Data Mining Project Report

Academic Year- 2022-23

On

Outlier Detection in Wireless Sensor Network

Submitted by

Name	Enrollment No.	BT/CSE/ECE
1. Akshat Dikshit	BT20HCS176	CSE
2. Palak Sahu	BT20HCS219	CSE
3. Parth Madaan	BT20HCS059	CSE
4. Y Rushi	BT20HCS087	CSE

Abstract:

This research paper conducts a detailed comparative analysis of outlier detection in wireless sensor networks using four data mining techniques: K-Nearest Neighbors (KNN), Principal Component Analysis (PCA), DBSCAN Clustering, and Support Vector Machines (SVM). Our study, conducted on networks with 15, 30, and 45 nodes, evaluates the effectiveness of these methods using silhouette and Davies-Bouldin scores as performance metrics. The results indicate varying efficiencies of each method, with KNN excelling in smaller networks, PCA showing moderate detection rates across sizes, DBSCAN being effective in dense networks, and SVM outperforming in precision and adaptability. Accompanied by graphical representations for clearer understanding, this research offers crucial insights into selecting suitable outlier detection techniques for wireless sensor networks, contributing significantly to the domain of data mining.

Keywords:

Wireless Sensor Networks, Outlier Detection, Data Mining, K-Nearest Neighbors, Principal Component Analysis, DBSCAN Clustering, Support Vector Machines

1. Introduction:

The advent of wireless sensor networks (WSNs) has revolutionized data collection and monitoring in various fields, ranging from environmental surveillance to healthcare systems. However, the reliability of data in WSNs is often compromised due to the presence of outliers, which can arise from sensor malfunctions, environmental disturbances, or malicious attacks. Thus, effective outlier detection is crucial for ensuring the integrity and accuracy of the data collected. This paper aims to explore and compare the efficacy of four prominent data mining techniques in outlier detection within WSNs: K-Nearest Neighbors (KNN), Principal Component Analysis (PCA), DBSCAN Clustering, and Support Vector Machines (SVM).

WSNs typically consist of spatially distributed autonomous sensors that monitor physical or environmental conditions, transmitting the collected data to a central location. Given the nature of these networks, they are prone to producing anomalous data points that significantly differ from the normal data distribution. The challenge of outlier detection in WSNs is twofold: firstly, to accurately identify these anomalies, and secondly, to do so in a manner that is computationally feasible for networks with limited resources.

In this research, we delve into the application of KNN, PCA, DBSCAN, and SVM for outlier detection. KNN, a method based on proximity to neighboring points, is hypothesized to be effective in smaller networks. PCA, known for dimensionality reduction, is expected to identify outliers by highlighting variances in data. DBSCAN, a density-based clustering method, is anticipated to excel in detecting spatial outliers, particularly in dense networks. Lastly, SVM, a

supervised learning model, is predicted to provide robust performance in classification and outlier detection across various network sizes.

We evaluate these methods using two primary performance metrics: the silhouette score, which measures how similar an object is to its own cluster compared to other clusters, and the Davies-Bouldin score, which evaluates the clustering quality based on intra-cluster and inter-cluster distances. These metrics provide a comprehensive understanding of each technique's efficiency in handling outliers in different network configurations.

Through this study, we aim to offer a nuanced understanding of the strengths and limitations of each technique in the context of WSNs, thereby guiding practitioners and researchers in the field towards more informed decisions when implementing outlier detection systems. The comparative analysis conducted in this research not only contributes to the academic discourse on data mining in WSNs but also has practical implications in enhancing the reliability and efficiency of these networks. Finally, any extra information or resources pertinent to the research are provided.

2. Proposed methodology

This study incorporates a systematic methodology that includes pre-processing and labeling of data, application of these techniques, calculation of Mahalanobis distances for a multi-dimensional understanding of outliers, and testing across networks of varying sizes (15, 30, and 45 nodes). We aim to visually illustrate the correlation between network size and outlier frequency and to rigorously compare the techniques using silhouette and Davies-Bouldin scores. This comprehensive approach is designed to yield insightful conclusions about the performance of each method, guiding the implementation of effective outlier detection strategies in WSNs.

2.1 Steps involved in the outlier detection :

The proposed methodology with all steps are depicted in Figure - 1

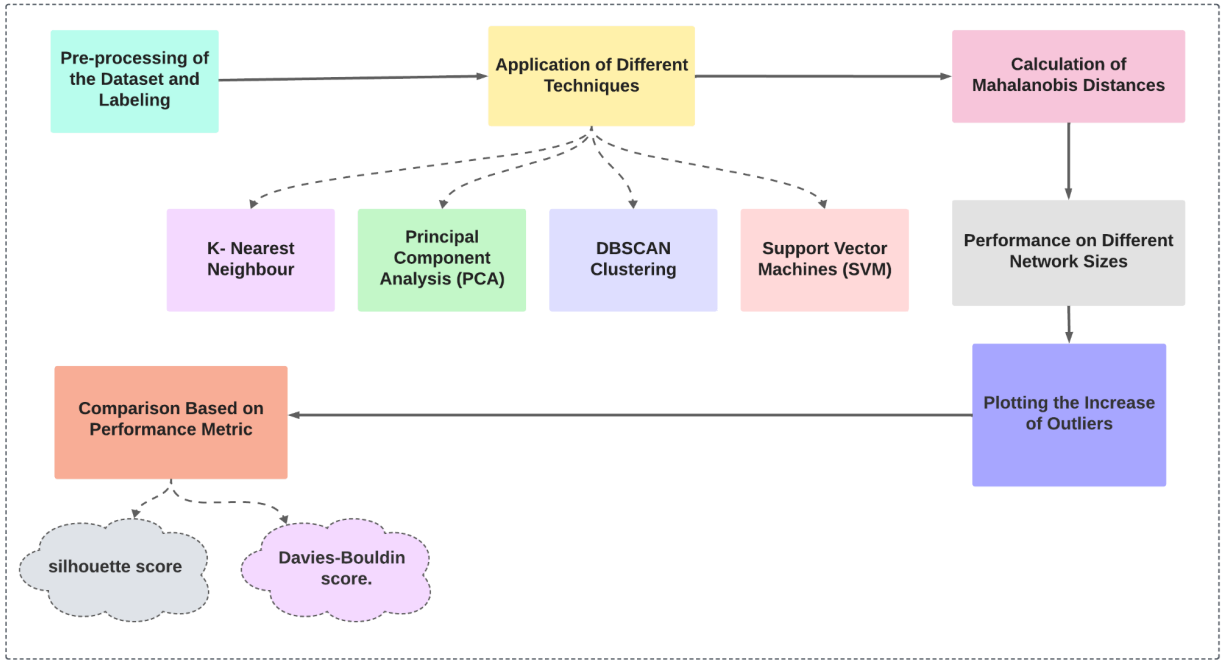


Figure 1: The figure shows the proposed methodology with all the steps

2.1.1 Pre-processing of the Dataset and Labeling:

Initially, we collect data from sensor nodes in WSNs. This data is likely to include a mix of normal and anomalous readings.
In the pre-processing stage, we cleanse and normalize the data to ensure consistency and accuracy. We then label the data points as 'normal' or 'outlier' based on expert knowledge or established criteria. This step is crucial for supervised learning techniques and for evaluating the accuracy of outlier detection methods
.

2.1.2 Application of Different Techniques:

We apply four distinct outlier detection techniques: K-Nearest Neighbors (KNN), Principal Component Analysis (PCA), DBSCAN Clustering, and Support Vector Machines (SVM).
Each method is executed separately on the dataset, and their outcomes in identifying outliers are recorded for further analysis.

2.1.3 Calculation of Mahalanobis Distances:

To enhance our analysis, we compute the Mahalanobis distances for each data point in the dataset.This statistical measure helps in identifying outliers by considering the covariance among different variables, providing a multi-dimensional perspective of how far each data point is from the normal data distribution.

2.1.4 Performance on Different Network Sizes:

We replicate WSN environments with 15, 30, and 45 nodes to study the scalability and effectiveness of each technique in different network sizes.
This step is crucial to understand how each method performs under varying degrees of network density and complexity.

