

A Novel Machine Learning Framework for Outlier Detection in High Dimensional Data

Deepak

Former Assistant Professor of Mathematics

C.R.A. College, Sonipat, Haryana, India

deepakdahiya707@gmail.com

Abstract: Outlier detection in high-dimensional datasets is a critical task in data mining and machine learning, particularly in applications involving large-scale and complex data. Traditional outlier detection techniques often face limitations related to high dimensionality, the availability of labeled data, detection accuracy, and scalability. This study presents a learning-based framework for improving outlier identification by integrating supervised, unsupervised, hybrid, and ensemble learning strategies. The proposed approach, CSS-EOD, employs an effective selection strategy to combine complementary outlier detection techniques and improve the identification of anomalous instances. The framework is designed to exploit the advantages of both labeled and unlabeled data while providing improved applicability to large-scale datasets and cloud-based big-data environments. The effectiveness of the proposed method was evaluated using benchmark high-dimensional datasets, including Arrhythmia, Cardio, Letter, MNIST, Satellite, Stamps, and Thyroid. The performance was assessed using ROC-AUC and mean average precision (mAP), with experimental results averaged over 30 independent runs. The results demonstrate that CSS-EOD consistently achieves competitive or superior performance compared with established outlier detection methods. In particular, on the Thyroid dataset, CSS-EOD achieved a ROC-AUC of 97.48%, outperforming Aggarwal (97.04%), Aggarwal and Sathe (96.95%), Kumar (95.57%), Rayana et al. (94.31%), and Zimek et al. (97.13%). Similarly, CSS-EOD achieved the highest mAP of 46.74%, compared with 40.65%, 41.43%, 35.05%, 28.64%, and 41.50% for the corresponding existing methods. These findings demonstrate that the proposed framework can effectively enhance outlier detection accuracy and provide a promising scalable solution for high-dimensional and large-scale data environments.

Keywords: Outlier Detection, High-Dimensional Data, Anomaly Detection, Big Data, Cloud Computing, Scalable Outlier Detection.

I. INTRODUCTION

Latent patterns are referred to as outliers when they differ from the remainder of the data in real world business data. Stated differently, unique data instances vary significantly from those inside a specific dataset. Outliers are now recognized as a substantial subject for research across a range of fields and applications, although in the past, they were considered noisy data. To enhance the caliber of data analytics and generate the necessary business intelligence (BI), several academics have created outlier identification techniques that help generate new knowledge and remove outliers. In other words, outliers can be used in two separate ways. To use the quality of the mined data, they must first be removed from the dataset. The second objective is to extract fresh data related to outliers. Applications have different criteria for how they employ outliers. However, both strategies must identify outliers in the supplied data. The data utilized in many real-world applications is high-dimensional. It has a large number of dimensions, which makes calculations more challenging. Handling such data is necessary in order to get knowledge that is valuable or in order to raise the data's quality by removing outliers.



The topic of outlier identification research benefited from the several outlier detection techniques now in use, including Local Outlier Factor (LOF) [2], Local Distance-based Outlier Factor (LDOF) [5], Local Projection Score (LPS) [1], and others. However, as the most recent literature indicates, further study is required in areas such as ensemble approaches with the optimal selection of candidate methods and seamless and efficient integration of Learning techniques, both supervised and unsupervised, including outlier identification based on clustering. By proposing a novel framework and set of algorithms, this thesis aims to address the shortcomings of existing methods for spotting outliers in high dimensional data, advancing the state of the art.

The study presented in this paper was motivated by essential discoveries found during a literature survey on contemporary outlier identification techniques. More precisely, the literature has highlighted three significant issues. Firstly, the issue of adaptive clustering is covered in [5] and [8]. It has limits when it comes to using local density and distance information to its full potential for outlier detection. Secondly, unsupervised learning techniques are more likely to represent data, whereas supervised techniques use training data with high-quality prediction skills. However, the issue noted is how to combine these two approaches with a unique framework [7]. Third, as elaborated in [6], since current ensemble techniques do not have an optimal strategy for identifying constituent detectors to be included in the ensemble, ensemble based optimization of the outlier detection methodology is required. Fourth, as mentioned in [19] and [20], scalability is another critical consideration in outlier identification. To enhance the state of the art, this thesis suggests a unique architecture with underlying algorithms to handle these issues.

