Recent Research Trends of AI in Various Streams of Engineering and Management
May-2025
ISBN Number: 978-93-95305-47-1

A REVIEW ON OUTLIER DETECTION TECHNIQUES: TECHNIQUES AND APPLICATION

Km. Pooja Yadav
Research Scholar
Department of Computer Science and Engineering,
Bansal Institute of Engineering & Technology, Lucknow - India
Email: poojay1799@gmail.com
Dr. C.L.P. Gupta
Professor
Department of Computer Science and Engineering,
Bansal Institute of Engineering & Technology, Lucknow
Email: clpgupta@gmail.com
https://orcid.org/0000-0002-7202-3872

ABSTRACT

Once thought of as noisy data in statistics, outliers have emerged as a significant issue that is being studied in a variety of application areas and research sectors. While some methods are more general, several outlier detection strategies have been developed specifically for particular application domains. Certain application sectors, like studies on crime and terrorism, are being studied under stringent anonymity. Such methods' methods and outcomes are not easily available. Outlier detection methods are covered in great detail in a number of publications, research and review articles, surveys, and machine learning and statistical areas. Our goal in this study is to provide a systematic and general overview of different outlier detection approaches. Through this activity, we intend to gain a better grasp of the various paths that outlier analysis research is going in, both for ourselves and for newcomers to the field who will be able to pick up the ties to various areas of applications in depth.

Keywords: Event detection, Outlier detection, QoS, Detection efficiency

1. Introduction:

Outlier detection involves recognizing patterns in data that deviate from expected behavior. It is widely applied across various domains, including fault detection in safety-critical equipment, credit card fraud detection, and fraud identification in insurance or healthcare sectors, military monitoring for opponent actions, and cyber security intrusion detection. They are important in data because they can be transformed

into information that is helpful for a variety of applications. A compromised machine may be transferring private information to an unapproved location if there is an unusual traffic pattern in the network [1]. Malignant tumours may be present in an aberrant MRI image [2]. A malfunction in a spacecraft component could be indicated by anomalous readings from a sensor, or credit card or identity fraud could be indicated by outliers in credit card transaction data [3–4]. Outliers in statistical data have been studied since the 1800s [5]. Since then, various research groups have developed a diverse range of outlier identification techniques, some tailored for specific applications and others designed to be more general-purpose. These methods continue to evolve, incorporating advancements in machine learning, statistical modeling, and data mining to enhance their accuracy and effectiveness. We hope to learn more about the different directions that outlier analysis research is taking as a result of this effort. Additionally, even if those applications were not initially planned, we want to take into account modifying methods from other disciplines to our areas of interest, which are crime detection and counterterrorism.

Outliers are patterns of data that substantially depart from a well-defined notion of typical behaviour. Outliers in a straightforward 2-dimensional data set are shown in Figure 1. Since the majority of observations are located in these two locations, the data has two normal regions, N1 and N2. Points o1 and o2, as well as points in area O3, are examples of outliers. They all differ significantly from the regions x y N1 N2 o1 o2 O3, yet they all have one thing in common: the analyst finds them all fascinating. The "interestingness" or applicability of outliers to actual circumstances is a crucial component of outlier detection. Although they both deal with unwanted noise in the data, outlier identification is not the same as noise reduction [6] and noise accommodation [7]. Noise is a phenomenon in data that obstructs data analysis even though it is not of interest to the analyst. Noise reduction is driven by the need to remove unwanted components from the data before performing any data analysis. The term "noise accommodation" describes how to protect a statistical model estimate against unusual observations [8]. Novelty detection [9,10,11] is a subject connected to outlier identification that looks for patterns in data that have never been seen before (emergent, novel), such as a fresh conversation topic in a news group. Novel patterns are distinguished from outliers by the fact that, once identified, they are usually integrated into the standard model. Novelty detection is a subject connected to outlier detection [9]. The key difference between novel patterns and outliers is that novel patterns represent previously unrecognized trends, which, once identified—such as a new discussion topic emerging in a news group—integrate into

the regular model over time. In contrast, outliers remain anomalies, deviating significantly from expected behavior without becoming part of the established pattern.