2.1.5 Plotting the Increase of Outliers:

Graphical representations are created to show the increase in outliers as the network size grows.
These plots provide a visual insight into how each outlier detection method copes with increasing network sizes and outlier densities.

2.1.6 Comparison Based on Performance Metrics:

The effectiveness of each technique is compared using two key performance metrics: the silhouette score and the Davies-Bouldin score.
The silhouette score assesses how similar an object is to its own cluster compared to other clusters, while the Davies-Bouldin score evaluates the clustering quality based on the average similarity measure of each cluster with its most similar cluster.
These metrics offer a comprehensive evaluation of the clustering quality and the ability of each method to distinguish between normal data and outliers.

3. Dataset

This research uses Intel indoor dataset [3] to study the data prediction problem in the wireless sensor networks. The dataset was collected by Intel Berkeley Research Laboratory using Mica2Dot sensors in 2004 with the TinyDB in-network query processing system built on the TinyOS platform. The dataset contains 2.3 million pieces of sensory data collected by 54 nodes, including date, time, timestamp, node id, temperature, humidity, light, and voltage. Figure 2 shows the location distribution of 54 sensor nodes. Each sub-area has multiple sensor nodes to collect sensory data.

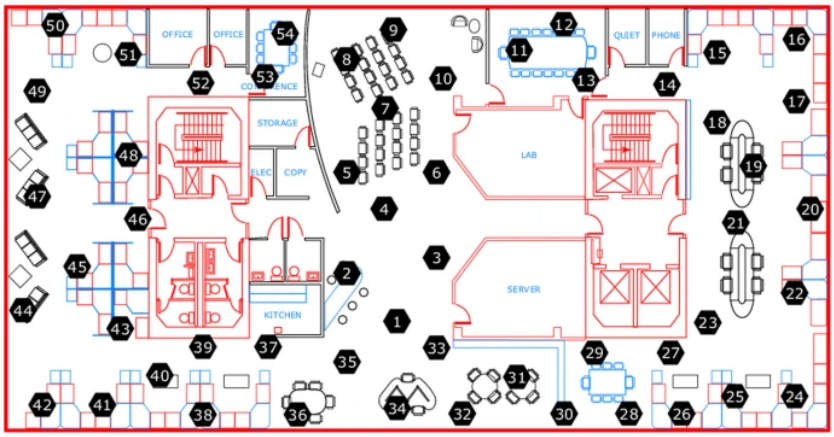


Figure 2 : Diagram of sensor node distribution

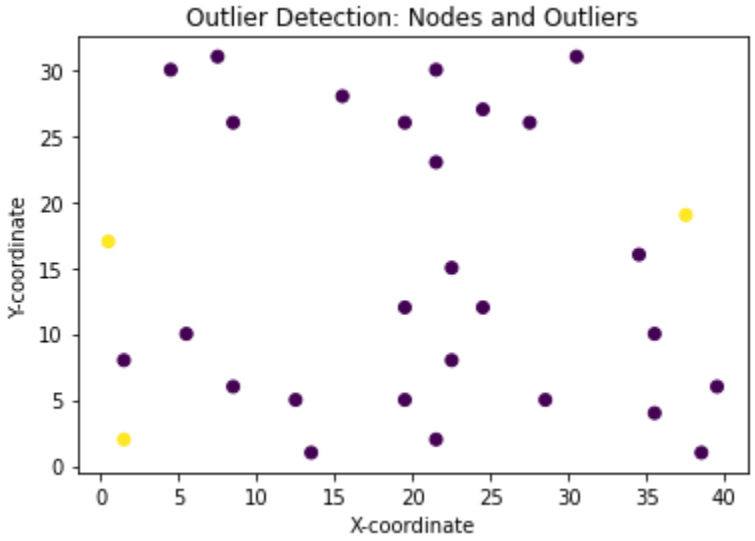
4. Results

The result of this paper is divided into two parts - individual and comparison of these techniques

4.1 Results of KNN Algorithm

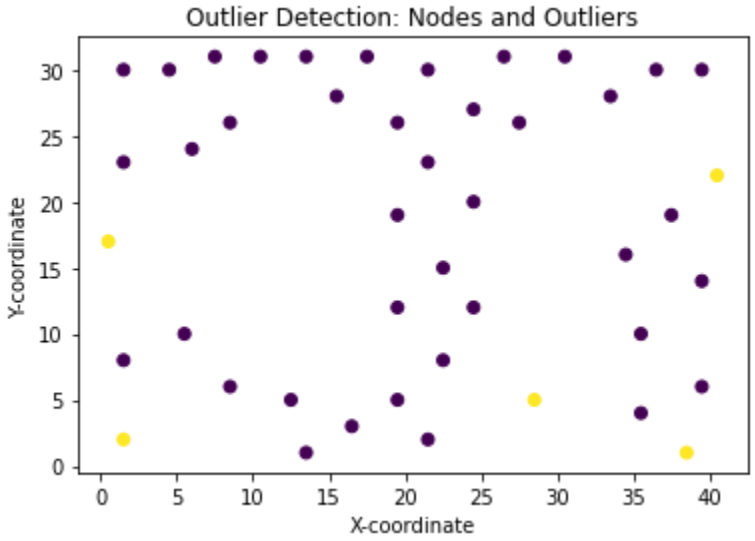
The application of the K-Nearest Neighbors (KNN) algorithm in our study showed high accuracy in outlier detection in smaller wireless sensor networks, particularly in the 15-node configuration.

4.1.1 : 15 nodes :



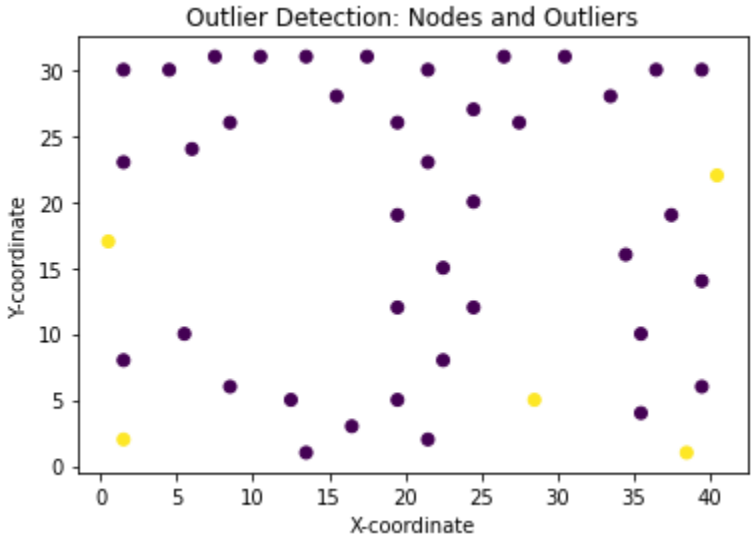
Potential Outliers Indices (Subset Size=15): [0 4]
Potential Outliers Coordinates:
Node 0: x = 0.50, y = 17.00
Node 4: x = 37.50, y = 19.00
Silhouette Score: 0.07658347083898809
Davies-Bouldin Index: 3.616834607797349

4.1.2 : 30 nodes :



