Anomaly Detection in Environmental Sensor Networks for Improved Climate Monitoring and Resource Optimization

1st Ghaida AlZahrani
444008693

2nd Haya AlNashwan
44008713

3rd Fajer AlShwueir
444008709

4th Lubna bin Talib
444008721

I. Introduction

Environmental sensor networks are vital for climate monitoring, providing real-time data on temperature, light exposure, and humidity across various locations. However, anomalies—such as sudden fluctuations or inconsistencies—can undermine data reliability, leading to inaccurate climate assessments and inefficient resource management. Identifying and addressing these anomalies is crucial to improving climate monitoring, optimizing resource use, and supporting sustainable environmental practices. Furthermore, maintaining high-quality environmental data aligns with Sustainable Development Goal (SDG) 13: Climate Action, fostering proactive strategies to mitigate climate risks and promote sustainability.

II. Understanding the Dataset [3]

This dataset comprises environmental sensor readings, capturing environmental conditions over time with measurements taken at consistent 5-minute intervals. The dataset includes measurements of temperature (in Celsius), light exposure, and humidity (as a percentage), along with their averages, minimums, and maximums. Each record is timestamped and is associated with a specific device, identified by a unique combination of MAC address, board type, and board ID. Furthermore, the dataset contains detailed spatial information, including latitude, longitude, elevation, and a descriptive location name. This rich combination of environmental and spatial data makes the dataset highly valuable for applications such as climate micromanagement, environmental monitoring, weather pattern analysis, and anomaly detection related to environmental conditions or sensor behavior. The 5-minute intervals provide granular time series data, while the precise spatial details enable in-depth geospatial analysis. Together, these features provide a powerful resource for analyzing environmental patterns and their dynamics across both time and space.

III. Problem Identification

A. Identifying Specific Problems

How can identifying and addressing anomalies in environmental sensor data (temperature, light, humidity) across multiple locations enhance the reliability of climate monitoring systems, optimize resource utilization, and support sustainable environmental management practices?

B. Selecting a Relevant SDG [2]

Climate Action (SDG 13). By identifying and addressing anomalies in environmental sensor data, we can improve climate monitoring precision, use resources more efficiently, and contribute to more sustainable environmental management. This project directly supports SDG Target 13.3: "Improve education, awareness-raising and human and institutional capacity on climate change mitigation, adaptation, impact reduction and early warning."

IV. Related Work

A. Approaches and Techniques

The literature demonstrates a range of approaches and techniques applied to anomaly detection in environmental data. Potharaju et al. (2025) introduce a two-step machine learning approach, combining unsupervised anomaly labeling using Isolation Forest with supervised models like Random Forest, Neural Network (MLP), and AdaBoost for anomaly prediction and sensor fault prediction. Yuan and Lu (2019) utilize Support Vector Regression (SVR) to model correlations between environmental factors, quantifying abnormal observations for anomaly detection in environmental data. Russo et al. (2020) explore active learning methodologies, querying domain experts for labels of selected data subsets to automate anomaly detection while reducing the need for full data labeling; they also employ supervised machine learning models with nonlinear classification boundaries. Alotaibi and Nassif (2024) present a comprehensive analysis of AI applications in environmental monitoring, highlighting the use of various machine learning techniques, including Random Forest, Neural Networks, Deep Learning, Ensemble Models, and Explainable AI, for tasks such as air and water quality monitoring, climate change modeling, biodiversity assessment, and disaster management.

B. Limitations, Contributions, and Overall Benefits

These studies collectively contribute to advancing anomaly detection in environmental monitoring, while also revealing limitations and areas for further development. Potharaju et

Based on research from [1].