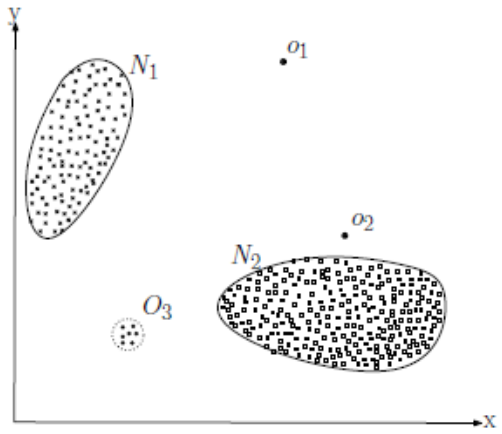

Fig. 1. An elementary illustration of outliers in a two-dimensional data set.

Outlier detection, also referred to as anomaly detection, is a vital component of data analysis focused on identifying data points that significantly deviate from expected patterns. With the rapid expansion of big data across diverse fields, the need for robust and efficient outlier detection techniques has become more critical than ever. This review delves into various methodologies, examining their underlying principles, real-world applications, strengths, and limitations, while also highlighting emerging trends and future research directions in the field.

A. Statistical Methods

Statistical approaches are grounded in probability theory. Techniques such as z-score, modified z-score, and Grubbs' test are popular for detecting outliers in normally distributed data. These methods are straightforward and easy to implement. However, they often assume a specific distribution, which can lead to inaccuracies in non-Gaussian datasets.

B. Machine Learning Approaches

Machine learning techniques have gained traction due to their ability to handle large datasets and complex relationships. Common methods include:

ISBN Number: 978-93-95305-47-1

- **Isolation Forest:** This machine learning algorithm detects outliers by randomly selecting a feature and a split value to isolate anomalies from normal data points. By constructing multiple decision trees, it efficiently separates outliers in fewer steps than regular data points. Its scalability, speed, and effectiveness make it particularly well-suited for high-dimensional datasets and large-scale anomaly detection tasks.
- **One-Class SVM:** This method is effective in high-dimensional spaces but can be sensitive to the choice of kernel and parameters.
- **Autoencoders:** Used primarily in deep learning, autoencoders can learn a representation of normal data and flag deviations as outliers. They are powerful but require substantial data and computational resources.

C. Distance-Based Methods

Distance-based techniques, such as k-nearest neighbors (KNN), identify outliers by measuring the distance between data points, where those significantly distant from their neighbors are flagged as anomalies. These methods are intuitive and effective for low-dimensional data; however, they face challenges in high-dimensional spaces due to the 'curse of dimensionality,' which can reduce distance-based distinctions and impact detection accuracy. To mitigate this, advanced variations incorporating dimensionality reduction techniques or adaptive distance metrics have been proposed.

D. Clustering-Based Techniques

Clustering methods like DBSCAN and k-means can also identify outliers. DBSCAN, for instance, identifies clusters of varying density, designating points that do not belong to any cluster as outliers. While effective, these methods require careful tuning of parameters and may not perform well on datasets with varying densities.

E. Ensemble Methods

Ensemble techniques, which combine multiple algorithms, can enhance detection accuracy. Methods like Random Cut Forest (RCF) leverage multiple trees to improve robustness against noise and improve detection rates. However, they can be computationally intensive.

F. Applications

Outlier detection techniques find applications across various domains, including fraud detection in finance, network security, fault detection in manufacturing, and healthcare for monitoring patient vitals. Each application requires careful consideration of the method used, balancing accuracy, efficiency, and interpretability.

2. Related Work:

The literature suggests a wide range of pertinent attempts to classify outlier identification techniques. For example, authors in [11] group the approaches by study areas, problem characteristics, and application domains, including machine learning, data mining, statistics, etc. The authors in [17] classify outlier detection techniques into five categories—statistical, closest neighbor, clustering, classification, and spectral decomposition—based on a taxonomy specifically designed for Wireless Sensor Networks (WSNs). Similarly, the authors in [1,22] present a comparable categorization for outlier detection in WSNs but further expand it by incorporating an additional category focused on artificial intelligence-based approaches. Meanwhile, the authors in [23–25] propose a distinct taxonomy for machine learning techniques applied to IoT outlier detection. They classify existing methods into four key groups: hybrid algorithms, dimensionality reduction techniques, clustering-based approaches, and classification models, highlighting the evolving role of machine learning in anomaly detection for IoT systems. The authors of [26] put out a worldwide categorisation study of the advancements in outlier detection methods. They give the performance, advantages, disadvantages, and difficulties that various techniques confront. Nevertheless, it is not tailored for IoTs or WSN; rather, it is a generic survey about outlier identification methods.