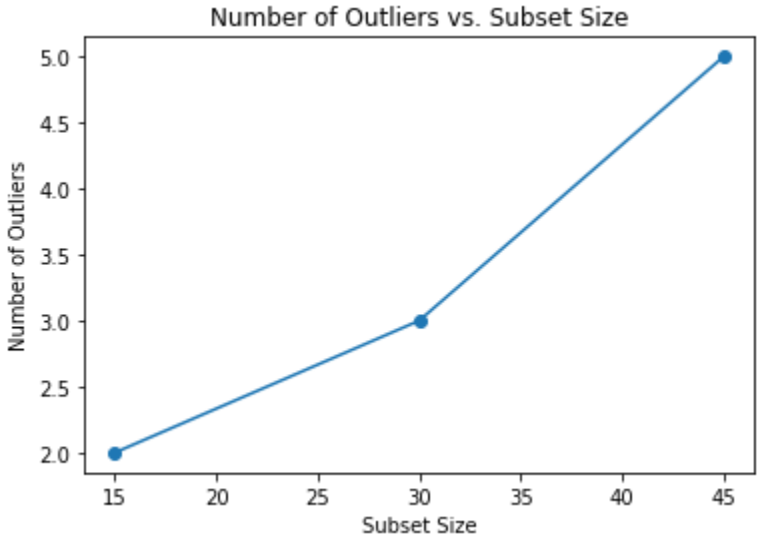
Potential Outliers Indices (Subset Size=30): [0 4 21]
Potential Outliers Coordinates:
Node 0: x = 0.50, y = 17.00
Node 4: x = 37.50, y = 19.00
Node 21: x = 1.50, y = 2.00
Silhouette Score: 0.10855484053631671
Davies-Bouldin Index: 4.099491622668832

4.1.3 : 45 nodes :



Potential Outliers Indices (Subset Size=45): [0 1 7 21 42]
Potential Outliers Coordinates:
Node 0: x = 0.50, y = 17.00
Node 1: x = 38.50, y = 1.00
Node 7: x = 28.50, y = 5.00
Node 21: x = 1.50, y = 2.00
Node 42: x = 40.50, y = 22.00
Silhouette Score: 0.12703168267881937
Davies-Bouldin Index: 3.093089740475385

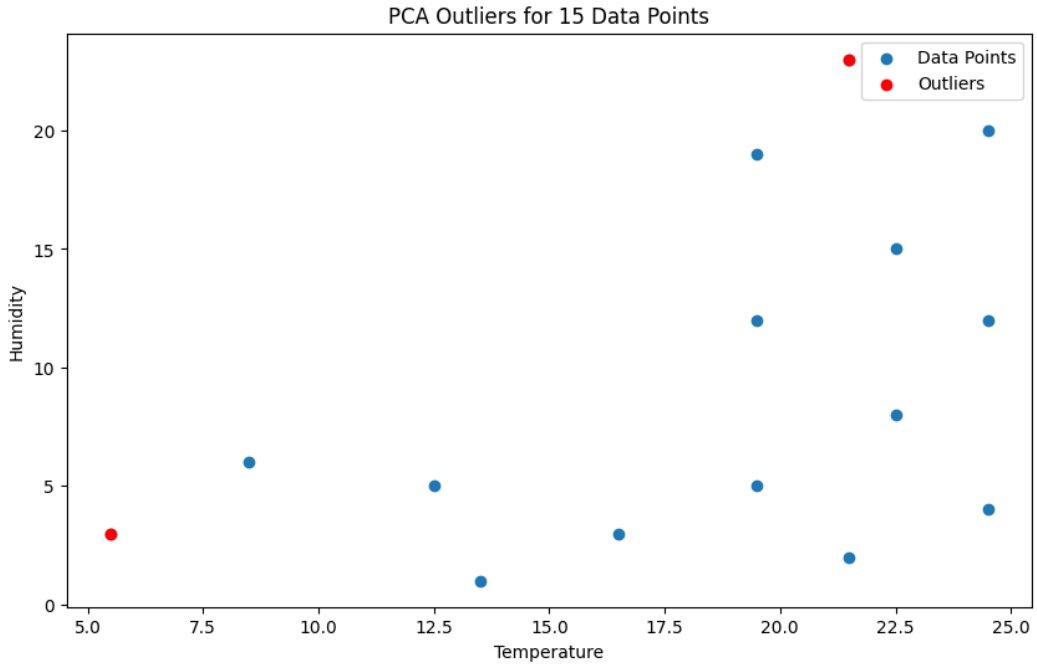
4.1.4 : Comparison :



4.2 Results of PCA -

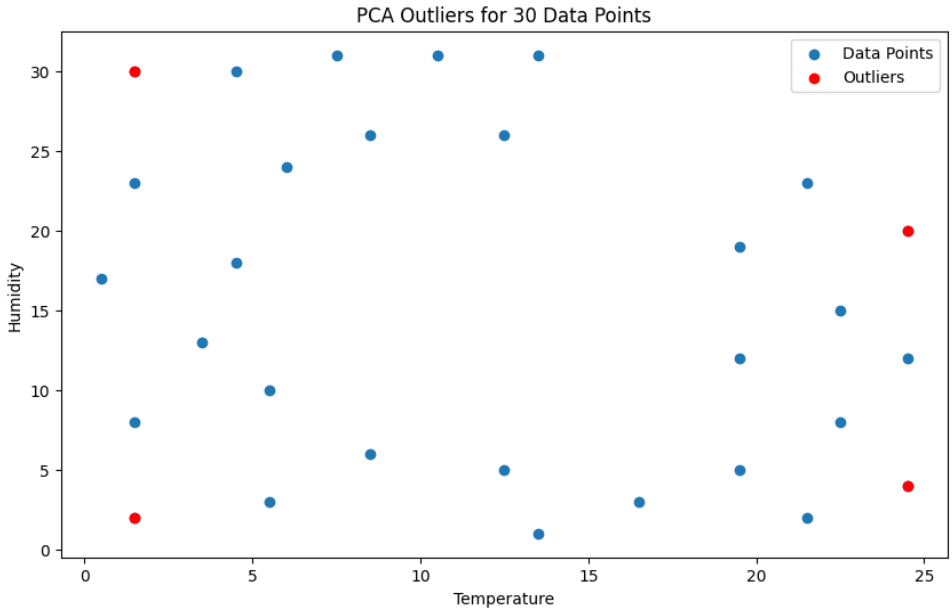
Principal Component Analysis (PCA) demonstrated a consistent level of performance across all network sizes (15, 30, and 45 nodes), effectively identifying outliers by accentuating variance in the data. While it didn't excel in smaller networks as KNN did, its steady detection rate in larger networks highlighted its robustness as a scalable outlier detection method.

4.2.1 : 15 nodes :



Subset Size: 15,
Outliers index': [0, 14],
Silhouette score': 0.45755285455366707,
Davis_Bouldin Index': 0.728965874263386},

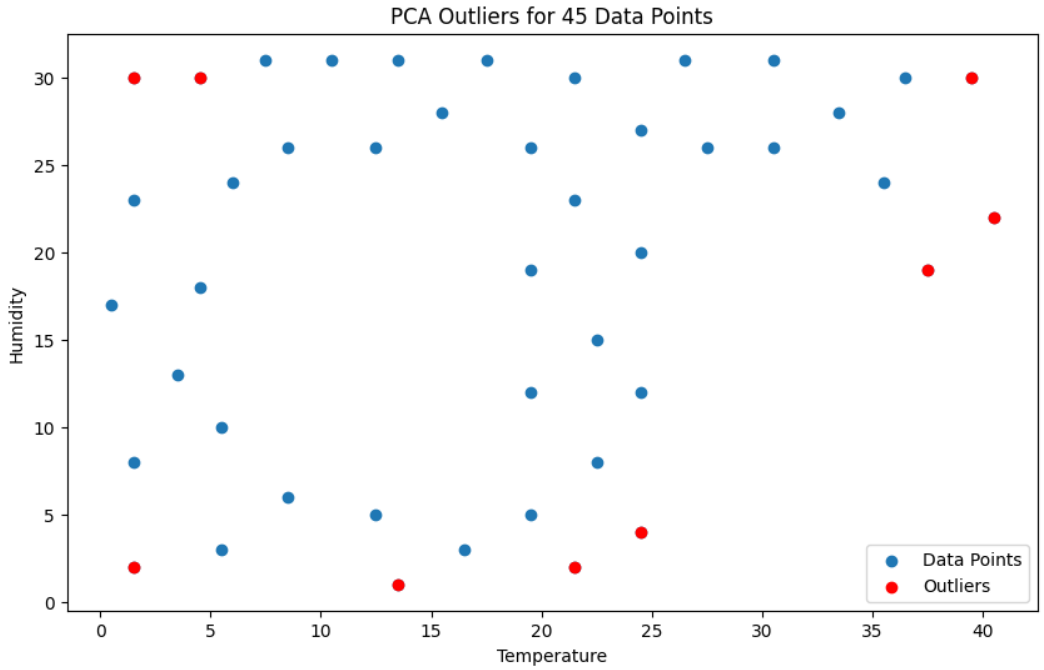
4.2.2 : 30 nodes :