II. RELATED WORK

The written works have several contributions to the field of outlier identification research. M. Bai, X. Wang, J. Xin, and G. Wang [4] developed the Grid-Based Partition method (GBP) and Distributed Local Outlier Factor Computing (DLC) to identify outliers. Compared to previous methods, this approach processed huge datasets in less processing time. This approach uses fewer resources when working with large-scale data. However, the considerable overhead of the network resources used by the current approaches suggests that their performance is constrained.

Furthermore, the developed method's increased dimensionality causes significant time consumption and performance degradation while calculating the distance between two tuples. R. J. Campbell, D. Moulavi, A. Zimek, and J. Sander [5] used the Global-Local Outlier Scores from Hierarchies (GLOSH) to determine whether outliers were present. A comprehensive hierarchy of density-based clustering was built to identify answers to an infinite number of density criteria. This made the GLOSH as reliable as other methods for outlier identification, such as kNN, the LDOF approach, and local outlier probabilities (LoOP). Unfortunately, the findings show that this approach is not able to produce more accurate density estimates, which leads to inaccurate computing estimates.

B. Tang and H. He [6] developed a local outlier detection method based on kernel density estimation (KDE). KDE was utilized to measure the object's local outlier using the density distribution at the item's location. To estimate the density distribution of an item, the KDE technique considered its shared closest neighbor and reverse nearest neighbor. More excellent performance was shown when the KDE approach was evaluated on artificial and real-time datasets. Unfortunately, real-world applications cannot utilize the KDE approach as it is only effective on standard datasets.

M. A. Rahman, K. L. M. Ang, and K. P. Seng [7] created parameters using the two independent density adjacent clustering techniques. The clustering method included two distinct approaches: a unique neighborhood set and a unique nearest neighborhood. While the other technique produced better results even with the explicit outlier present, the first method simulated using a dataset devoid of explicit outliers. However, the random structure of the clusters and their incapacity to handle complicated datasets lowered the computational complexity of the techniques developed.

A. Abid, A. Kachouri, and A. Mahfoudhi [8] developed a technique called Density-Based Spatial Clustering of Applications with Noise Outlier Detection (DBSCANOD). This approach was used to calculate the lowest approved cluster with radius value to create the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) procedure to group data. Choosing each point, This approach determined whether the end was typical or atypical. However, amelioration is crucial in finding outliers, particularly in the vicinity of the outlier spots. According to an



examination of available methods, the main problem is that when two data sets are analyzed, their dimensionality rises, resulting in lengthy calculation times. These methods rely only on the clustering process' random selection of cluster centers.

Liu et al. [9] examined the challenge of handling high-dimensional data to forecast anomalies. While there have been some studies, High-dimensional data has not received as much attention when it comes to outlier spotting or taking neighborhood size into account as it might be. They established the Local Projection Score (LPS) to show the departure from the area. They were able to get LPS data by using low-rank approximation. A greater probability of detecting outliers is made possible by high LPS. They created and assessed the LPOD method based on this idea.

Their goal was to enhance LPOD in the future. Kriegel et al. [10] suggested the Angle Based Outlier Detection (ABOD) and related techniques for multidimensional data. They discovered that ABOD outperforms the distance-based approach LOF. Li and Kitagawa [11] suggested a strategy for detecting outliers based on distance. Examples and subspaces are both its foundations. To locate appropriate subspaces, choose parameters, and identify outliers, a genetic algorithm (GA) is employed. Compared to earlier models, it has a greater accuracy rate and convergence. Regarding the fitness value function that GA employs, their approach needs refinement. Verma et al. [12] investigated techniques for identifying suspicious conduct without supervision. That's what the video surveillance system does. Using deep learning models, they hoped to enhance their approach.

Zhang et al. [13] developed a technique for determining outlierness called the LDOF. Stated differently, it shows how much an object differs relative to its neighbors across distance. They aimed to enhance dispersed real-world datasets. The research mentioned above clarifies that unsupervised learning techniques offer more options to represent data. For the purpose of discovering outliers, the suggested approach in this research takes that into account and uses hybrid methods that combine supervised and unsupervised learning.