The authors in [27] proposed a comprehensive approach for outlier detection in IoT by conducting a systematic study of the dataset through a five-step framework, alongside classifying various outlier detection methodologies. The first step involves establishing a scenario for generating a labeled dataset on a real-world Internet of Things (IoT) system using mathematical modeling. This foundational step ensures a structured dataset, enabling more accurate training and evaluation of anomaly detection techniques. Their approach emphasizes the importance of realistic data representation in improving the

reliability and effectiveness of IoT-based outlier detection methods. The second stage's goal is to intercept data packets and change a portion of them in order to inject anomalous packets. Reintroducing these packets into the network is the third step. Subsequently, the fourth phase involves data sniffing in order to examine, evaluate, and create a model that illustrates the actions of network outliers. In order to assess and identify anomalies, we must lastly use a machine learning algorithm like logistic regression, support vector machines (SVM), linear discriminant analysis (LDA), K closest neighbours (KNN), or tree-based techniques.

The authors in [31] developed a distributed approach for real-time outlier detection in Wireless Sensor Networks (WSNs) using time-series data. Their method leverages spatial-temporal correlation to distinguish between normal data variations, sensor errors, and actual anomalies. Each node employs the Autoregressive and Moving Average (ARMA) prediction model to identify temporal outliers. Subsequently, nodes communicate with their neighbors to verify whether detected outliers also deviate spatially. This technique, known as Temporal and Spatial Real Data-Based Outlier Detection (TSOD), improves detection accuracy but introduces some communication overhead.

In [32], the authors proposed a distributed, online outlier detection approach for hierarchical WSNs using histograms. Unlike TSOD, this method does not require a verification step to confirm outliers. Through a theoretical analysis, they demonstrate that the error margin of their estimation remains minimal. Their approach is computationally efficient, making it well-suited for real-time applications in resource-constrained environments.

Additionally, the authors in [33] introduced an Approximation Adaptive Kernel Density Estimator (AKDE) technique for online outlier detection in data streams. This method applies Kernel Density Estimation (KDE) to compute the probability density function (PDF) of incoming data, enabling real-time anomaly identification with improved adaptability to dynamic data distributions. They provide evidence that their method outperforms KDE. Based on the ARMA model, the authors of [34] created an online adaptive technique that could dynamically detect and replace outliers. Additionally, this approach can meet the healthcare application's real-time radar requirements. In order to simulate them using ARMA, their approach may examine the correlation of neighbourhood data. Their model outperforms SVM and neural networks in terms of modelling and prediction speed.

The authors in [35] proposed an IoT architecture for error and event detection based on four statistical models, emphasizing the crucial role of spatial-temporal correlation. Their approach begins by segmenting the data using the Classification and Regression Trees (CART) model, which partitions the dataset into meaningful subsets. For each partition, a predictive model is then developed to enhance anomaly detection. This process results in a structured decision tree that facilitates classification. Additionally, they incorporate prediction error analysis to refine their model, ensuring higher accuracy in distinguishing between normal data variations, sensor errors, and significant events within IoT systems. The Random Forest (RF) model is then used to produce a number of trees. Another model for classification and regression is the Gradient Boosting Machine (GBM). Lastly, a linear classifier that divides data into classes according to features or parameters is the Lin ear Discriminant Analysis (LDA) model.

Based on a modified K-means algorithm, the authors of [36] presented an unsupervised method for outlier detection and clustering (ODC). A data value is considered an outlier if its average distance is p times greater than its centroid. To improve the clustering process, they then eliminate the data set's outlier values.