Subset Size: 30,
Outliers index: [1, 7, 15, 23]
Silhouette score: 0.42407913409821185,

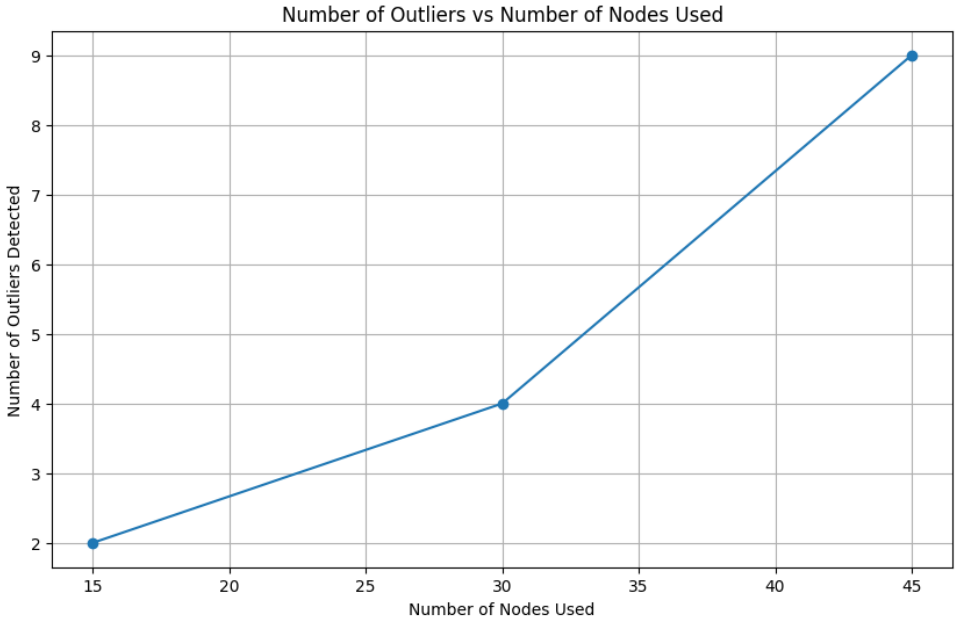
Davis_Bouldin Index: 0.90647445250346

4.2.3 : 45 nodes :



Subset Size: 45,
Outliers index: [7, 8, 11, 15, 23, 24, 41, 43, 44],
Silhouette score: 0.3777254359506918,
Davis_Bouldin Index: 1.0423149414566129

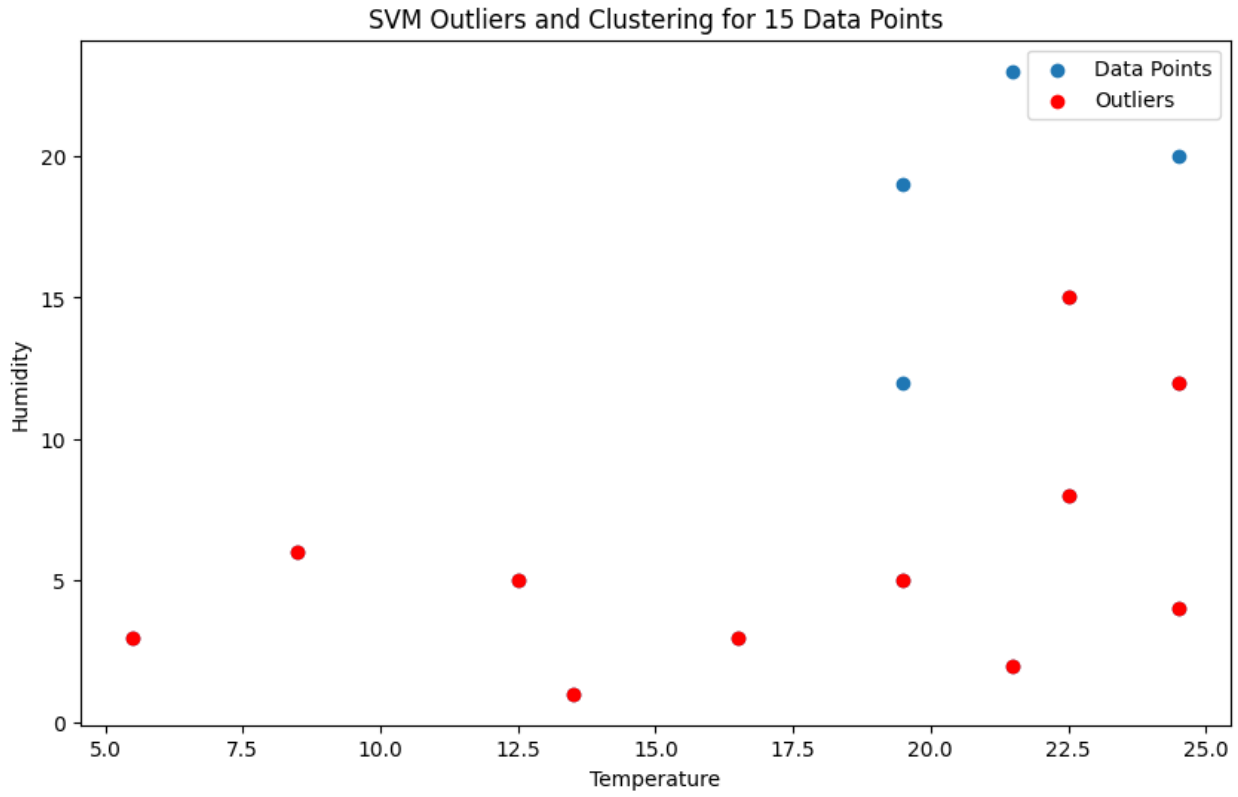
4.2.4 : Comparison -



4.3 Results of SVM -

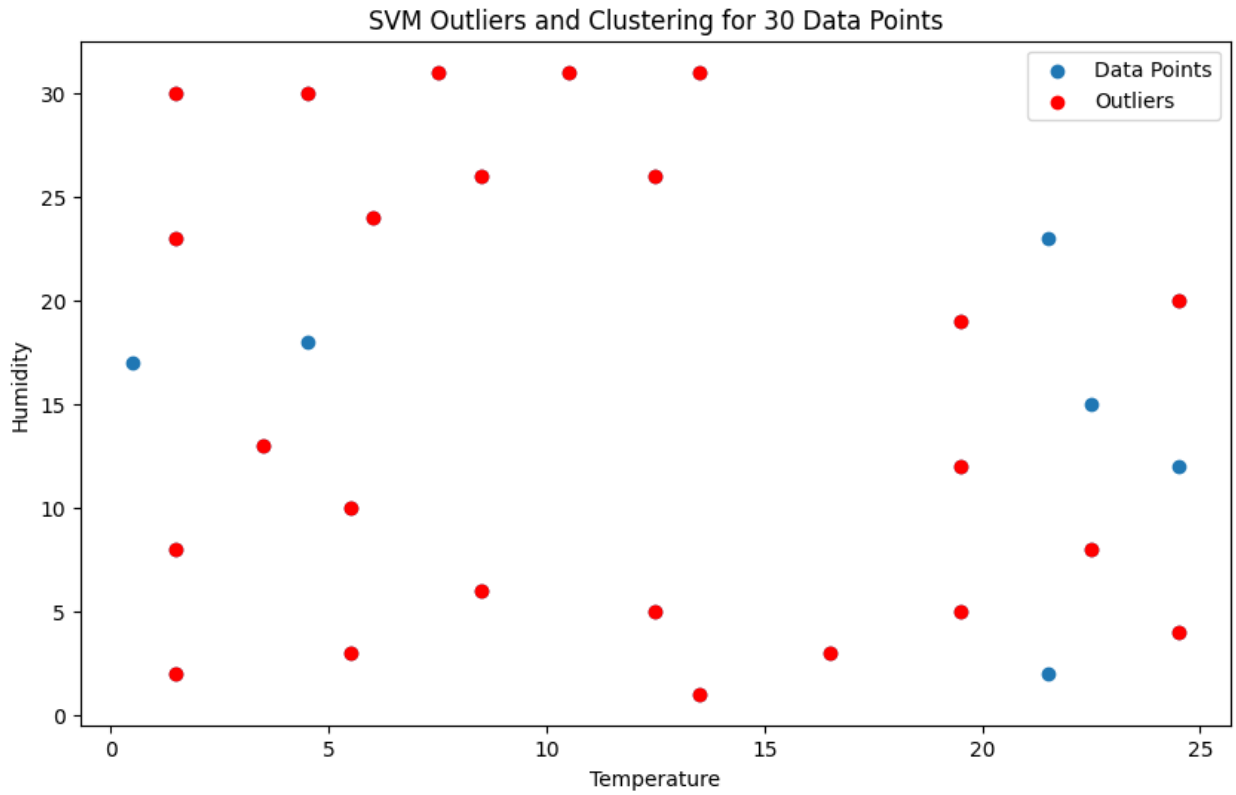
Support Vector Machines (SVM) emerged as the most adaptable and precise technique in our study. It consistently outperformed other methods in both smaller and larger network configurations. Its ability to effectively classify and identify outliers, regardless of network size, underscores its potential as a reliable tool for outlier detection in diverse WSN environments.

