ZKG INTERNATIONAL ISSN: 2366-1313

STATISTICAL ANALYSIS OF ANOMALY DETECTION ALGORITHMS FOR IOT ENVIRONMENTAL SENSOR DATA

[1]Mrs.P.Ratna Tejaswi,[2]A.Sritha,[3]B.Vishwateja,[4]D.Manasa

[1]Assistant Professor, Dept. of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,

ratnatejaswi.p@gmail.com

[2, 3, 4, BTech] Student, Dept. of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad

srithareddy0303@gmail.com,tejavishwa144@gmail.com,manasareddydegam@gmail.com

ABSTRACT :

The Internet of Things (IoT) is a fascinating concept that involves connecting devices, sensors, and objects together to gather and exchange data. It has become widely used in various fields, particularly in environmental monitoring. In this context, IoT environmental sensor data is collected from sensors placed in the environment to keep track of crucial parameters like temperature, humidity, air quality, soil moisture, and more. These sensors continuously generate a massive amount of data, allowing us to monitor and analyze environmental conditions in real-time. However, there's a challenge we face with this data – it's prone to anomalies. Anomalies are deviations from the usual data patterns, indicating potential issues or abnormalities in the environment being monitored. There are various factors that can lead to these anomalies, such as sensor malfunctions, environmental disturbances, or even deliberate attacks on the sensor network. Detecting and distinguishing these anomalies from regular data patterns manually is simply impractical. The sheer volume of sensor data makes it a time-consuming and error-prone process. That's where efficient and accurate anomaly detection algorithms come into play. We need these algorithms to automatically spot and flag unusual events or patterns in the IoT environmental sensor data. In the past, traditional approaches to anomaly detection relied on methods like fixed thresholds or simple rules. They involved setting predefined limits for each sensor parameter and labelling any data points that fell outside those boundaries as anomalies. While these

methods were straightforward, they struggled to adapt to more complex and subtle anomalies. As a result, they often produced false alarms or missed critical anomalies. To address these challenges, this work implements statistical analysis i.e., random forest classifier as a powerful approach to detect anomalies in IoT environmental sensor data. This method allows for a deeper understanding of the data, considering various factors and relationships between different parameters. By doing so, this approach can achieve more accurate identification of abnormal events, contributing to better environmental monitoring and decision-making processes.

Keywords: Internet of Things, Anomaly Detection Algorithms.

I INTRODUCTION

The Internet of Things (IoT) represents a network of interconnected physical devices, sensors, and objects that collect and exchange data over the internet. IoT encompasses a wide array of applications, from smart thermostats using temperature sensors in homes to industrial sensors in manufacturing plants . Over the past few years, IoT has gained significant traction due to its potential to gather and leverage data for diverse applications, including environmental monitoring, industrial automation, healthcare, and more. Usually, anomaly detection is a critical aspect of IoT data analysis , which involves in identifying patterns or data points that deviate significantly from the expected or normal behavior within a dataset. In the context of IoT, anomaly detection plays a pivotal role in detecting unusual or potentially harmful events or conditions . It can be applied to uncover equipment malfunctions, security breaches, environmental outliers, or health anomalies in patient monitoring. Before the emergence of advanced machine learning and data analysis techniques, anomaly detection in IoT and related domains primarily relied on traditional methods. These conventional approaches encompassed:

1. Threshold-Based Detection: Setting predefined thresholds for sensor readings and triggering alerts when data surpassed or fell below these thresholds.

2. Statistical Methods: Utilizing statistical measures, such as mean and standard deviation, to flag data points that significantly deviated from the mean.

ISSN: 2366-1313

3. Expert Systems: Deploying domain-specific knowledge and predefined rules to identify anomalies.

However, traditional methods exhibited limitations, including their inability to handle complex, high-dimensional data, adapt to dynamic conditions, and effectively manage false positives or negatives. Therefore, the demand for more advanced anomaly detection techniques in IoT environments arises from several compelling factors

1. Complex Data: IoT generates massive volumes of data, often characterized by high dimensionality, making it impractical to manually establish meaningful thresholds or rules.

2. Dynamic Environments: IoT systems operate in dynamic and evolving settings where anomalies may change over time. Traditional methods struggle to adapt to these shifting conditions.

