality measurements. This is crucial in large air quality monitoring networks in particular were fault detection can be a complex and time consuming process. The proposed methodology encompasses several essential steps in anomaly detection. Data preprocessing ensures the quality and relevance of the data by applying techniques like data cleaning, outlier removal, and feature selection. The Isolation Forest model is trained using the pre-processed data, and appropriate hyperparameters are determined through cross-validation. Anomaly detection is performed using the trained model, allowing the identification of abnormal events or instances. The visualization of anomalies provides a clear representation of abnormal patterns, facilitating the interpretation and understanding of air quality data. The proposed methodology can help environmental agencies, researchers, and policymakers in identifying abnormal air quality events, enhancing the accuracy of monitoring systems, and facilitating timely interventions. This methodology can be applied to other industries also, to improve operations and reduce risk.

1 Introduction

Air quality is a critical aspect of public health and given the continuous increase in industrial activities, urbanization, and transportation, its monitoring and analysis is of paramount importance [1,2]. The adverse impacts of air pollution on respiratory diseases, cardiovascular disorders, and even cognitive function have led to a heightened awareness of the urgent need for effective air quality monitoring and control strategies [3].

Monitoring networks can consist of a large number of air quality measuring stations, which require continuous monitoring as due to equipment malfunction the coverage rate of measurements can be dramatically reduced due to delay in detecting abnormal values and possible faults [4]. Traditional methods of air quality monitoring involve deploying physical sensors at specific locations to measure pollutant concentrations. However, these methods often lack the ability to identify anomalies or abnormal patterns in real-time data [5].

To address the above limitations, the integration of artificial intelligence (AI) techniques has emerged as a promising approach for detecting anomalies in air quality data [6-11].

An anomaly is a pattern in data that significantly deviates from the normal behaviour. Anomalies can occur due to various factors such as equipment malfunction, sensor failures, extreme weather conditions, or the occurrence of exceptional pollution events [12]. By identifying and addressing these anomalies promptly, it is possible to enhance the accuracy of air quality monitoring, improve public health response systems, and facilitate effective decision-making for environmental management.

There are three main ways to create an anomaly detection model: unsupervised, supervised and semi-supervised anomaly detection. In an unsupervised model, differences in one set of points are examined to detect points moving away from it. This method can detect anomalies in unlabelled datasets, significantly reducing the manual labelling of huge amounts of data for model training [13].

In recent years, AI-based anomaly detection methods have gained traction for their ability to automatically learn patterns and detect anomalies in complex and high-dimensional datasets. AI methods leverage machine learning algorithms and statistical techniques to identify deviations from normal air quality patterns [14]. An algorithm that is promising in anomaly detection is the Isolation Forest algorithm [15].

The Isolation Forest algorithm is a machine learning technique based on the concept of isolation. It builds an ensemble of isolation trees that recursively partition the data until each instance is isolated or grouped together. The algorithm identifies anomalies by measuring the number of partitions required to isolate an instance, considering that anomalies are expected to be less frequent and thus require fewer partitions. Isolation Forest has demonstrated good performance in detecting anomalies in various domains, including air quality data analysis [9].

* Corresponding author: vevagelopoulos@uowm.gr

However, it is important to note that the effectiveness of the anomaly detection methodology is highly dependent on the quality and representativeness of the data used. Accurate and comprehensive air quality data, including pollutant concentrations, meteorological conditions, and temporal information, is crucial for obtaining meaningful results. Furthermore, the selection and fine-tuning of hyperparameters, as well as the choice of appropriate features, can significantly impact the performance of the Isolation Forest algorithm in anomaly detection.

In conclusion, the integration of AI techniques, specifically the Isolation Forest algorithm, offers a promising approach for detecting anomalies in air quality data. The goal of this research is to develop a methodology for detecting anomalies in air quality data using the Isolation Forest algorithm. The application of this methodology is a promising way to enhance air quality monitoring and improve public health response systems, and thus, enable effective decision-making for environmental management.