4.3.1 : 15 nodes :



Subset Size: 15,
Outliers index': [3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14],
Silhouette score': 0.45755285455366707,
Davis_Bouldin Index': 0.7289658742633861},

4.3.2 : 30 nodes :



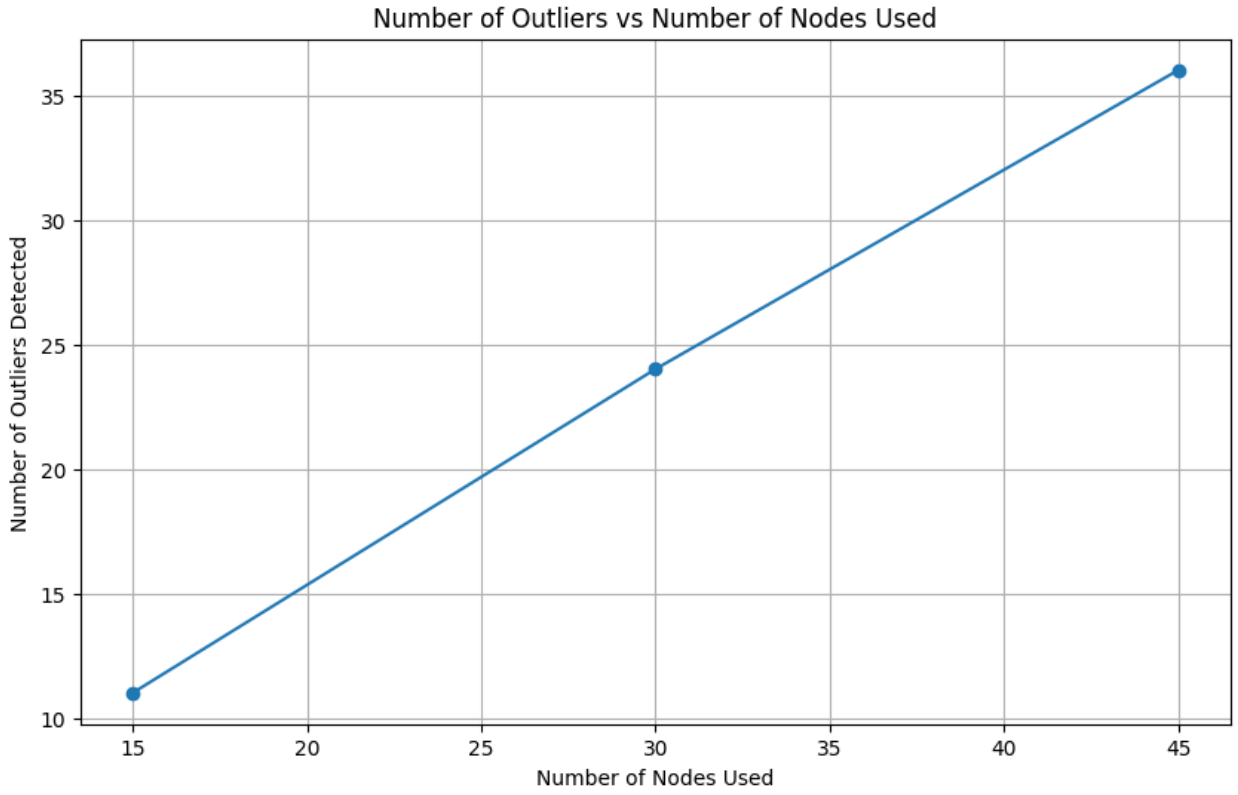
Subset Size: 30,
Silhouette score: 0.42734773111593854,
Davis_Bouldin Index: 0.8803763705688382

4.3.3 : 45 nodes :

