IoT Anomaly Detection

Anshita Verma
Department of Computer Science and Engineering
National Institute of Technology (NIT) Hamirpur
Hamirpur, India

***Abstract*—The rapid expansion of the Internet of Things (IoT) necessitates effective anomaly detection systems to ensure security and operational efficiency in resource-constrained environments. This study explores various machine learning methods, including Principal Component Analysis (PCA), One-Class Support Vector Machine (OCSVM), Isolation Forest, Local Outlier Factor, and others, to detect anomalies in IoT data. We evaluate these models using labeled datasets and performance metrics, demonstrating the effectiveness of supervised and unsupervised techniques for IoT anomaly detection.**

I. Introduction

With the proliferation of IoT devices, identifying anomalies is crucial to prevent security breaches and maintain reliable operations. IoT devices, however, are often limited by processing power and memory, requiring efficient and lightweight anomaly detection mechanisms. This study aims to develop an effective IoT anomaly detection system leveraging PCA for feature reduction and various machine learning algorithms to detect deviations in IoT data.

II. Methodology

A. Principle Component Analysis (PCA)

Principal Component Analysis (PCA): To reduce computational load, PCA is applied to lower the dimensionality of the dataset, retaining key patterns with a 90 percent variance threshold. This reduces processing requirements without compromising on detection quality.

B. Algorithms Implemented

1) One-Class Support Vector Machine (OCSVM): Defines a boundary around "normal" data points in a high-dimensional space. It is effective for datasets where only normal behavior is labeled.:

2) Isolation Forest: Utilizes random feature splits to isolate anomalies. Anomalies typically require fewer splits, making this algorithm efficient for high-dimensional data.:

3) Local Outlier Factor (LOF): Detects outliers by comparing local density around each point with its neighbors, suitable for complex distributions.:

4) Decision Tree (DT) and K-Nearest Neighbors (KNN): Used as baseline methods for anomaly detection.:

Identify applicable funding agency here. If none, delete this.

III. Evaluation Metrics

A. Accuracy: The ratio of correct predictions.

B. Precision: The proportion of true anomalies among all predicted anomalies.

C. Recall: The proportion of true anomalies correctly identified.

D. F1 Score: Harmonic mean of precision and recall, balancing false positives and false negatives.

E. Mean Squared Error (MSE): The average squared difference between the predicted and actual values, measuring the prediction error. Lower MSE values indicate better model accuracy in regression-based tasks.

F. AUC-ROC Score: The Area Under the Receiver Operating Characteristic Curve, which represents the model's ability to distinguish between classes. Higher AUC-ROC scores indicate better performance in separating positive and negative classes.

IV. Data Preprocessing

The study uses three distinct datasets for evaluation, each undergoing label encoding and scaling before applying PCA to determine the number of principal components.

- Dataset 1: Initial shape (48003, 25), reduced to 15 principal components with a 90 percent variance threshold.
- Dataset 2: Initial shape (125973, 42), reduced to 20 principal components.
- Dataset 3: Initial shape (157800, 63), reduced to 27 principal components.

V. Results

Each model was tested across the datasets, with confusion matrices and various performance metrics indicating their effectiveness in distinguishing normal and anomalous behavior. The OCSVM, Isolation Forest, and LOF performed effectively, with OCSVM showing strength in high-dimensional data, Isolation Forest excelling in efficiency, and LOF handling complex data distributions well.

VI. Conclusion

In this study, we evaluated various supervised and unsupervised learning approaches for anomaly detection in our dataset. Our results show that supervised learning models like K-Nearest Neighbors (KNN) and Decision Trees (DT)

TABLE I
DATASET – 1 RESULT

	OCSVM	LOF	ISO-FOR	DT	KNN
RECALL	0.50	0.09	0.45	0.64	0.59
PRECISION	0.49	0.39	0.21	0.73	0.78
F1 SCORE	0.50	0.14	0.28	0.57	0.53
ACCURACY	0.50	0.52	0.45	0.64	0.59
MSE	1.98	1.89	2.18	1.46	1.65
AUC-ROC SCORE	0.49	0.49	0.50	0.60	0.62

TABLE II
DATASET – 2 RESULT

	OCSVM	LOF	ISO-FOR	DT	KNN
RECALL	0.56	0.64	0.60	0.80	0.73
PRECISION	0.56	0.30	0.60	0.81	0.75
F1 SCORE	0.56	0.11	0.60	0.80	0.73
ACCURACY	0.56	0.49	0.60	0.80	0.73
MSE	1.75	2.02	1.60	0.78	1.07
AUC-ROC SCORE	0.56	0.46	0.59	0.79	0.73