Aggarwal [14] looked at several outlier ensemble types. Both classification and clustering were taken into consideration during their analysis. They discovered many frequently used combination functions, including the pruned, damped, and maximal functions. They discovered mid-range assessment, diversity, and consensus problems using unsupervised techniques. Model averaging, bagging, random forests, boosting, and bucketing are some categorization techniques utilized in ensemble analysis. They aimed to achieve more valuable business information (BI) using ensemble outlier approaches. Aggarwal [15] explored various methods based on the subspace approach for outlier detection. They consist of unbiased, aggregation-based, and rarity based methods. Of all the subspace-related problems, they believed outlier identification was the most difficult. Chakraborty et al. [16] investigated deep feature learning techniques in conjunction with ensemble settings with outliers. Stacking autoencoders are used for feature extraction. Outliers are identified using majority voting. They want to concentrate on several instances of this kind of outlier in the future.

Zhang et al. [17] suggested the Locality Sensitive Hashing (LSH) forest genetic architecture. It is a quick tree isolation ensemble approach. They planned to enhance it to facilitate anomaly detection with reduced latency. Trittenbach and Bohm [18] suggested a dimension-based subspace search technique for identifying outliers. Additionally, they offered a heuristic called Greedy Maximum Deviation (GMD). Runtime, number of subspaces, robustness, and quality of search results are the particular parameters that determine its basis. Deng et al. [19] noted that finding outliers in large, multidimensional datasets is challenging. Tensor-train layers and ELM are integrated into the proposed ETFLM, significantly reducing high-dimensional data and improving detection accuracy by more than 60%. Muneer et al. [20] A hybrid model that utilizes a Deep Autoencoder Neural Network (DANN) effectively handles high-dimensional anomaly detection in a gas turbine dataset to obtain increased accuracy with Adamax optimization. The literature review above demonstrates that many unsupervised learning algorithms' ensemble approaches require an ideal mechanism for choosing suitable candidates to act as component detectors. As a consequence, the model's stability and accuracy were just average.



III. RESEARCH METHODOLOGY

This part presents the research for this thesis, which tries to meet all of the research goals. The study included several methods that developed novel outlier identification strategies by applying learning-based approaches. It is thought that performance in the outlier identification process might be improved by employing learning-based techniques to find anomalies in data with several dimensions. This study looked at several theories, including the validation of an unsupervised approach and a hybrid approach that blends supervised and unsupervised methods, an ensemble learning strategy that merges the findings of several constituent outlier detection methods selected using an innovative selection strategy, and a methodology to make outlier detection methods scalable for handling massive amounts of data in the big data and cloud computing era.