The authors in [37] proposed an outlier identification method for the Internet of Things (IoT) by leveraging the k-means clustering algorithm and big data processing techniques. Their approach efficiently handles large-scale distributed data by integrating the Mahout machine learning library, MapReduce, and the Hadoop architecture. To further enhance their framework, they improve the LinkSmart middleware, a key component of the Hydra middleware project, to seamlessly integrate their algorithm within IoT systems [38]. This enhancement ensures better scalability and real-time anomaly detection in massive IoT networks.

Furthermore, the authors in [39] introduced a novel outlier detection algorithm based on the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) technique. This method effectively identifies anomalies in complex datasets by clustering similar data points while isolating noise as potential outliers. Their approach enhances traditional DBSCAN by optimizing its parameters for IoT environments, improving detection accuracy in dynamic and high-dimensional data streams.

They use Apache Spark and the SCALA programming language to implement their modified DBSCAN method. Additionally, RDD-DBSCAN, a Resilient Distributed Dataset, is used by their technique. They

can handle two-dimensional data sets using their distributed approach. It gets around the drawbacks of the standard DBSCAN and MapReduce models when handling large amounts of data.

In [40], the authors suggested two methods for identifying and eliminating outliers using an outlier score. The first is distance-based and determines how comparable data values are by using the Euclidean distance. The second approach is cluster-based, utilizing the k-means algorithm along with Euclidean distance to group similar data points while identifying anomalies. By measuring the distance between points and their respective cluster centroids, this method effectively detects outliers that deviate significantly from the cluster structure. It is widely used due to its simplicity and efficiency, particularly in large-scale datasets where quick and accurate anomaly detection is essential. To assess the quality of clusters and eliminate outliers, they employ two outlier scores: likelihood ratio and F-score. They make use of the R statistical computing project's health care dataset. They demonstrate that when selecting the right outlier score, the second method performs more accurately than the first.

An online technique for density-based outlier detection on large data sets was presented by the authors in [41]. This process, which is focused on determining the Local Outlier Factor (LOF), consists of two steps [42]. In the initial phase, the data is divided into grids and dispersed among the network's dispersed nodes using Grid-Based Partitioning (GBP). During the second phase, density outliers are found in parallel using the Distributed LOF Computing technique (DLC). Their method saves a lot of network resources and can manage high complexity when working with huge dimensional data collections.

3. Conclusion

Choosing the right outlier detection technique depends on the specific characteristics of the dataset and the application at hand. Statistical methods offer simplicity, while machine learning techniques provide scalability and adaptability to complex datasets. Distance-based and clustering methods can be effective but require more careful parameter tuning. As data continues to grow in complexity and volume, ongoing research into hybrid and ensemble methods may provide the best balance of performance and reliability.

References:

1. Kumar, V. 2005. Parallel and Distributed Computing for Cybersecurity. Distributed Systems Online, IEEE 6, 10.

2. Spence, C., Parra, L., and Sajda, P. 2001. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. IEEE Computer Society, Washington, DC, USA, 3.
3. Aleskerov, E., Freisleben, B., and Rao, B. 1997. Cardwatch: A neural network based database mining system for credit card fraud detection. In Proceedings of IEEE Computational Intelligence for Financial Engineering. 220-226.
4. Fujimaki, R., Yairi, T., and Machida, K. 2005. An approach to spacecraft outlier detection problem using kernel feature space. In Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM Press, New York, NY, USA, 401-410
5. Edgeworth, F. Y. 1887. On discordant observations. Philosophical Magazine 23, 5, 364 -375.
6. Teng, H., Chen, K., and Lu, S. 1990. Adaptive real-time outlier detection using inductively generated sequential patterns. In Proceedings of IEEE Computer Society Symposium on Re-search in Security and Privacy. IEEE Computer Society Press, 278-284.
7. Rousseeuw, P. J. and Leroy, A. M. 1987. Robust regression and outlier detection. John Wiley & Sons, Inc., New York, NY, USA.
8. Huber, P. 1974. Robust Statistics. Wiley, New York.
9. Markou, M. and Singh, S. 2003a. Novelty detection: a review-part 1: statistical approaches. Signal Processing 83, 12, 2481-2497.
10. Markou, M. and Singh, S. 2003b. Novelty detection: a review-part 2: neural network based approaches. Signal Processing 83, 12, 2499-2521.
11. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. ACM Comput. Surv. (CSUR) 2009, 41, 1–58. [CrossRef]
12. Castillo, A.; Thierer, A.D. Projecting the Growth and Economic Impact of the Internet of Things. SSRN Electron. J. 2015. [CrossRef]
13. Branch, J.W.; Giannella, C.; Szymanski, B.; Wolff, R.; Kargupta, H. In-network outlier detection in wireless sensor networks. Knowl. Inf. Syst. 2013, 34, 23–54.
14. Martincic, F.; Schwiebert, L. Distributed event detection in sensor networks. In Proceedings of the International Conference on Systems and Networks Communications (ICSNC 2006), Tahiti, French Polynesia, 29 October–3 November 2006; p. 43. [CrossRef]