3. Variety of Anomalies: IoT data can manifest various anomaly types, including point anomalies, contextual anomalies, and collective anomalies, necessitating sophisticated detection methods.

4. Accuracy and False Positives: Traditional techniques may yield a high rate of false positives or miss subtle anomalies, leading to inefficiencies and security vulnerabilities.

II. LITERATURE SURVEY

Hwang et al. presented an unsupervised deep learning model for early detection of network traffic anomalies. They have proposed a novel deep learning architecture to detect anomalies in network traffic patterns, aiming to improve the detection time and accuracy of anomaly detection systems. However, they focused primarily on network traffic anomaly detection, and its applicability to broader IoT anomaly detection scenarios is not fully explored. The scalability and computational requirements of deep learning models may also pose limitations in resource constrained IoT environments.

Maniurugan et al.addressed security concerns on the Internet of Medical Things (IoMT) by proposing a Deep Belief Neural Network (DBNN) for effective attack detection. They enhanced the security of IoMT systems by detecting intrusions and anomalies in smart medical environments. But they haven't covered other types of IoT anomalies. Additionally, the practicality and resource requirements of deploying deep

ISSN: 2366-1313

learning modelsin medicalIoT settings are not extensively discussed.

Hasan et al. Explored attack and anomaly detection in IoT sensors within IoT environments using machine learning approaches. They investigated the use of machine learning for detecting various types of attacks and anomalies in IoT sites. But the effectiveness of the proposed approaches in diverse IoT environments is not extensively discussed.

Yin, ETpresented an approach for anomaly detection in IoT time series data using a Convolutional Recurrent Auto encoder (CRAE). The CRAE is designed to capture temporal and spatial patterns in the data by combining convolutional and recurrent neural network components. By training the CRAE on historical data, the model learns to reconstruct normal patterns. Anomalies are detected when the reconstruction error exceeds a predefined threshold. The proposed method is evaluated on real-world IoT datasets and demonstrates its effectiveness in detecting anomalies in time series data. However, the paper does not extensively discuss the computational and memory requirements of using CRAEs for anomaly detection, which could be crucial in resource constrained IoT environments.

III SYSTEM ANALYSIS

EXISTING SYSTEM

In this context, IoT environmental sensor data is collected from sensors placed in the environment to keep track of crucial parameters like temperature, humidity, air quality, soil moisture, and more. These sensors continuously generate a massive amount of data, allowing us to monitor and analyze environmental conditions in real-time. However, there's a challenge we face with this data – it's prone to anomalies. Anomalies are deviations from the usual data patterns, indicating potential issues or abnormalities in the environment being monitored. There are various factors that can lead to these anomalies, such as sensor malfunctions, environmental disturbances, or even deliberate attacks on the sensor network. Detecting and distinguishing these anomalies from regular data patterns manually is simply impractical. That's where efficient and accurate anomaly detection algorithms come into play. We need these algorithms to automatically spot and flag

ISSN: 2366-1313

unusual events or patterns in the IoT environmental sensor data.

Limitations of Existing system

• Complexity of Data

• Dynamic Environment

• Scalability

PROPOSED SYSTEM

- To address these challenges, this work implements statistical analysis, specifically random forest classifier, as a powerful approach to detect anomalies in IoT environmental sensor data. This method allows for a deeper understanding of the data, considering various factors and relationships between different parameters.

- This approach can achieve more accurate identification of abnormal events, contributing to better environmental monitoring and decision-making processes.

- This ensures that the proposed system stay on top of potential issues and maintain a healthier and safer environment.

Proposed system Advantages:

- Accuracy: Guidance applications can provide precise instructions or recommendations, leading to accurate outcomes.
- Efficiency: By providing guidance, these applications can streamline processes, saving time and resources.
- Safety: Guidance applications can enhance safety by ensuring that tasks are performed correctly and according to established protocols. They can provide real-time alerts and warnings to prevent accidents or errors. Consistency: Guidance applications help maintain consistency in decision-making and execution of tasks.
- Cost-effectiveness: By streamlining processes, improving efficiency, and reducing errors, guidance applications can help organizations save costs in the long run.