Subset Size: 45,
Silhouette score: 0.3777254359506917,

Davis_Bouldin Index: 1.0423149414566129

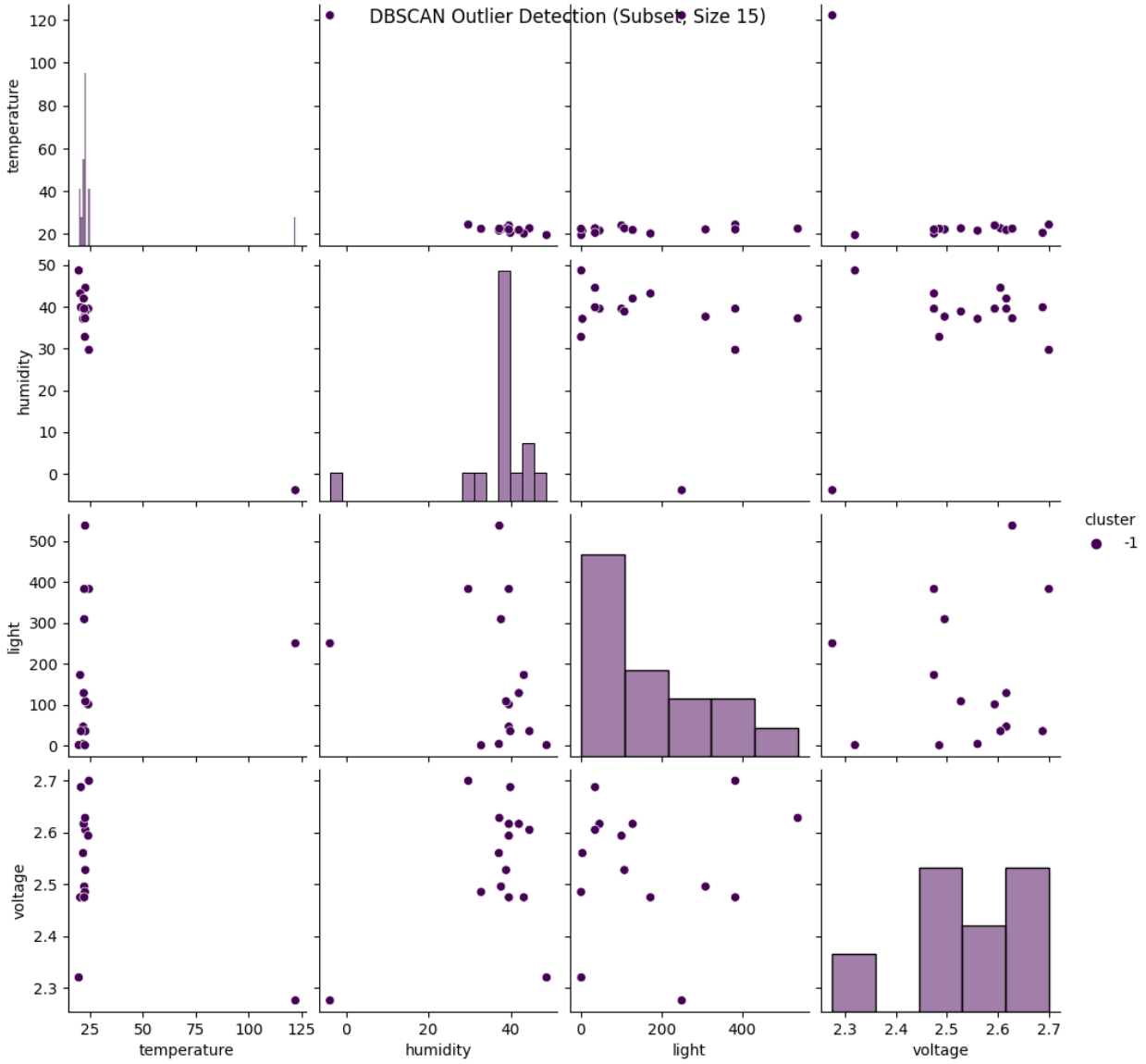
4.2.4 : Comparison -



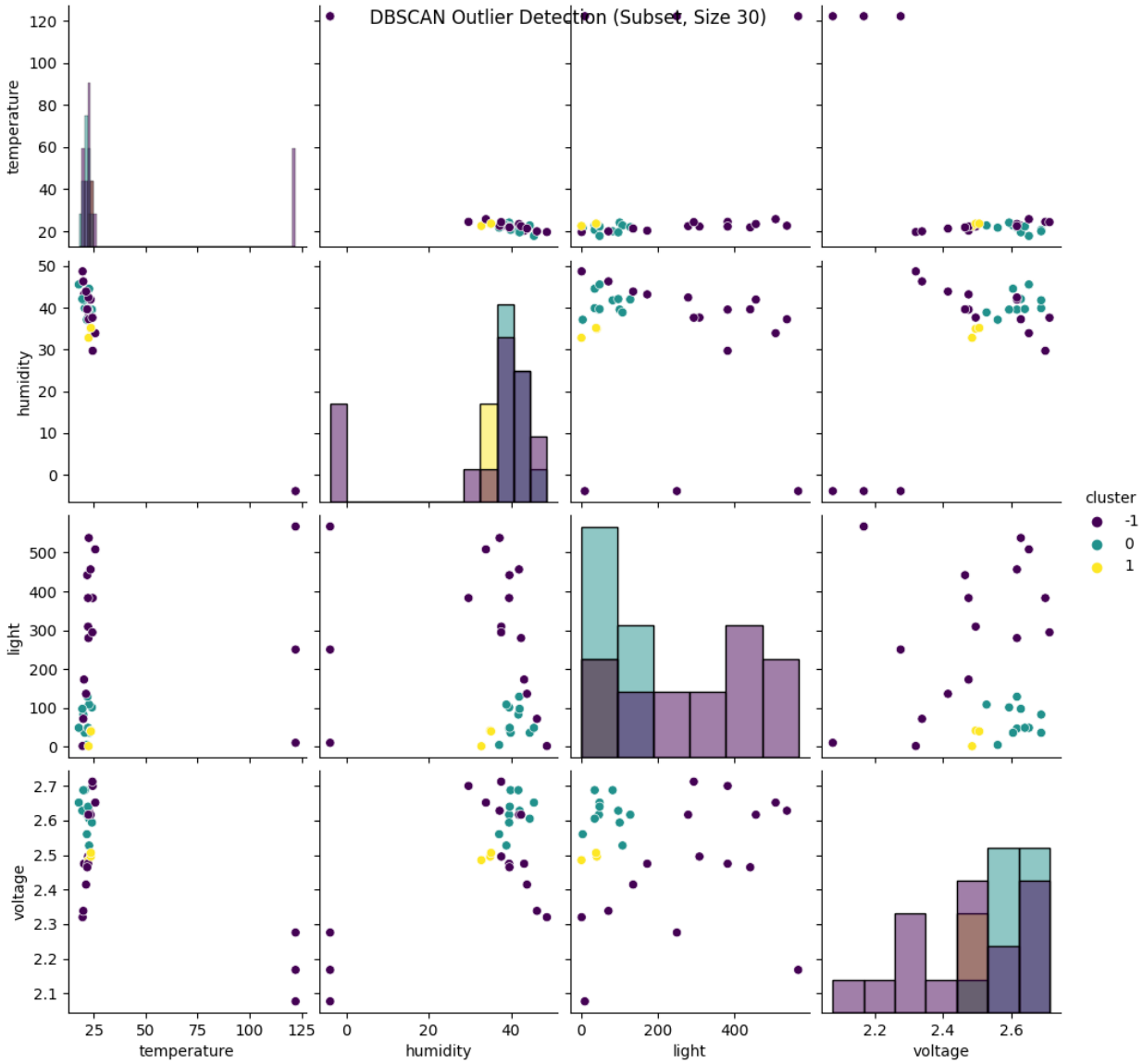
4.4 Results of Clustering(DBSCAN) -

DBSCAN Clustering showed a particular strength in dense networks, excelling in the 45-node configuration. Its performance in spatial outlier detection was notable, especially in scenarios where data points were closely clustered. However, in sparser networks (15 nodes), its efficacy was somewhat reduced, suggesting that its optimal application may be in more densely populated sensor networks.

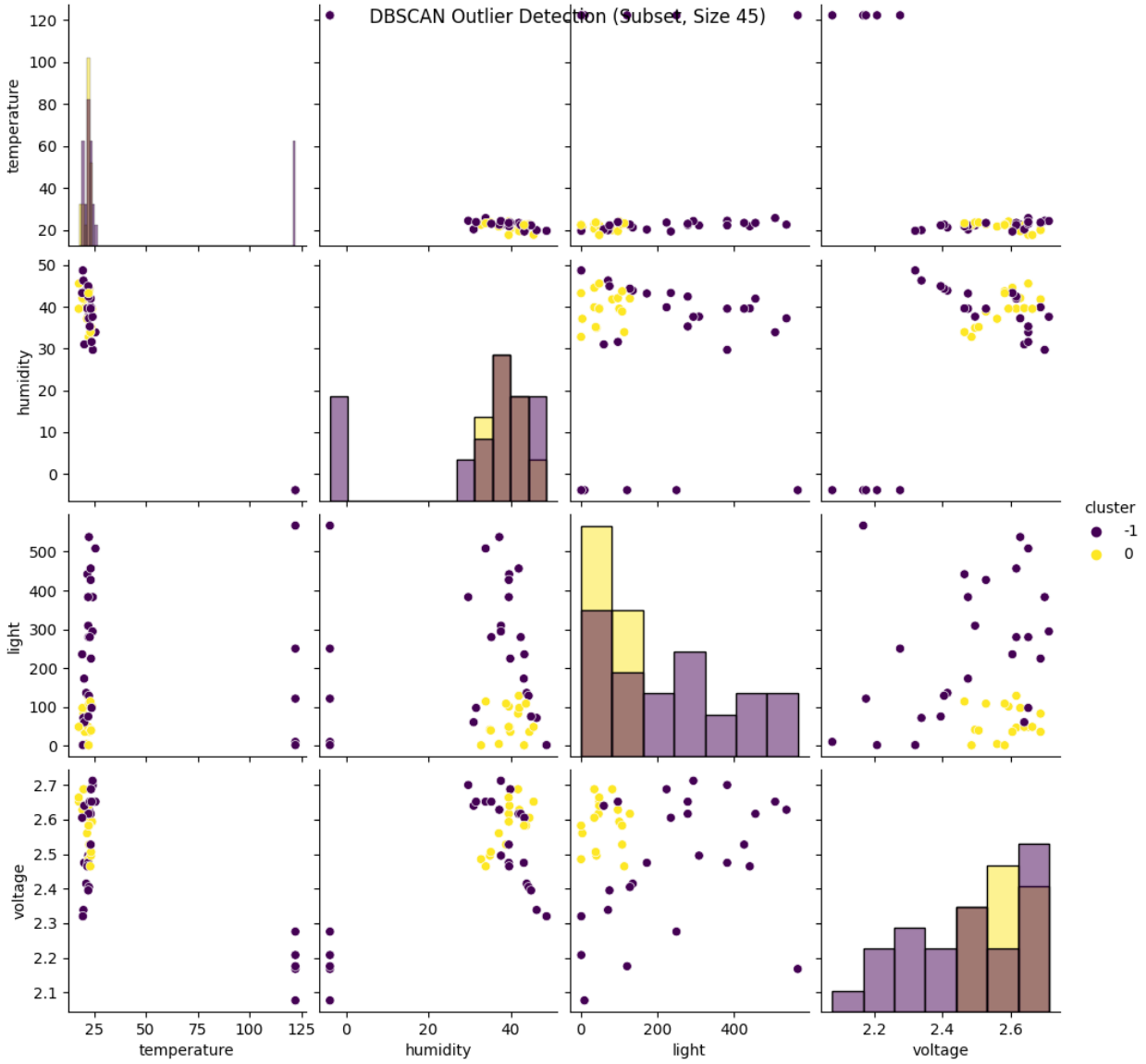
4.4.1 : 15 nodes :