15. Krishnamachari, B.; Iyengar, S. Distributed Bayesian algorithms for fault-tolerant event region detection in wireless sensor networks. IEEE Trans. Comput. 2004, 53, 241–250. [CrossRef]
16. Shahid, N.; Naqvi, I.H.; Qaisar, S.B. Characteristics and classification of outlier detection techniques for wireless sensor networks in harsh environments: A survey. Artif. Intell. Rev. 2012, 43, 193–228. [CrossRef]
17. Zhang, Y.; Meratnia, N.; Havinga, P. Outlier detection techniques for wireless sensor networks: A survey. IEEE Commun. Surv. Tutor. 2010, 12, 159–170. [CrossRef]
18. Ding, M.; Chen, D.; Xing, K.; Cheng, X. Localized fault-tolerant event boundary detection in sensor networks. In Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, Miami, FL, USA, 13–17 March 2005; Volume 2, pp. 902–913. [CrossRef]
19. Lazarevic, A.; Ertoz, L.; Kumar, V.; Ozgur, A.; Srivastava, J. A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection. In Proceedings of the 2003 SIAM International Conference on Data Mining, San Francisco, CA, 1–3 May, 2003; pp. 25–36. [CrossRef]
20. Ganesh Kumar, D.; Insozhan, N.; Parthasarathy, V. Recognition of faulty node detection using fuzzy logic in iot. Int. J. Sci. Technol. Res. 2019, 8, 1112–1116.
21. Cook, A.A.; Misirli, G.; Fan, Z. Anomaly Detection for IoT Time-Series Data: A Survey. IEEE Internet Things J. 2020, 7, 6481–6494. [CrossRef]
22. Ayadi, A.; Ghorbel, O.; Obeid, A.M.; Abid, M. Outlier detection approaches for wireless sensor networks: A survey. Comput. Netw. 2017, 129, 319–333. [CrossRef]
23. Jiang, J.; Han, G.; Liu, L.; Shu, L.; Guizani, M. Outlier detection approaches based on machine learning in the internet-of-things. IEEE Wirel. Commun. 2020, 27, 53–59. [CrossRef]
24. Kumar Dwivedi, R.; Pandey, S.; Kumar, R. A Study on Machine Learning Approaches for Outlier Detection inWireless Sensor Network. In Proceedings of the 8th International Conference Confluence 2018 on Cloud Computing, Data Science and Engineering (Confluence), Noida, India, 11–12 January 2018; pp. 189–192. [CrossRef]
25. Ghosh, N.; Maity, K.; Paul, R.; Maity, S. Outlier detection in sensor data using machine learning techniques for IoT framework and wireless sensor networks: A brief study. In Proceedings of the 2019 International Conference on Applied Machine Learning (ICAML'19), Bhubaneswar, India, 25–26 May 2019; pp. 187–190. [CrossRef]
26. Wang, H.; Bah, M.J.; Hammad, M. Progress in Outlier Detection Techniques: A Survey. IEEE Access 2019, 7, 107964–108000. [CrossRef]