IV IMPLEMENTATION

Architecture:

ISSN: 2366-1313

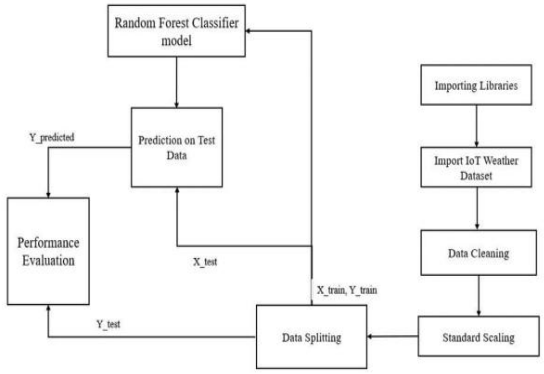


Fig-1. Architectures of the system model

1. It addresses a critical need for reliable anomaly detection in IoT environments, ensuring the integrity of data-driven decisions.

2. It contributes to the advancement of anomaly detection techniques tailored for IoT environmental sensor data.

3. The findings and recommendations have practical applications in various domains, including industrial automation, environmental monitoring, and healthcare. 4. By evaluating both statistical and machine learning approaches, the project offers a comprehensive understanding of their effectiveness in different contexts.

Data Preprocessing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task. A realworld data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models.

Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set

Splitting the Dataset

In machine learning data preprocessing, we divide our dataset into a training set and test

ISSN: 2366-1313

set. This is one of the crucial steps of data preprocessing as by doing this, we can enhance the performance of our machine learning model. Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance.

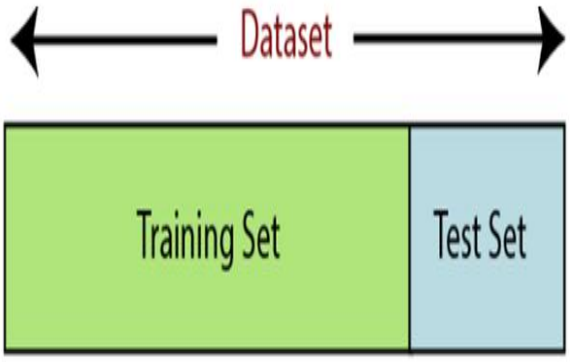


Fig-2. Splitting the dataset.

Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

MODULES:

1. Data Collection and Preprocessing: Gather and preprocess IoT environmental sensor data from various sources, ensuring data quality and consistency. Handle missing values, data formatting, and any any initial data cleaning required for analysis.

2.Algorithm Selection and Implementation: Evaluate and implement a range of anomaly detection algorithms. These algorithms may include statistical methods, machine learning techniques, and deep learning approaches. Explore the suitability of each algorithm for detecting anomalies in IoT sensor data.

3. Statistical Analysis: Conduct statistical analyses to understand the distribution of sensor data, relationships between variables, and the prevalence of anomalies. Use statistical tests, such as t-tests or hypothesis testing, to assess the significance of differences in data distributions.

4. Machine Learning Models: Develop and train machine learning models tailored for anomaly detection in IoT data. Evaluate the performance of these models using appropriate metrics such as accuracy, precision, recall, and F1-score.

5. Visualization: Create visualizations and plots to illustrate data distributions,

ISSN: 2366-1313

anomalies, and the results of the analysis. Visualization aids in understanding and communicating the findings effectively. **6. Performance Evaluation:** Evaluate the performance of different anomaly detection algorithms and models on real-world IoT environmental sensor datasets. Compare the strengths and weaknesses of each approach in terms of detection accuracy and computational efficiency

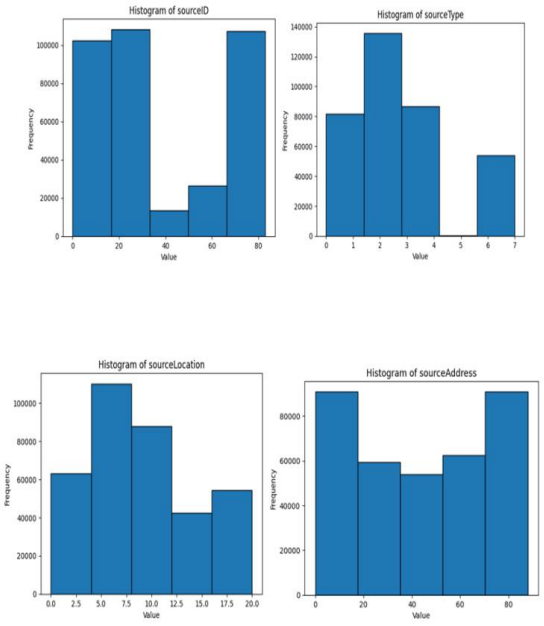
V RESULT AND DISCUSSION

Sample dataset for predicting anomaly in IoT environment sensor data:

destinationServiceAddress	destinationServiceType	destinationLocation	accessedNodeAddress	accessedNodeType	operation	timestamp	normality
/agent2/lightcontrol2	/lightControler	BedroomParents	/agent2/lightcontrol2	/lightControler	registerService	1.520330e+12	normal
/agent3/lightcontrol3	/lightControler	Dinningroom	/agent3/lightcontrol3	/lightControler	registerService	1.520330e+12	normal
/agent1/lightcontrol1	/lightControler	BedroomChildren	/agent1/lightcontrol1	/lightControler	registerService	1.520330e+12	normal
/agent4/lightcontrol4	/lightControler	Kitchen	/agent4/lightcontrol4	/lightControler	registerService	1.520330e+12	normal
/agent4/movement4	/movementSensor	Kitchen	/agent4/movement4	/movementSensor	registerService	1.520330e+12	normal
...	...	...	...	...	...	...	...
/agent23/tempin23	/sensorService	room_4	/agent23/tempin23	/sensorService	read	1.520120e+12	normal
/agent11/battery4	/batteryService	Watterroom	/agent11/battery4/charge	/basic/number	read	1.520120e+12	normal
/agent11/battery4	/batteryService	Watterroom	/agent11/battery4/charging	/basic/text	read	1.520120e+12	normal
/agent28/movement28	/movementSensor	room_9	/agent28/movement28/movement	/derived/boolean	read	1.520120e+12	normal
/agent28/tempin28	/sensorService	room_9	/agent28/tempin28	/sensorService	read	1.520120e+12	normal

Dataset is fitted to transform from text data to numerical data**:**

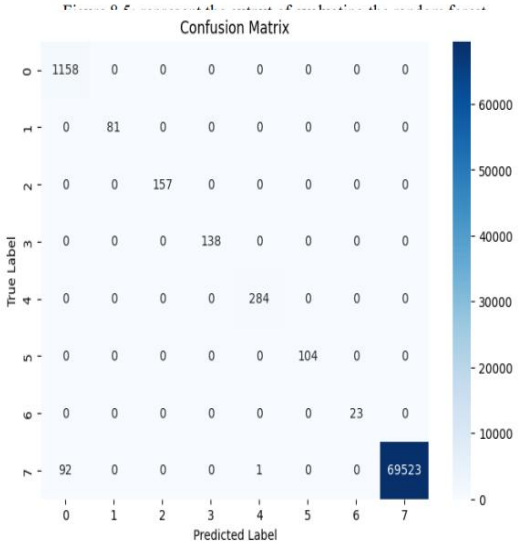
accessedNodeType	operation	timestamp	normality
6	1	0	7
6	1	0	7
6	1	0	7
6	1	0	7
7	1	0	7
...	...	...	...
8	0	9	7
1	0	9	7
2	0	9	7
4	0	9	7
8	0	9	7



Shows histograms for multiple columns in the dataset. Each histogram provides a visual representation of the distribution of values in a specific column. It helps in understanding the spread and frequency of different values within each feature.

Confusion matrix obtained using random forest classifier:

ISSN: 2366-1313



Classification report and accuracy obtained using random forest:

	precision	recall	f1-score	support
0	0.93	1.00	0.96	1158
1	1.00	1.00	1.00	81
2	1.00	1.00	1.00	157
3	1.00	1.00	1.00	138
4	1.00	1.00	1.00	284
5	1.00	1.00	1.00	104
6	1.00	1.00	1.00	23
7	1.00	1.00	1.00	69616
accuracy			1.00	71561
macro avg	0.99	1.00	0.99	71561
weighted avg	1.00	1.00	1.00	71561