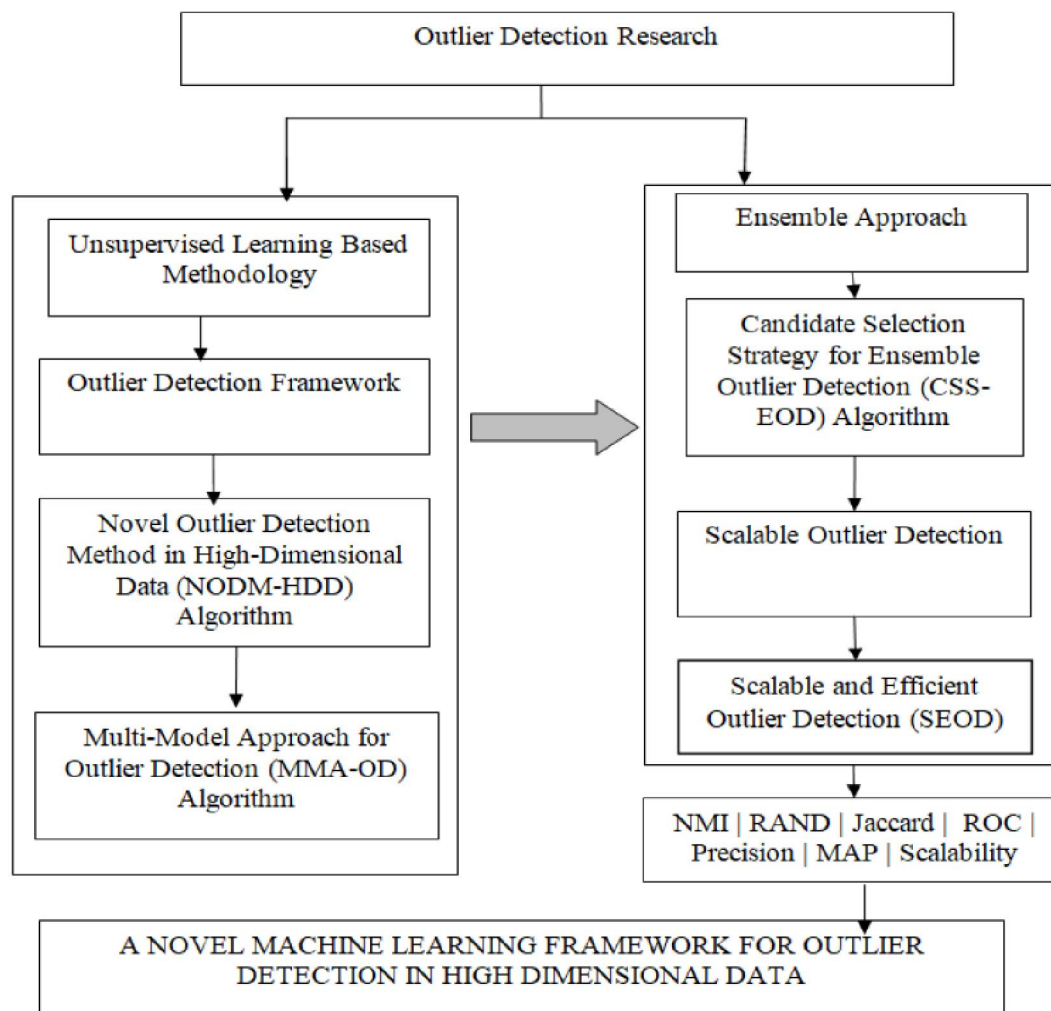


Figure 1: Overview of the research carried out

This paper employed various approaches to find outliers, including ensemble learning, supervised learning, unsupervised learning, and a hybrid technique that blends the two. Several approaches were put forth to enhance outlier identification in high-dimensional data. A hybrid strategy incorporating both supervised and unsupervised learning systems has benefits; outlier detection is possible with unsupervised learning even when labeled data is unavailable. The ensemble approach also uses suitable techniques and a unique selection procedure to enhance outlier detection. In



cloud big data ecosystems, a scalability-based methodology is handy for processing bigger amounts of data via improving outlier identification tools.

IV. RESULT ANALYSIS

The suggested algorithm was assessed using several datasets sourced from [26]. In the empirical investigation, nine datasets were used. The Arrhythmia dataset has 274 dimensions, 452 points, 14.60% outliers, and 65 outliers. The Cardio dataset has 9.61% outliers, 176 outliers, 21 dimensions, and 1831 points. The Letter dataset has 6.25% outliers, 100 outliers, 32 dimensions, and 1600 points. The MNIST dataset consists of 7603 points, 100 dimensions, 700 outliers, and 9.21% outliers. The Satellite dataset has 6435 points, 36 dimensions, 2036 outliers, and 31.64% outliers. The Stamps dataset consists of 340 points, 9 dimensions, 31 outliers, and 9.12% outliers. There are 93 outliers, 6 dimensions, 3772 points, and 2.47% outliers in the Thyroid dataset. Comparisons are made between the experimental results and the most advanced techniques regarding mean average precision (mAP) and receiver operating characteristic (ROC-AUC). The average of 30 experiments' worth of observations is given in this section.

Table 1: ROC-AUC performance comparison

Dataset	ROC-AUC(%)	Aggarwal [21]	Aggarwal and Sathe [22]	Kumar [23]	Rayana et al. [24]	Zimek et al. [25]	CSS EOD
Arrhythmia	0.779679	0.778775		0.772845	0.769428	0.779679	0.780182
Cardio	0.881385	0.890933		0.894752	0.884199	0.882591	0.905807
Letter	0.796463	0.807116		0.83415	0.847617	0.794754	0.790634
MNIST	0.859979	0.863094		0.859577	0.839075	0.860582	0.867617
Satellite	0.591041	0.602196		0.62511	0.628929	0.590538	0.604508
Stamps	0.899073	0.897164		0.880682	0.86018	0.899777	0.902993

As Table 1 shows, the suggested approach CSS-EOD outperforms the current methods in terms of ROC-AUC. Seven distinct datasets measure performance.