27. Morales, L.V.V.; López-Vizcaíno, M.; Iglesias, D.F.; Díaz, V.M.C. Anomaly Detection in IoT: Methods, Techniques and Tools. Proceedings 2019, 21, 4. [CrossRef]
28. Sheng, B.; Li, Q.; Mao,W.; Jin,W. Outlier detection in sensor networks. In Proceedings of the International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc), Montreal QC Canada, 9–14 September 2007; pp. 219–228. [CrossRef]
29. Palpanas, T.; Papadopoulos, D.; Kalogeraki, V.; Gunopulos, D. Distributed deviation detection in sensor networks. ACM Sigmod Rec. 2003, 32, 77–82. [CrossRef]
30. Panda, M.; Khilar, P.M. Distributed soft fault detection algorithm in wireless sensor networks using statistical test. In Proceedings of the 2012 2nd IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC 2012), lSolan, India, 6–8 December 2012; pp. 195–198. [CrossRef]
31. Zhang, Y.; Hamm, N.A.; Meratnia, N.; Stein, A.; van de Voort, M.; Havinga, P.J. Statistics-based outlier detection for wireless sensor networks. Int. J. Geogr. Inf. Sci. 2012, 26, 1373–1392. [CrossRef]
32. Xie, M.; Hu, J.; Tian, B. Histogram-based online anomaly detection in hierarchical wireless sensor networks. In Proceedings of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom-2012—11th IEEE International Conference on Ubiquitous Computing and Communications, IUCC-2012, Liverpool, UK, 25–27 June 2012; pp. 751–759. [CrossRef]
33. Boedihardjo, A.P.; Lu, C.T.; Chen, F. Fast adaptive kernel density estimator for data streams. Knowl. Inf. Syst. 2015, 42, 285–317. [CrossRef]
34. Lv, Y. An Adaptive Real-time Outlier Detection Algorithm Based on ARMA Model for Radar's Health Monitoring. In Proceedings of the 2015 IEEE AUTOTESTCON, National Harbor, MD, USA, 2–5 November 2015.
35. Nesa, N.; Ghosh, T.; Banerjee, I. Outlier detection in sensed data using statistical learning models for IoT. In Proceedings of the IEEEWireless Communications and Networking Conference, WCNC, Barcelona, Spain, 15–18 April 2018; pp. 1–6. [CrossRef]
36. Ahmed, M.; Mahmood, A.N. A novel approach for outlier detection and clustering improvement. In Proceedings of the 2013 IEEE 8th Conference on Industrial Electronics and Applications, ICIEA 2013, Melbourne, VIC, Australia, 19–21 June 2013; pp. 577–582. [CrossRef]
37. Souza, A.M.; Amazonas, J.R. An outlier detect algorithm using big data processing and Internet of Things architecture. Procedia Comput. Sci. 2015, 52, 1010–1015. [CrossRef]

38. Hydra Technology Project—In-JeT ApS. Available online: https://www.in-jet.eu/portfolio-items/hydra/ (accessed on 20 December 2021).
39. Cordova, I.; Moh, T.S. DBSCAN on Resilient Distributed Datasets. In Proceedings of the 2015 International Conference on High Performance Computing and Simulation, HPCS 2015, Amsterdam, The Netherlands, 20–24 July 2015; pp. 531–540. [CrossRef]
40. Christy, A.; Gandhi, M.G.; Vaithyasubramanian, S. Cluster based outlier detection algorithm for healthcare data. Procedia Comput. Sci. 2015, 50, 209–215. [CrossRef]
41. Bai, M.; Wang, X.; Xin, J.; Wang, G. An efficient algorithm for distributed density-based outlier detection on big data. Neurocomputing 2016, 181, 19–28. [CrossRef]
42. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying Density-Based Local Outliers. Int. J. Gynecol. Obstet. 2009, 107, S93.
43. Tian, H.X.; Liu, X.J.; Han, M. An outliers detection method of time series data for soft sensor modeling. In Proceedings of the 28th Chinese Control and Decision Conference, CCDC 2016, Yinchuan, China, 28–30 May 2016; pp. 3918–3922. [CrossRef]
44. Xie, M.; Hu, J.; Guo, S.; Zomaya, A.Y. Distributed Segment-based Anomaly Detection with Kullback-Leibler Divergence in Wireless Sensor Networks. IEEE Trans. Inf. Forensics Secur. 2016. 12, 101–110. [CrossRef]
45. Lyu, L.; Jin, J.; Rajasegarar, S.; He, X.; Palaniswami, M. Fog-empowered anomaly detection in IoT using hyperellipsoidal clustering. IEEE Internet Things J. 2017, 4, 1174–1184. [CrossRef]