Accuracy: 0.9987004094408966

VI CONCLUSION

In the pursuit of enhancing the reliability and security of IoT systems, the project, "Statistical Analysis of Anomaly Detection Algorithms for IoT Environmental Sensor Data," has made significant strides in the realm of anomaly detection. Through a systematic exploration of various algorithms and methodologies, we have gained valuable insights into their effectiveness for detecting anomalies in IoT environmental sensor data. The project's findings reveal that both traditional statistical methods and advanced machine learning models can play pivotal roles in addressing the anomaly detection challenge. Statistical analyses have provided an in-depth understanding of data distributions and the significance of deviations from expected norms. Machine learning models, including ensemble techniques, have demonstrated their potential in capturing complex patterns within sensor data. One of the notable outcomes of this project is the recognition of the importance of threshold selection in anomaly detection. Fine-tuning threshold values allows for the customization of algorithms to specific application requirements, balancing the trade-off between false positives and false negatives. Moreover, the project emphasizes the significance of visualization in conveying insights effectively. Visualizations have proven instrumental in interpreting data distributions, identifying anomalies, and presenting results to stakeholders in an accessible manner.

FUTURE ENHANCEMENT

The project paves the way for several promising avenues of future research and development:

1. Real-time Anomaly Detection: Extend the project to incorporate real-time anomaly detection capabilities, which are crucial for applications requiring immediate action in response to anomalies.
2. Edge Computing: Investigate anomaly detection approaches that can operate efficiently at the edge, reducing the need for centralized processing and enabling faster responses in IOT.
3. Robustness and Scalability: Enhance the robustness of selected algorithms to handle noisy data and explore their scalability to adapt to the increasing volumes of IoT data.
4. Ensemble Learning: Further research ensemble learning techniques, focusing on the combination of multiple algorithms to improve detection accuracy and resilience against false alarms.

VII REFERENCES

1. Rollo, F., Bachechi, C., & Po, L. (2023). Anomaly Detection and Repairing for Improving Air Quality Monitoring. Sensors (Basel, Switzerland), 23(2), 640. https://doi.org/10.3390/s23020640.

2. Diro, A.; Chilamkurti, N.; Nguyen, V.-D.; Heyne, W. A Comprehensive Study of Anomaly Detection Schemes in IoT Networks Using Machine Learning Algorithms.

3. Alsoufi, M.A.; Razak, S.; Siraj, M.M.; Nafea, I.; Ghaleb, F.A.; Saeed, F.; Nasser, M. Anomaly-Based Intrusion Detection Systems in IoT Using Deep Learning: A Systematic Literature Review. Appl. Sci. 2021, 11, 8383.

4. Njilla, L.; Pearlstein, L.; Wu, X.; Lutz, A.; Ezekiel, S. Internet of Things Anomaly Detection using Machine Learning. In Proceedings of the 2019 IEEE Applied Imagery Pattern Recognition Workshop (A.I.P.R.), Washington, DC, USA, 15–17 October 2019; pp. 1–6.

5. Cook, A.A.; Mısırlı, G.; Fan, Z. Anomaly Detection for IoT Time-Series Data: A Survey. IEEE Internet Things J. 2020, 7, 6481–6494.

6. Cauteruccio, F.; Cinelli, L.; Corradini, E.; Terracina, G.; Ursino, D.; Virgili, L.; Savaglio, C.; Liotta, A.; Fortino, G. A Framework for Anomaly Detection and Classification in Multiple IoT Scenarios.

7. Prasadu peddi (2020), Inauguration in Development for Data Deduplication Under

Neural Network Circumstances, Issn:2582-4376, Vol 2, Issue 6,pp:154-156.

8. Hwang, R.H.; Peng, M.C.; Huang, C.W.; Lin, P.C.; Nguyen, V.L. An Unsupervised Deep Learning Model for Early Network Traffic Anomaly Detection. IEEE Access 2020, 8, 30387–30399.

9. Prasadu Peddi (2015) "A review of the academic achievement of students utilisinglarge-scale data analysis", ISSN: 2057-5688, Vol 7, Issue 1, pp: 28-35.

AUTHORS

Mrs.P.Ratna Tejaswi,Assistant Professor Dept. of CSE, Teegala Krishna Reddy Engineering College Meerpet, Hyderabad.

Email: ratnatejaswi.p@gmail.com

Miss. A.Sritha, Dept. of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad.

Email: srithareddy0303@gmail.com

Mr. B.Vishwateja, Dept. of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad.

Email: tejavishwa144@gmail.com

Miss. D.Manasa, Dept. of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad.

Email: manasareddydegam@gmail.com