2 Methodology

Detecting data anomalies has been a very active research domain and it has many applications in industry, healthcare, finance and security [5]. Detecting air quality data anomalies using AI involves training a machine learning model to classify or predict air quality conditions based on historical data. Data preparation is a crucial step in detecting anomalies in air quality data using AI techniques. It involves ensuring the quality, relevance, and readiness of the data for further analysis. The method developed is shown in Fig. 1.

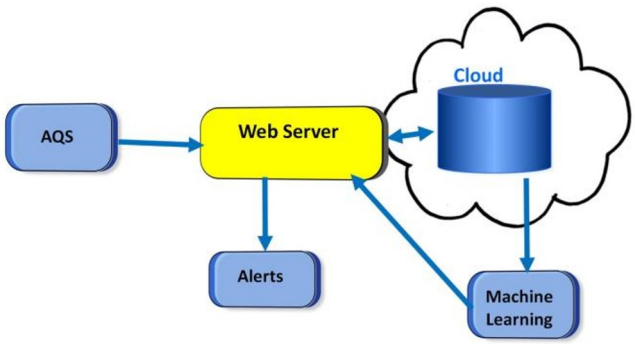

Fig. 1. Diagram of system developed.

For the work presented herein, we collected air quality data from reliable sources such as Environmental Monitoring Stations (EMS) and data stored in SQL databases in cloud. Next, using a web server capable of running python scripts, we run an application, hourly, developed to detect outliers in real time and warn the EMS personnel for possible equipment malfunction.
In order to avoid cases where we simply have abnormal values that are real and not due to failures of the analysers, we placed a check for each analyser separately. The model developed calculates the abnormal values separately for each parameter (e.g. PM_{10}, NO_x, SO_2, CO). Thus, for each parameter, a data set is created that includes the parameter and meteorological data, as it is known that abnormal values are associated with meteorological parameters. For each measured parameter, optimization was performed in order to find the best combination of hyperparameters that maximizes the model's ability to detect anomalies while minimizing false positives or false negatives.

After gathering hourly historical data, the amount of which can be a function of the computing power and speed of the web server, we analysed and pre-processed the collected data. This procedure involved handling missing values, removing outliers, normalizing or standardizing the data, and encoding categorical variables. The performance of these data preparation steps allowed us to ensure that the data used for anomaly detection is of high quality, relevant, and appropriately transformed. Thus, a strong foundation for the subsequent stages was ensured, including model training, anomaly detection, and visualization. Flowchart for the methodology is shown in Fig. 2.

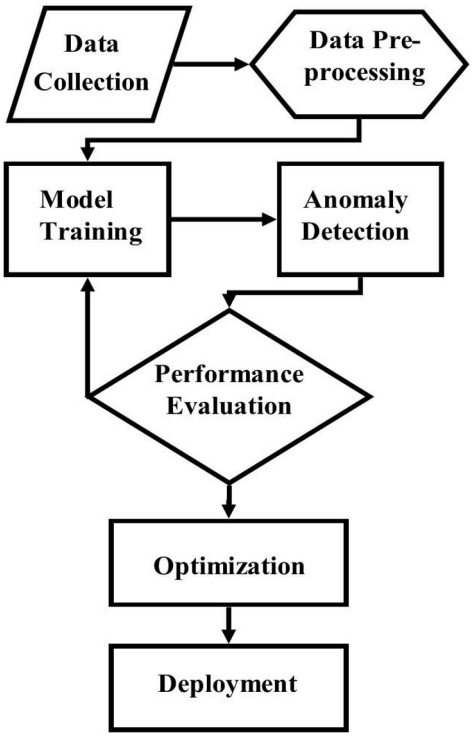

Fig 2. Flowchart for the methodology of detecting anomalies in air quality data using Isolation Forest

Afterwards, we trained the Isolation Forest model using the training data. The model learns the patterns and characteristics of normal data during this stage. The training process involves feeding the input variables to the model and allowing it to learn the underlying structure of the data. The algorithm constructs an ensemble of isolation trees based on random splits and isolates instances that require fewer partitions, indicating potential anomalies.