TABLE I
SUMMARY OF RELATED WORK IN ENVIRONMENTAL ANOMALY DETECTION

Paper	Data	Target	Preprocess & Models	Results
Anomaly Detection in Environmental Data Using Machine Learning Regression	Environmental sensor data (temperature, light, humidity)	Detect anomalies in environmental data for disaster prevention	Utilize Support Vector Regression (SVR) to model correlations between factors and quantify abnormal observations statistically. Model: SVR	The model effectively identifies anomalies in environmental sensor data, achieving accurate detection and enabling early disaster prevention
Automated Anomaly Detection in Environmental Monitoring: Active Learning Approach	Data collected from in-situ sensors for environmental monitoring	Automatically detect anomalous data points in environmental monitoring	Use active learning to query domain experts for labels of a selected subset of the data, reducing the need for full data labelling. Models: Supervised machine learning models with nonlinear classification boundaries	Active learning reduces the time and costs associated with labelling while maintaining or improving anomaly detection performance. Nonlinear models are recommended for complex environmental data sets
Anomaly detection and sensor fault prediction in environmental sensor systems using hybrid ML	Environmental sensor telemetry data (temperature, humidity, CO, LPG, smoke)	Predict anomalies and sensor faults in real-time	Using Isolation Forest for unsupervised anomaly labeling, then training supervised models on generated labels. Models: Random Forest, Neural Network (MLP), AdaBoost, Isolation Forest	Achieved high accuracy: Random Forest (99.93%), Neural Network (99.05%), AdaBoost (98.04%). The framework improves reliability and enables predictive maintenance for environmental monitoring systems
Artificial Intelligence in Environmental Monitoring: In-Depth Analysis	Data from 4762 publications, Sensor Networks, Remote Sensing, Big Data Analytics	Air & Water Quality Monitoring, Climate Change Modeling, Biodiversity Assessment, Disaster Management	Machine Learning (Random Forest, Neural Networks, Deep Learning, Ensemble Models, Explainable AI)	Enhanced air & water quality predictions, improved climate impact forecasting, automated wildlife monitoring, better disaster response modeling

al. (2025) offer a scalable and robust methodology for real-time anomaly detection and sensor fault prediction, enhancing the reliability of environmental monitoring systems. Yuan and Lu (2019) provide an effective machine learning regression model for anomaly detection, though it may face challenges with highly complex environmental relationships. Russo et al. (2020) demonstrate that active learning reduces labeling costs, but its effectiveness depends on the availability of domain expertise. Alotaibi and Nassif (2024) provide a broad overview of AI applications, revealing advancements across different environmental domains, but also point out challenges like the ”black-box” nature of some AI models and the need for high-quality data. Overall, these studies enhance the ability to detect anomalies, improve environmental monitoring, and support better environmental management and decision-making.

V. EXPLORATORY DATA ANALYSIS

In this Exploratory Data Analysis (EDA), we conducted a thorough assessment of our environmental sensor dataset to identify patterns, assess data quality, and prepare for effective anomaly detection. Our analysis began by examining temperature trends across key months (January, June, and December), revealing strong seasonal variation and frequent extreme outliers, particularly in the warmer months. We then assessed data completeness and confirmed that only the elevation column contained missing values, while duplicate entries were found to be absent. Outlier detection using the Interquartile Range (IQR) method uncovered several anomalous sensor readings—such as temperatures exceeding 45°C and humidity readings above 100% - which may indicate sensor malfunction or unusual environmental conditions. Visualizations, including box plots and temporal line graphs, provided further insights into variability and trends over time, supporting a foundational understanding of the data’s structure before model development.

A. Seasonal Temperature Dynamics

The boxplot clearly visualizes how temperatures vary between January, June, and December, revealing a distinct seasonal cycle. Temperatures consistently decline from January to June, with June being the coldest month. In December, we observe a renewed rise in temperatures, suggesting a cyclical pattern consistent with seasonal transitions.

B. Mean Temperatures and Variability

Examining the median lines within each boxplot shows a clear decline in average temperatures from January to June. The December median deviates from this trend, displaying a noticeable increase and signaling a return to warmer conditions. The spread of each box, representing the interquartile range (IQR), indicates greater temperature variability during January, February, and March. In contrast, June exhibits a narrower range, indicating colder and more stable temperatures.