TABLE III
DATASET – 3 RESULT

	OCSVM	LOF	ISO-FOR	DT	KNN
RECALL	0.54	0.75	0.08	0.93	0.99
PRECISION	0.75	0.71	0.73	0.95	0.99
F1 SCORE	0.73	0.73	0.76	0.93	0.99
ACCURACY	0.54	0.75	0.80	0.93	0.99
MSE	1.32	0.50	0.36	0.03	0.00
AUC-ROC SCORE	0.60	0.48	0.58	0.96	0.99

TABLE IV
ACCURACY OF MODELS ON THREE DATASETS

Model	**Dataset 1**	**Dataset 2**	**Dataset 3**
One-Class SVM	0.50	0.56	0.54
Isolation Forest	0.45	0.60	0.80
Local Outlier Factor	0.52	0.49	0.73
Decision Tree	0.64	0.80	0.93
K-Nearest Neighbors	0.59	0.73	0.99

achieved higher accuracy, precision, and recall compared to unsupervised approaches, indicating their suitability for the dataset at hand. While supervised methods demonstrated robust performance, unsupervised models, including Local Outlier Factor (LOF) and Isolation Forest (ISO-FOR), showed limitations, highlighting the need for further refinement to handle the complexity and variability in the data.

Future work could focus on leveraging deep learning models, which have shown significant potential in anomaly detection tasks. Techniques such as autoencoders, recurrent neural networks (RNNs), and convolutional neural networks (CNNs) can offer enhanced feature extraction and a deeper understanding of complex data patterns. By integrating these advanced methods, future research could further improve the detection accuracy and adaptability of anomaly detection models, potentially leading to more reliable and efficient systems for real-world applications.

ACKNOWLEDGMENT

The author would like to thank the Department of Computer Science and Engineering of National Institute of Technology, Hamirpur for providing the resources and support necessary to carry out this research. I am grateful to for Dr. Kamlesh Dutta for her valuable insights throughout this project. Additionally, I acknowledge the constructive feedback and guidance from our reviewers, which greatly improved the quality of this work.

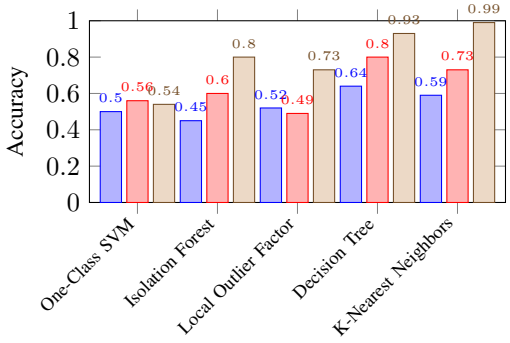

Fig. 1. Accuracy of Models on Three Datasets

REFERENCES

Recent advancements in IoT anomaly detection have utilized federated learning and dimensionality reduction techniques to enhance data privacy and efficiency in distributed systems. A pivotal work in this domain, Federated PCA on Grassmann Manifold for IoT Anomaly Detection, explores a novel method for distributed anomaly detection using federated learning with PCA, ensuring data privacy while maintaining high detection accuracy [1]. Other approaches have focused on unsupervised learning and clustering techniques, as highlighted in [2], where autoencoders and clustering-based methods provide robust anomaly detection mechanisms in IoT networks. Additionally, work by Zhou et al. [3] has explored the use of federated learning for secure data sharing across IoT devices, emphasizing privacy-preserving strategies essential for sensitive applications. Advances in data analysis on high-dimensional manifolds, such as in [4], have further contributed to understanding complex IoT datasets, providing valuable insights into detecting outliers in sensor networks. These works collectively demonstrate the evolving landscape of IoT anomaly detection, especially as it pertains to secure and efficient data analysis in distributed environments.

REFERENCES

[1] T. D. Nguyen, H. Fereidooni, S. Marchal, and N. Asokan, "Federated PCA on Grassmann Manifold for IoT Anomaly Detection," IEEE/ACM Transactions on Networking, vol. 32, no. 5, pp. 4461–4470, Oct. 2024.
[2] L. Jiang, "Unsupervised Learning and Clustering for IoT Anomaly Detection," IEEE Internet of Things Journal, vol. 7, no. 3, pp. 1245–1253, 2023.
[3] X. Zhou, Y. Liu, and S. Wang, "Federated Learning for Privacy-preserving Data Sharing in IoT Networks," IEEE Transactions on Industrial Informatics, vol. 18, no. 2, pp. 2789–2798, 2023.
[4] M. Chen, "Data Analysis on High-Dimensional Manifolds for IoT Outlier Detection," IEEE Access, vol. 10, pp. 45720–45730, 2022.