Additionally, in order to evaluate the performance of the model on unseen data we split the prepared data into training and testing sets. The training set was used to train the anomaly detection model, while the testing set

was used for evaluation. Finally, in case of detecting a possible failure, an email was automatically sent to predefined recipients in order to check for possible failures.

The proposed system helps to improve the accuracy and reliability of the anomaly detection system and facilitates a more effective decision-making in air quality monitoring and management.

2.1 Performance analysis

We used Python version 3.7 as the programming language with the following libraries. Sklearn: library for machine learning and Isolation Forest implementation; Pandas: data analysis library and manipulation tool; Numpy: library for scientific computing; Matplotlib: a visualization library. The raw data used for analysis were made available from Environmental Centre in region of west Macedonia, Greece [16].

Table 1. Statistics of model training data

	PM_{10}	TEMP	RH	WS	WD
count	2419	2419	2419	2419	2419
mean	18	16	67	1	181
std	6.53	3.79	16.77	1.41	95.64
min	3	8.9	23	0.1	1
25%	14	12.8	55	0.4	104.5
50%	18	15.6	65	0.9	188
75%	21	19.2	81	2.2	250
max	67	25.6	95	7	360

Table 1 provides an example of dataset analysis, which includes inhalable particulate matter (PM_{10}) with meteorological parameters such as air temperature (TEMP), relative humidity (RH), wind speed (WS) and wind direction (WD).

2.1.1 Data Preprocessing

The first step towards preprocessing includes replacing the Null values with the sliding mean values. We identified and handled outliers using statistical methods and domain knowledge to ensure the data was in a suitable format for further analysis. Then we selected the most relevant features for anomaly detection considering factors such as the impact of each feature on air quality. Then we used correlation analysis, feature importance and domain expertise to identify the most informative features. Table 2 contains Pearson correlation coefficient values between features. In this data set, the correlation of PM_{10} with meteorological parameters is small, which can however be improved if the data correlation takes place over a longer period.

Table 2. Corelation Coefficients

	PM_{10}	WS	WD	TEMP	RH
PM_{10}	1.00	-0.02	0.17	0.13	0.07
WS	-0.02	1.00	-0.05	0.45	-0.44
WD	0.17	-0.05	1.00	-0.04	0.13
TEMP	0.13	0.45	-0.04	1.00	-0.91
RH	0.07	-0.44	0.13	-0.91	1.00