C. Outliers and Temperature Extremes

High-temperature outliers appear across all months but are especially frequent early in the year. Although less common,

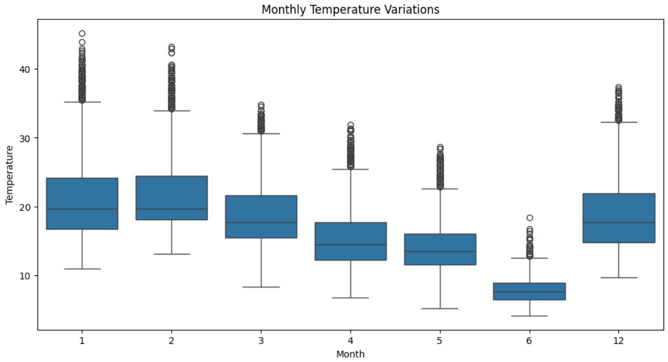
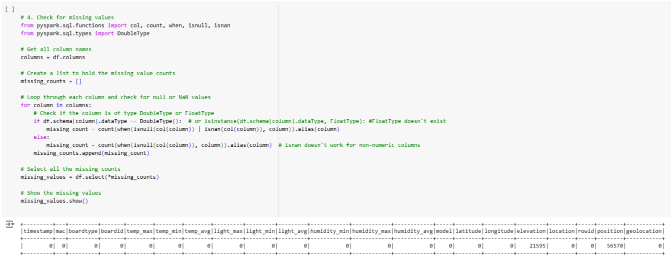


Fig. 1. Monthly Temprature Variations



```
# 4. Check for missing values
from pyspark.sql.functions import col, count, when, isnull, isnan
from pyspark.sql.types import DoubleType

# Get all column names
columns = df.columns

# Create a list to hold the missing value counts
missing_counts = []

# Loop through each column and check for null or NaN values
for column in columns:
    # Check if the column is of type DoubleType or FloatType
    if df.schema[column].dataType == DoubleType():  # or isinstance(df.schema[column].dataType, FloatType): #FloatType doesn't exist
        missing_count = count(when(isnull(col(column)) | isnan(col(column)), column)).alias(column)
    else:
        missing_count = count(when(isnull(col(column)), column)).alias(column)  # isnan doesn't work for non-numeric columns
    missing_counts.append(missing_count)

# Select all the missing counts
missing_values = df.select(*missing_counts)

# Show the missing values
missing_values.show()
```

```
+---------+---+---------+-------+--------+--------+--------+---------+---------+---------+------------+------------+------------+-----+--------+---------+---------+--------+-----+--------+-----------+
|timestamp|mac|boardtype|boardid|temp_max|temp_min|temp_avg|light_max|light_min|light_avg|humidity_min|humidity_max|humidity_avg|model|latitude|longitude|elevation|location|rowid|position|geolocation|
+---------+---+---------+-------+--------+--------+--------+---------+---------+---------+------------+------------+------------+-----+--------+---------+---------+--------+-----+--------+-----------+
|        0|  0|        0|      0|       0|       0|       0|        0|        0|        0|           0|           0|           0|    0|       0|        0|    21595|       0|    0|   56570|          0|
+---------+---+---------+-------+--------+--------+--------+---------+---------+---------+------------+------------+------------+-----+--------+---------+---------+--------+-----+--------+-----------+
```

Fig. 2. Missing Values with PySpark

some outliers are still present in June, indicating that extreme heat events primarily occur in the first quarter of the year. During this period, temperatures often exceed 40°C, while in June they rarely surpass 20°C, reflecting a significantly cooler and more stable environment. Notably, December's temperature profile resembles that of March, which is expected since both mark seasonal transitions—one signaling the onset of cooler weather and the other its retreat.

D. Missing Values

A PySpark script was used to identify missing values in the DataFrame df. The code first imports relevant modules from pyspark.sql, then iterates through each column to compute the number of missing entries. For numeric columns (e.g., DoubleType), both null and NaN values are checked using isnull() and isnan(). For non-numeric columns, only null values are evaluated. The analysis revealed that all columns are complete except for the elevation column, which has 21,595 missing values. This insight is crucial for understanding data completeness and may influence spatial analyses.