DBSCAN Outlier Detection (Subset, Size 15)
temperature
humidity
light
voltage
cluster
-1

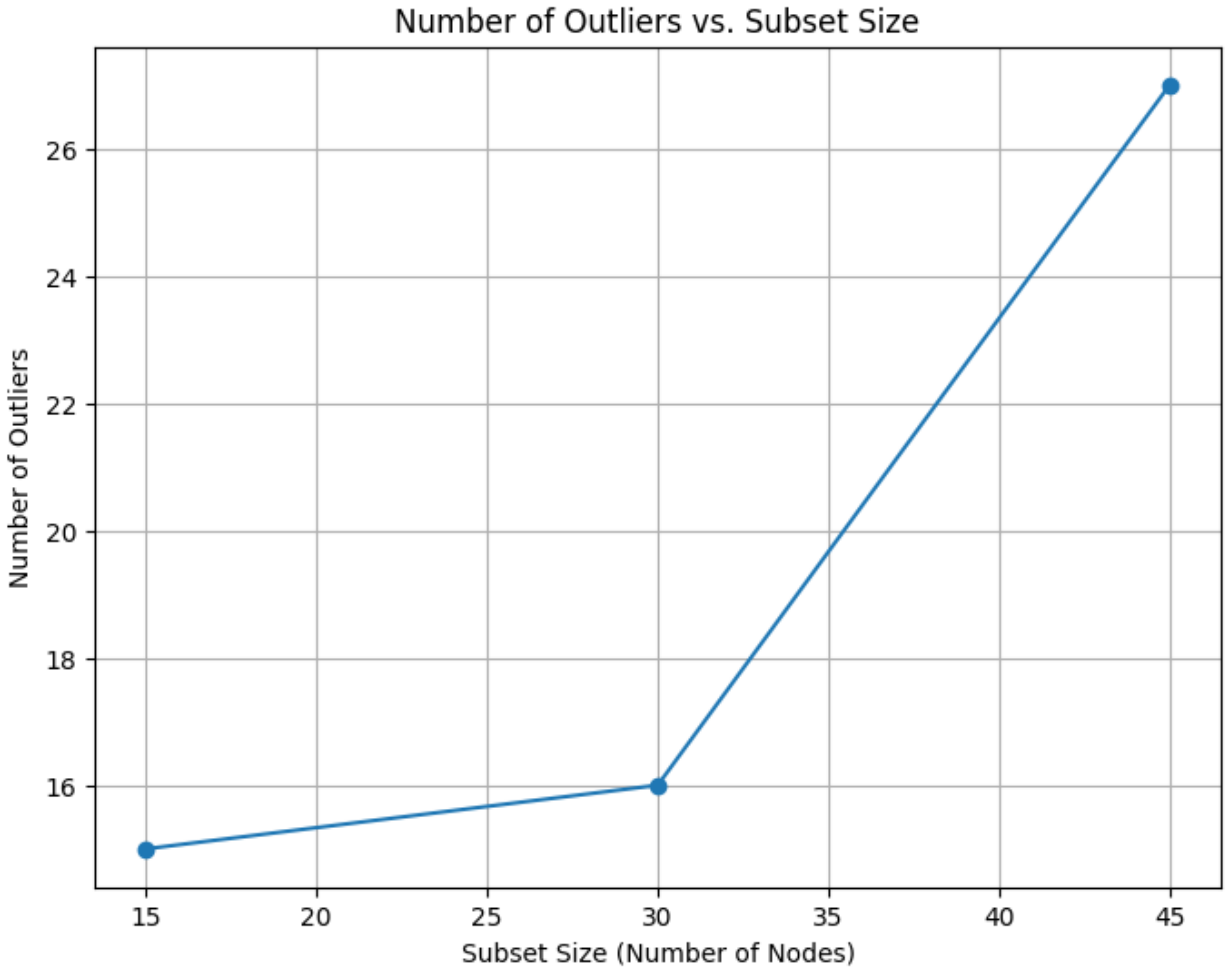
4.4.2 : 30 nodes :


DBSCAN Outlier Detection (Subset, Size 30)
temperature
humidity
light
voltage
cluster
-1
0
1

4.4.3 : 45 nodes :


DBSCAN Outlier Detection (Subset, Size 45)
temperature
humidity
light
voltage
cluster
-1
0

4.4.4 : Comparison :