Table 2: Mean average Precision Comparison

Dataset	Map	Aggarwal [21]	Aggarwal and Sathe [22]	Kumar [23]	Rayana et al. [24]	Zimek et al. [25]	ISEC-OD
Arrhythmia	0.378483	0.378785		0.374061	0.370845	0.378483	0.381498
Cardio	0.353358	0.372654		0.388332	0.368433	0.355268	0.413759
Letter	0.239994	0.248537		0.288134	0.31758	0.238386	0.241904
MNIST	0.393056	0.396071		0.391548	0.371951	0.393759	0.39989
Satellite	0.406724	0.41597		0.437376	0.440693	0.406724	0.421698
Stamps	0.371247	0.36783		0.340394	0.315972	0.372453	0.37979
Thyroid	0.406523	0.414362		0.350544	0.286425	0.415065	0.467426

The analysis of the suggested methods' respective mAP performances technique CSSEOD with that of known methods is presented in Table 5.2. With seven distinct datasets, performance is seen.



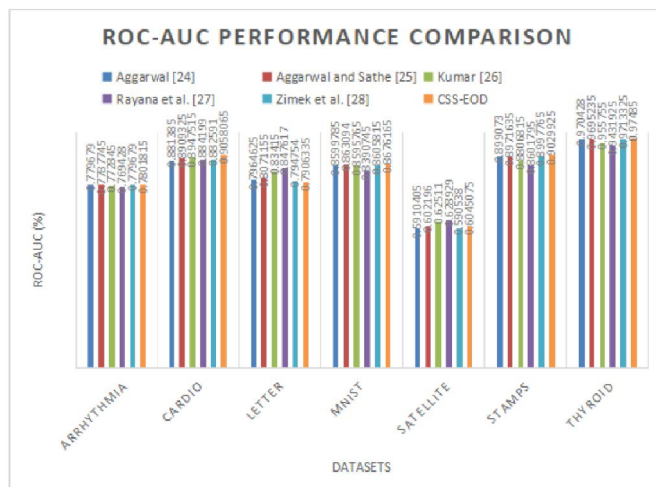


Figure 2: ROC-AUC performance comparison

The ROC-AUC metric is used to compare the effectiveness of various outlier identification methods, as shown in Figure 2. Seven datasets employed in the empirical investigation are the subject of observations. Due to their various modes of operation, it was discovered that every outlier detection model had distinct performance statistics. Nonetheless, all state-of-the-art techniques were outperformed by the suggested CSS-EOD algorithm. Aggarwal [21]'s technique yielded a ROC-AUC of 97.04% for the Thyroid dataset; Aggrawal and Sathe [22] reported 96.95%; Kumar [23] reported 95.57%; Rayana et al. [24] reported 94.31%; Zimeket al. [25] reported 97.13%; and CSS-EOD reported 97.48%. It is discovered that, in terms of ROC-AUC, the suggested method, CSS-EOD, performs best.