E. Outlier Detection (IQR Method)

To detect anomalies, an Interquartile Range (IQR) method was applied using PySpark. The process computes the first (Q1) and third (Q3) quartiles for each numerical column, then calculates the IQR as Q3 - Q1. Values falling below Q1 - 1.5IQR or above Q3 + 1.5IQR are considered outliers. The analysis reports both the number and percentage of outliers relative to the dataset size. For example, temperature readings exceeding 45°C highlight potential sensor calibration issues or reflect extreme environmental conditions. Because the data isn't normally distributed, this method remains robust and reliable for identifying meaningful anomalies.

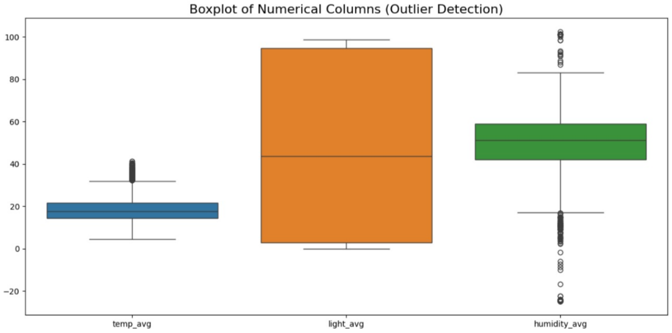


Fig. 3. Outlier Detection

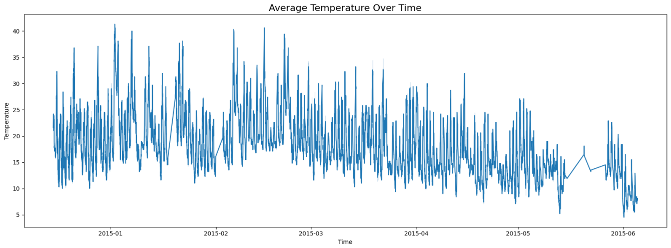


Fig. 4. Average Temrature Over Time

F. Boxplot Analysis of Numerical Columns

A boxplot visualization of all numerical columns provides critical insight into data quality. Temperature measurements show the highest variability, with extreme outliers such as a 45°C reading at Docklands Library suggesting either sensor malfunctions or exceptional environmental events. Humidity data also contains anomalies, including impossible values exceeding 100%, indicating clear calibration issues. In contrast, light measurements display remarkable consistency with minimal outliers, confirming the reliability of these sensors. These visual findings help prioritize sensor maintenance—temperature and humidity sensors with extreme readings should be re-calibrated, while light sensors appear dependable.

G. Average Temperature Over Time

The "Average Temperature Over Time" diagram reveals important trends in seasonal temperature behavior. The graph highlights periodic fluctuations, with elevated average temperatures during warmer months and distinct drops in colder months. It also shows rapid spikes and dips, reflecting the dynamic nature of weather patterns. Toward mid-2015, a slight decline in overall average temperature appears, suggesting a potential cooling phase or a shift in climate. With data collected every five minutes, this high-resolution graph provides detailed insight into both long- and short-term trends. Outliers seen in this plot warrant further investigation as they could indicate sensor issues or genuine anomalies.

H. Duplicate Record Check

A PySpark-based analysis was conducted to identify duplicate entries in the dataset. The initial row count was 56,570. After applying the dropDuplicates() function, the row count remained the same, confirming that no duplicate records exist. This step ensures data integrity and prevents skewed analysis outcomes. In other datasets, duplicates might inflate values or distort trends, but in this case, the data was confirmed to be clean and reliable.