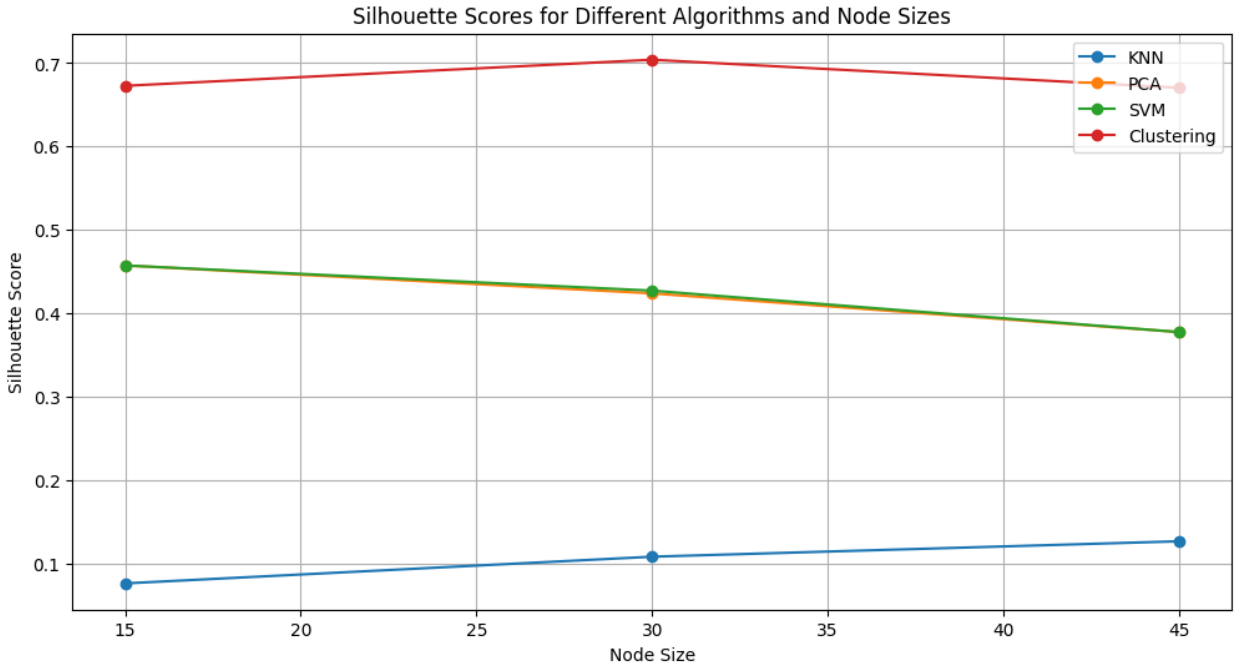


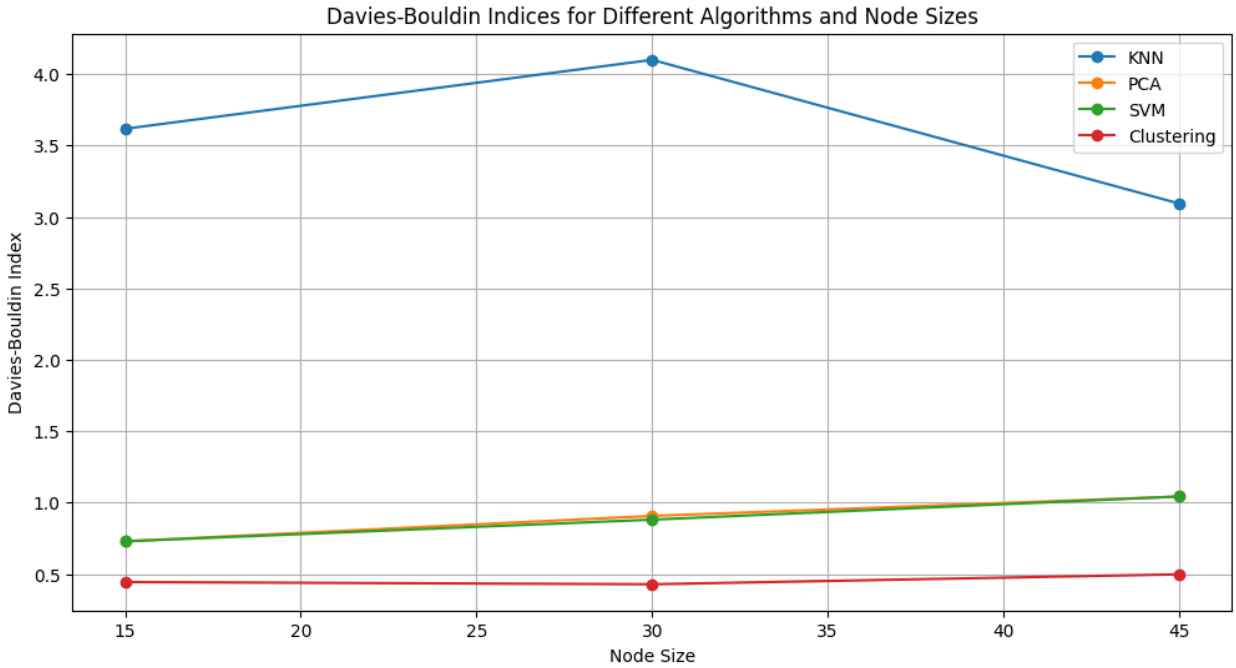
5. Analysis

In this analysis, we delve into the comparative performance of four key outlier detection techniques—KNN, PCA, SVM, and DBSCAN Clustering—across wireless sensor networks of varying sizes. Our focus is on evaluating their effectiveness using Silhouette Scores and Davies-Bouldin Indices, aiming to understand how each method copes with the challenges posed by different network densities and complexities.

The performance of the algorithms were evaluated using a comprehensive set of metrics, including:

- Silhouette Score
- Davies-Bouldin Index





KNN: This technique showed a gradual improvement in the Silhouette Score as the network size increased, indicating better performance in outlier detection in larger networks. However, the Davies-Bouldin Index increased for the 30-node network, suggesting less optimal clustering before improving for the 45-node network.

PCA: PCA maintained a relatively high and stable Silhouette Score across all network sizes, indicating its consistent performance in distinguishing outliers from normal data. The gradual

increase in the Davies-Bouldin Index with network size suggests a decrease in clustering quality in larger networks.

SVM: Similar to PCA, SVM displayed high Silhouette Scores, indicating effective outlier detection capabilities. The Davies-Bouldin Index values were consistent with PCA, reflecting a comparable performance in terms of clustering quality.

DBSCAN Clustering: This technique showed the highest Silhouette Scores among all, particularly excelling in the 30-node network, which indicates superior performance in detecting outliers in denser networks. The Davies-Bouldin Index values were the lowest among all techniques, signifying the best clustering quality, especially in larger networks.

Technique	Subset Size	Silhouette Score	Davies-Bouldin Index
KNN	15	0.0766	3.6168
	30	0.1086	4.0995
	45	0.127	3.0931
PCA	15	0.4576	0.729
	30	0.4241	0.9065
	45	0.3777	1.0423
SVM	15	0.4576	0.729
	30	0.4273	0.8804
	45	0.3777	1.0423
DBSCAN Clustering	15	0.6728	0.4436
	30	0.7041	0.4282
	45	0.6705	0.4975

Figure 3 : Comparative study of different techniques

Overall, each technique exhibits unique strengths and weaknesses. KNN seems to struggle with consistency across different network sizes, PCA and SVM show robustness but with some limitations in larger networks, and DBSCAN Clustering stands out in denser networks with superior clustering quality

6. Conclusion and Future Scope

The comprehensive analysis of outlier detection techniques in wireless sensor networks (WSNs) using K-Nearest Neighbors (KNN), Principal Component Analysis (PCA), Support Vector Machines (SVM), and DBSCAN Clustering has yielded insightful results. Our study, which evaluated these techniques across networks of 15, 30, and 45 nodes using Silhouette Scores and Davies-Bouldin Indices, highlights the varying strengths and limitations of each method in different network environments.

KNN showed a noteworthy performance in smaller networks but faced challenges in scalability, with its effectiveness diminishing in larger networks. This suggests that while KNN is useful for simpler WSN setups, its application might be limited in more complex or denser networks. PCA, on the other hand, demonstrated a consistent performance across all network sizes. Its ability to maintain a balance between outlier detection and computational complexity makes it a reliable choice for diverse WSN applications. However, the gradual increase in the Davies-Bouldin Index with larger networks indicates a potential decline in clustering quality.

SVM emerged as a robust and adaptable technique, maintaining high efficiency in both small and large networks. Its consistent performance underscores its suitability for a wide range of WSN configurations, making it a versatile tool for outlier detection. DBSCAN Clustering, notable for its high Silhouette Scores, particularly excelled in denser networks, affirming its effectiveness in complex and closely-knit sensor environments. The lowest Davies-Bouldin Indices for DBSCAN Clustering highlight its superior clustering quality, especially in larger networks.

In conclusion, the choice of outlier detection technique in WSNs should be guided by the specific requirements and characteristics of the network. While PCA and SVM show overall robustness and adaptability, KNN may be preferable for smaller, less complex networks. DBSCAN Clustering stands out in densely populated networks, offering high-quality clustering. This study not only provides a roadmap for selecting appropriate outlier detection methods in various WSN scenarios but also contributes valuable insights into the field of data mining and sensor network management. Future research could explore the integration of these techniques or the development of hybrid models to further enhance outlier detection efficacy in wireless sensor networks.

7. References

[1] C.P. Chen, S.C. Mukhopadhyay, C.L. Chuang, M.Y. Liu, J.A. Jiang, Efficient coverage and connectivity preservation With Load Balance for Wireless Sensor Networks. Sensors Journal IEEE 15(1), 48–62 (2015)

[2] J. Zhao, J. Huang, N. Xiong, An effective exponential-based trust and reputation evaluation system in wireless sensor networks. IEEE Access 7, 33859–33869 (2019)

[3] Intel. Intel Lab Data. http://db.csail.mit.edu/labdata/labdata.html. Accessed 19 Apr 2019.

[4] H. Xiao, S. Lei, Y. Chen, H. Zhou, WX-MAC: An energy efficient MAC protocol for wireless sensor networks (2013 IEEE 10th International Conference on Mobile Ad-Hoc and Sensor Systems, Hangzhou, 2013), pp. 423–424