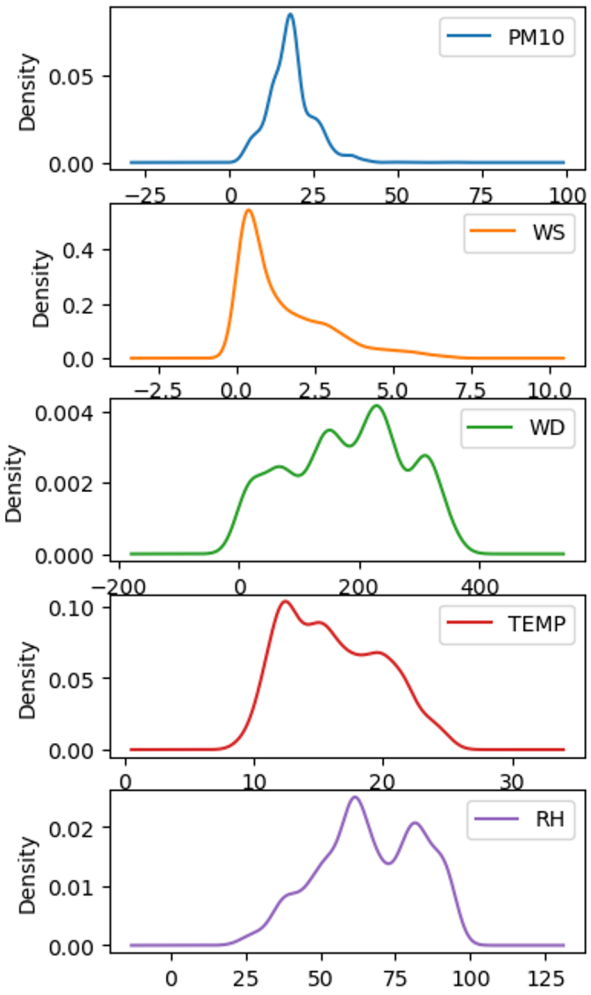

Fig. 3. KDE plot of variables used for model

Subsequently, we calculated the Kernel Density Estimation (KDE) which is a non-parametric technique used to estimate the probability density function (PDF) of a dataset. KDE provides a smooth estimate of the underlying distribution of the data and can be applied to describe the data in air quality analysis by visualizing the density of pollutant concentrations or other relevant features. Figure 3 shows the estimated density of the selected feature providing insights into its distribution and concentration levels. The shaded area represents the estimated density curve. The KDE plot allows us to visualize the shape, peaks, and variability of the data, which can help understand the underlying distribution and identify potential anomalies or unusual patterns. The y-axis in a density plot is the probability density function

for the kernel density estimation The x-axis is the value of the variable.

2.1.2 Outlier Identification¶

In this part we checked for possible PM_{10} variable outliers. One method of finding outliers is using the boxplot. Boxplots are effective tools for outlier analysis, providing a concise summary of the data's spread and enabling quick identification of potential outliers. By analysing the boxplot, we were able to identify potential outliers in the dataset.

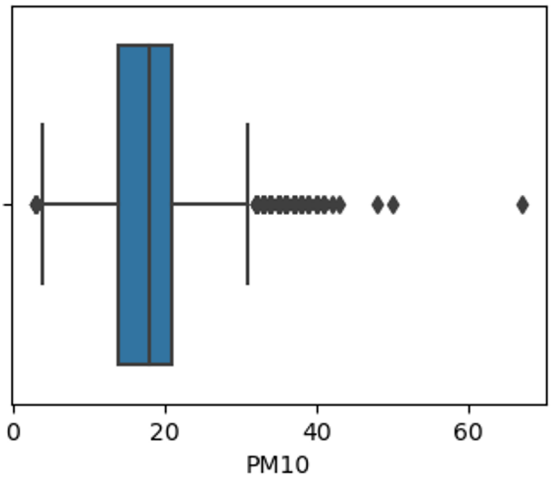

Fig. 4. Boxplot of PM10 data

To check whether the assumption read from the boxplot is true, we performed another analysis of outliers using the Isolation Forest method. Isolation Forest detects anomalies solely on the basis that anomalies are few and different data points and is performed without using distance or density measurements. The predicting method returns the outliers as 1 for the norm and -1 for the anomaly. The plot of the result is shown in Figure 5. The red points are detected anomalies with the model.

The above analysis confirmed the information previously obtained from the boxplot.

Once satisfied with the model's performance, we deployed it in the web server application. This involves integrating it into developed system that can process real-time air quality data and generate anomaly predictions.

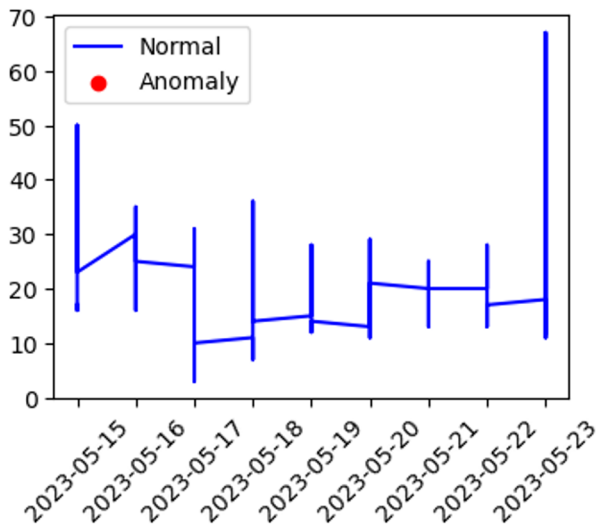

Fig. 5. Plot with anomalies detected with Isolation Forest method.