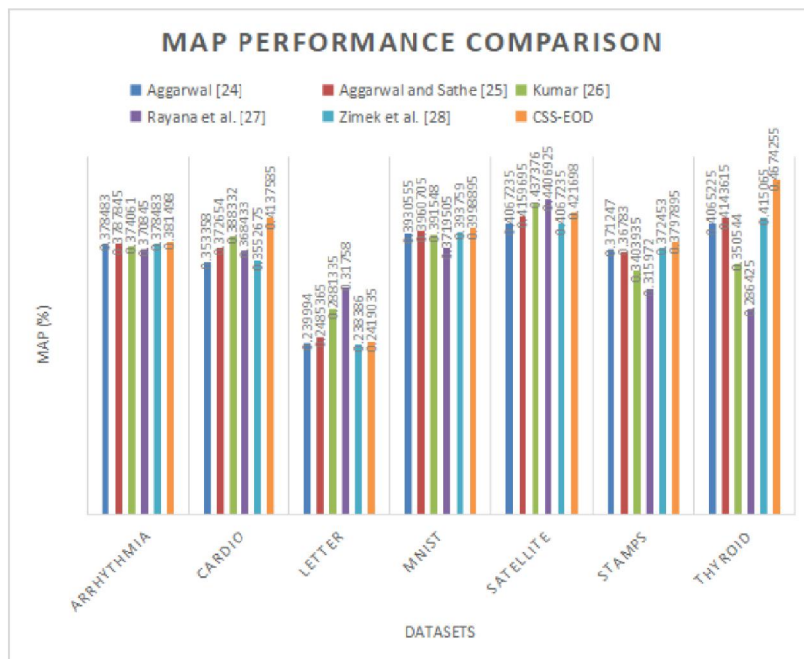


Figure 3: Mean average precision performance comparison



A comparison of several outlier identification methods is shown according to Figure 5.3's mean average precision (mAP) measure. Seven datasets used in the empirical study are the focus of the observations. It is shown that each outlier detection model exhibited unique performance statistics due to their different operating processes. The suggested CSS-EOD algorithm

demonstrated the best performance among all the cutting-edge methods. Performance is enhanced when the mAP value is more significant. The approach of Aggarwal [21] demonstrated 40.65% mAP for the Thyroid dataset; Aggarwal and Sathe [22] demonstrated 41.43%; Kumar [23] demonstrated 35.05%; Rayana et al. [24] demonstrated 28.64%; Zimeket al. [25] demonstrated 41.50%; and CSS-EOD demonstrated 46.74%. According to mAP, the suggested technique, CSS-EOD, performs best.

V. CONCLUSION

The study focused on implementing the Unsupervised Ensemble Outlier Detection (UEOD) framework. Despite not being a supervised learning technique, it incorporates implicit training processes to provide a global ground truth that directs the process of generating local ground truth. In order to create variety in a group of base detectors, a method known as Candidate Selection Strategy for Ensemble Outlier Detection (CSS-EOD) is employed. It gets the global ground truth, which is then used to generate the local ground truth. This method makes a fully unsupervised learning-based method possible. The empirical investigation used a large number of high-dimensional datasets.

The experiment's outcomes proved that CSS-EOD is superior to current techniques. The recommended doubt layer detection method for finding outliers in this chapter and the two previous chapters offer superior methods in high-dimensional data procedures. A mechanism for making any outlier identification technique that is now in use scalable while handling massive amounts of high-dimensional data must be developed.

REFERENCES

1. Kumar Dwivedi, R., Pandey, S., & Kumar, R. (2018). A Study on Machine Learning Approaches for Outlier Detection in Wireless Sensor Network. 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence). P1-4.
2. Jiang, J., Han, G., Liu, L., Shu, L., & Guizani, M. (2020). Outlier Detection Approaches Based on Machine Learning in the Internet-of-Things. IEEE Wireless Communications, 27(3), 53–59.
3. Deng, X., Jiang, P., Peng, X., & Mi, C. (2018). An Intelligent Outlier Detection Method with One Class Support Tucker Machine and Genetic Algorithm towards Big Sensor Data in Internet of Things. IEEE Transactions on Industrial Electronics, 1–11.
4. Liu, Y., Li, Z., Zhou, C., Jiang, Y., Sun, J., Wang, M., & He, X. (2019). Generative Adversarial Active Learning for Unsupervised Outlier Detection. IEEE Transactions on Knowledge and Data Engineering, 1–12.
5. Rayana, S., Zhong, W., & Akoglu, L. (2016). Sequential Ensemble Learning for Outlier Detection: A Bias-Variance Perspective. 2016 IEEE 16th International Conference on Data Mining (ICDM). P1-6.
6. Malini, N., & Pushpa, M. (2017). Analysis on credit card fraud identification techniques based on KNN and outlier detection. 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). P1-4.
7. Liu, H., Li, X., Li, J., & Zhang, S. (2017). Efficient Outlier Detection for High-Dimensional Data. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 1–11.
8. Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognition, 74, 406–421.
9. Nesa, N., Ghosh, T., & Banerjee, I. (2018). Non-parametric sequence-based learning approach for outlier detection in IoT. Future Generation Computer Systems, 82, 412–421.
10. Nguyen, H., Cai, C., & Chen, F. (2017). Automatic classification of traffic incident's severity using machine learning approaches. IET Intelligent Transport Systems, 11(10), 615–623.