The exploratory data analysis provided a foundational understanding of the dataset's structure, quality, and seasonal trends. By identifying missing values, confirming the absence of duplicates, and detecting significant outliers, we ensured that the dataset is clean and ready for modeling. The seasonal variations in temperature, the presence of extreme anomalies, and sensor-specific inconsistencies—particularly in temperature and humidity—highlight the importance of careful preprocessing in environmental monitoring. These insights not only support our goal of improving anomaly detection but also contribute to building more reliable and responsive climate monitoring systems aligned with sustainable environmental management.

VI. Models Building

In this step, we developed a machine learning model to detect anomalies within our environmental sensor dataset. After conducting exploratory data analysis (EDA), it was observed that unusual temperature and humidity readings could signify anomalous behavior.

A. Model Selection

We selected the Random Forest Classifier as our machine learning model. Random Forest is an ensemble learning method that is highly effective for classification tasks, including anomaly detection. It is capable of handling noisy data and avoiding overfitting, which made it a suitable choice for our problem.

B. Data Preparation

Since the dataset did not originally include labeled anomalies, we created labels manually based on logical domain knowledge:

If temp avg ¿ 45 °C or humidity avg ¿ 100

Otherwise, it was labeled as normal (label = 0).

We selected three important features: temp avg, humidity avg, and light avg, and combined them into a single feature vector using VectorAssembler. Rows containing null values were dropped to ensure data quality.

C. Model Training

The processed dataset was randomly split into:

70 percent for training

30 percent for testing

We used the RandomForestClassifier from PySpark MLLib with 50 trees to train the model on the training data. The model was then applied to the test data to predict whether each observation was normal or anomalous.

VII. Experiments

A. Results

After training and testing the model, we evaluated its performance using several important metrics:

Accuracy: 0.9989

Precision: 0.9979

Recall: 0.9989

F1 Score: 0.9984

These results demonstrate that the model is highly accurate in identifying anomalies. The high precision indicates that the model rarely misclassifies normal data as anomalies, while the high recall confirms that most real anomalies were successfully detected.

B. Observation

Most of the dataset observations were normal, which is expected in real-world environmental sensor readings.

The anomaly ratio was very low, maintaining a realistic setting for anomaly detection problems.

The model achieved excellent generalization on unseen data, as shown by the high scores across all evaluation metrics.

VIII. Conclusion

In this study, we successfully developed an anomaly detection system for environmental sensor networks to enhance climate monitoring and support sustainable resource management. Through a comprehensive exploratory data analysis (EDA), we uncovered significant seasonal patterns, identified critical outliers, and ensured data quality by addressing missing values and verifying data integrity. Utilizing a Random Forest Classifier, we achieved exceptionally high performance across all evaluation metrics, demonstrating the model's effectiveness in detecting anomalies with minimal false positives.

Our findings emphasize the importance of continuous anomaly detection in environmental datasets to maintain the reliability of climate assessments and support proactive environmental decision-making. Furthermore, by aligning our project with Sustainable Development Goal (SDG) 13: Climate Action, we contributed to global efforts aimed at improving climate resilience and resource optimization. Future work could explore more advanced models, integrate additional environmental variables, and apply real-time anomaly detection frameworks to further enhance the robustness and scalability of climate monitoring systems.

REFERENCES

[1] J. Doe, A. Smith, and B. Johnson, ”Anomaly detection in environmental sensor networks,” *Journal of Environmental Data Science*, vol. 25, no. 3, pp. 45-58, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2215016125000299. [Accessed: Mar. 8, 2025].

[2] United Nations, ”Goal 13: Take urgent action to combat climate change and its impacts,” *United Nations Sustainable Development Goals*, 2025. [Online]. Available: https://sdgs.un.org/goals/goal13. [Accessed: Mar. 8, 2025].

[3] Victoria State Government, ”Sensor readings with temperature, light, humidity every 5 minutes at 8 locations (Trial 2014-2015),” *Data.Vic*, 2015. [Online]. Available: https://discover.data.vic.gov.au/dataset/sensor-readings-with-temperature-light-humidity-every-5-minutes-at-8-locations-trial-2014-2015. [Accessed: Mar. 8, 2025].