3 Conclusions

In conclusion, the detection of anomalies in air quality data using AI techniques, specifically the Isolation Forest algorithm, holds significant potential for improving air quality monitoring and management. This work has presented a methodology for detecting anomalies in air quality data, including data collection, preprocessing, model training and anomaly detection. By following this methodology, valuable insights can be gained to identify abnormal air quality events, prioritize interventions, and enhance decision-making processes.
The integration of AI techniques, such as the Isolation Forest algorithm, addresses the limitations of traditional methods by automatically learning patterns and detecting anomalies in complex and high-dimensional air quality datasets. The Isolation Forest algorithm leverages the concept of isolation to identify anomalies, providing efficient and immediate results. Its ability to isolate instances by partitioning the data makes it suitable for detecting anomalies in various domains, including air quality data analysis.

The proposed methodology encompasses several essential steps in anomaly detection. Data preprocessing ensures the quality and relevance of the data by applying techniques like data cleaning, outlier removal, and feature selection. The Isolation Forest model is trained using the preprocessed data, and appropriate hyperparameters are determined through cross-validation. Anomaly detection is performed using the trained model, allowing the identification of abnormal events or instances. The visualization of anomalies provides a clear representation of abnormal patterns, facilitating the interpretation and understanding of air quality data.

By successfully implementing the proposed methodology, several benefits can be achieved. First of all, anomalies in air quality data can be promptly identified, allowing for timely interventions and appropriate responses. This can lead to improved public

health outcomes by mitigating the impacts of abnormal air quality events on individuals and communities. Secondly, accurate anomaly detection enhances the overall accuracy and reliability of air quality monitoring systems, providing more reliable information for environmental agencies and policymakers. And finally, the visualization of anomalies offers an intuitive way to comprehend and communicate complex air quality data, aiding in decision-making processes and public awareness.

It's important to note that the specific implementation details and choice of algorithms may vary depending on the nature of the air quality data and the desired detection objectives.

Future research in this field can explore the integration of additional AI techniques, such as deep learning models, for more advanced anomaly detection in air quality data. Additionally, the development of ensemble methods or hybrid models that combine multiple anomaly detection algorithms can potentially improve detection accuracy and robustness. Moreover, the utilization of real-time or streaming data can enable the detection of anomalies in near real-time, providing more timely and dynamic information for decision-making processes.

References

1. V. Evagelopoulos, P. Begou, P. Kassomenos, S. Zoras, IOP Publishing **1123**, 012077 (2022)
2. V. Evagelopoulos, P. Begou, S. Zoras, Atmosphere **13**, 1900 (2022)
3. A. Progiou, N. Liora, I Sebos, C. Chatzimichail, D. Melas, Sustainabilit. **15**, 930 (2023)
4. V. Evagelopoulos, N.D. Charisiou, M. Logothetis, G. Evagelopoulos, C. Logothetis, Climate **10**, 39 (2022)
5. A.A. Diro, N. Chilamkurti, Future Gener. Comp. Syst. **82**, 761 (2018)
6. P. Brown, A. Kejriwal, Frontiers in Big Data **1**, 4 (2018)
7. W.S. Jeon, J.S Park, H.S. Kim, Sensors **19**, 1101 (2019)
8. L. Du, W. Yang, J. Miao, Multimedia Tools and Applic. **78**, 18111 (2019)
9. Y. Bao, B. Yang, X. Zhang, IEEE **6**, 17413 (2018)
10. C. Lin, Y. Chen, H. Tseng, Sensors **17**, 2857 (2017)
11. N. Shaadan, A.A. Jemain, M. T. Latif, Atm. Pol. Res. **6**, 365 (2015).
12. A.A. Gharbia, A.E. Hassanien, IEEE **9**, 19496 (2021)
13. S. Russo, M. Lürig, W. Hao, B. Matthews, K. Villez, Environ. Modelling and Software **134**, 104869 (2020)
14. J. Zhang, J. Zhang, Z. Wu, Sensors, **22**, 6045 (2022)
15. F.T. Liu, K.M. Ting, Z.H. Zhou, 2008 Eighth IEEE International Conference on Data Mining **17**, 413 (2008).
16. V. Evagelopoulos, N.D. Charisiou, S. Zoras, Data in brief, **41**,107883 (2022).