11. Zhao, H., Liu, H., Ding, Z., & Fu, Y. (2018). Consensus Regularized Multi-View Outlier Detection. *IEEE Transactions on Image Processing*, 27(1), 236–248.
12. Chakraborty, D., Narayanan, V., & Ghosh, A. (2019). Integration of Deep Feature Extraction and Ensemble Learning for Outlier Detection. *Pattern Recognition*. P1-13.
13. Christy, A., Gandhi, G. M., & Vaithya subramanian, S. (2015). Cluster Based Outlier Detection Algorithm for Healthcare Data. *Procedia Computer Science*, 50, 209–215.
14. Souza, A. M. C., & Amazonas, J. R. A. (2015). An Outlier Detect Algorithm using Big Data Processing and Internet of Things Architecture. *Procedia Computer Science*, 52, 1010–1015.
15. Munoz-Organero, M., Ruiz-Blaquez, R., & Sánchez-Fernández, L. (2018). Automatic detection of traffic lights, street crossings and urban roundabouts combining outlier detection and deep learning classification techniques based on GPS traces while driving. *Computers, Environment and Urban Systems*, 68, 1–8.
16. Maniruzzaman, M., Rahman, M. J., Al-Mehedi Hasan, M., Suri, H. S., Abedin, M. M., El- Baz, A., & Suri, J. S. (2018). Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers. *Journal of Medical Systems*, 42(5). P1-17.
17. Ren, J., Guo, J., Qian, W., Yuan, H., Hao, X., & Jingjing, H. (2019). Building an Effective Intrusion Detection System by Using Hybrid Data Optimization Based on Machine Learning Algorithms. *Security and Communication Networks*, 2019, 1–11.
18. Erkuş, E. C., & Purutçuoğlu, V. (2020). Outlier Detection and Quasi-periodicity Optimization Algorithm: Frequency Domain Based Outlier Detection (FOD). *European Journal of Operational Research*. P1-19.
19. Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M., & Sun, J. (2017). Anomaly Detection for a Water Treatment System Using Unsupervised Machine Learning. 2017 IEEE International Conference on Data Mining Workshops (ICDMW). P1-8.
20. SADGALI, I., SAEL, N., & BENABBOU, F. (2019). Performance of machine learning techniques in the detection of financial frauds. *Procedia Computer Science*, 148, 45–54.
21. Sandosh, S., Govindasamy, V., & Akila, G. (2020). Enhanced intrusion detection system via agent clustering and classification based on outlier detection. *Peer-to-Peer Networking and Applications*. P1-8.
22. Bhatti, M. A., Riaz, R., Rizvi, S. S., Shokat, S., Riaz, F., & Kwon, S. J. (2020). Outlier detection in indoor localization and Internet of Things (IoT) using machine learning. *Journal of Communications and Networks*, 22(3), 236–243.
23. Moustafa, N., Choo, K.-K. R., Radwan, I., & Camtepe, S. (2019). Outlier Dirichlet Mixture Mechanism: Adversarial Statistical Learning for Anomaly Detection in the Fog. *IEEE Transactions on Information Forensics and Security*, 1–13.
24. Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., Nenadic, G. (2018). Machine learning methods for wind turbine condition monitoring: A review. *Renewable Energy*. P1-23.
25. Ma, B., Yuan, L., Xu, S., Zheng, K., Huang, F., Li, R., & Yuan, P. (2020). Positive Active Power Outlier Detection based on One-Class SVM. 2020 12th IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC). P1-4.
26. Zhao, Yue; Hryniewicki, Maciej K. (2018). IEEE 2018 International Joint Conference on Neural Networks (IJCNN) - Rio de Janeiro, Brazil (2018.7.8-2018.7.13)] 2018 International Joint Conference on Neural Networks (IJCNN) - XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning, p1–8.

