A Comparative Study of Outlier Detection Techniques for Unsupervised Models: LOF, OC-SVM, and IForest

Nur Syuhaidah Nor Azni1, 2, a) and Noraswaliza Abdullah3, b)

*1Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.*

*2Centre for Advanced Analytics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia*

*3Faculty of Information Technology and Communications, Universiti Teknikal Malaysia (UTeM), Melaka, Malaysia*

*a) Corresponding author: syuhaidah.azni@mmu.edu.my*

*b) noraswaliza@utem.edu.my*

**Abstract.** Detecting anomalous patterns is essential when working with real-life datasets, especially in sensitive sctors like healthcare and finance. This paper conducts a comparative evaluation of three widely used unsupervised outlier detection techniques: Local Outlier Factor (LOF), One-Class Support Vector Machine (OC-SVM), and Isolation Forest (IF). The assessment involved five health-related datasets: cervical cancer, diabetes, maternal health risk, caesarean, and dengue. Each technique was assessed using two evaluation metrics—Receiver Operating Characteristic (ROC) and top-ranked precision scored (P@N)—implemented via the PyOD library. Finding indicate that IF consistently surpassed the performance of the other two methods, achieving the highest average ROC (0.7297) and P@N (0.5975) across most datasets. While OC-SVM performed best on the cervical cancer dataset, while LOF showed superior results for dengue data. These findings suggest that IF is well-suited for unsupervised, small-scale medical datasets due to its efficiency and robustness. Future work will explore hybrid models integrating IF with clustering algorithms to enhance detection accuracy and scalability.

# INTRODUCTION

In data science, transforming unprocessed information into structured and meaningful insights is a core objective. However, this process is often disrupted by unusual or inconsistent records known as anomalies or outliers. These abnormal entries can distort analysis and model accuracy, making their detection as essential step in data preparation. Various anomaly detection strategies esixt, including Local Outlier Factor (LOF) [1]–[9], Isolation Forest (IF) [5], [10]–[24], One-Class Support Vector Machine [13], [14], [25], [26]. These techniques have been widely implemented across multiple sectors, including healthcare.

This study investigates the comparative performance of three prominent anomaly detection models – LOF, OC-SVM and IF – applied to five unlabeled healthcare-related datasets: cervical cancer, diabetics, maternal health risk, caesarean, and dengue. Each model employs a unique methodology for identifying unusual patterns in data. LOF for example, examines the local density around a point and contrasts it with the densities of its neighboring points [5]. OC-SVM identifies boundaries by separating the data distribution from the origin in a high-dimensional space to maximize margin [25]. Isolation Forest, on the other hand, it works by partitioning randomly selected subsets of features into trees, where anomalies are typically found closer to the root due to fewer splits needed to isolate them [5].

There are generally three categories of anomaly detection models: clustering-based, proximity-based, and density-based. Clustering-based approaches aim to organize data instances into cohesive groups by analyzing their mutual characteristics. This is done using information derived directly from the structure and distribution of the dataset. Examples of such techniques include K-Means, K-Medoids, and IF [13], [15], [16]. These algorithms classify items that fail to conform to any group as anomalies. Proximity-based detection methods focus on the distance between items in the dataset. These models identify anomalies by measuring how far a point lies from its nearest neighbors, often using metrics like the Euclidean distance. Any point that lacks a sufficient number of nearby neighbors – commonly fewer than k – is marked as an anomaly. OC-SVM belong to this category of model. Meanwhile, density-based approached detect unusual entries based on variations in the local data density around each record.

Various techniques have been proposed for identifying anomalies in datasets, such as the LOF, which measures the relative density of a data point compared to its neighboring values [5]. Numerous studies have explored anomaly detection across different domain and algorithm categories. For instance, H.C. Mandhare et al. [8] compared cluster-based, distance-oriented, and density-focused models to determine their effectiveness in spotting outliers. Their findings highlighted that cluster-based methods tend to be simpler and more computationally efficient, whereas distance and density approaches were associated with moderate to high complexity. In a separate analysis, E. H. Budiarto et al. [13] assessed three anomaly detection techniques – K-Means, LOF, and OC-SVM – on a pharmaceutical dataset, analyzing both computation time and memory utilization. Their results revealed that OC-SVM detected three outliers with a processing time of just 0.2 microseconds. Similarly, S. Behera et al. [14] benchmarked various density-based methods (LOF, OPTICS, DBSCAN, and DENCLUE) using a breast cancer dataset. The study focused on the accuracy of noise detection and execution time, ultimately identifying OPTICS as the most efficient in estimating outlier quantities within a short time frame.

S. Luan et. al. [5] investigated the detection of out-of-distribution (OOD) samples within deep neural networks (DNNs). Their approach involved runtime monitoring using IF and LOF, applied to well-known datasets such as MNIST and GTSRB. Evaluation was performed using standard metrics like precision, recall, F1 score, and overall accuracy. The results suggest that IF offers both strong predictive performance and low computational overhead. Another contribution by A. Belhadi et. al. [15] compared trajectory-based anomaly detection techniques across small, medium, and massive datasets. Their taxonomy divided the techniques into density-based and clustering-based methods. The framework yielded promising results for datasets of limited and moderate sizes, although it was computationally intensive for larger-scale databases.

This study aims to compare one representative anomaly detection method from each core category – IF (cluster-based), OC-SVM (distance-based), and LOF (density-based). Performance evaluation was carried out using two key indicators: the ROC curve and rank-based precision scores. These were computed based on the raw outlier scores generated by the detection algorithms. IF was selected due to its robustness against common issues such as masking and swamping, as well as its scalability. OC-SVM, known for training on single-class data distributions, is particularly effective on large volumes of data. Lastly, LOF is recognized for outperforming global techniques in detecting anomalies with minimal reliance on thresholds, as it effectively identifies subtle deviations within dense clusters.

# METHOD

In this paper, three different techniques from each approach have been applied, which are Local Outlier Factor (density-based), One Class Support Vector Machine (distance-based), and Isolation Forest (cluster-based), and five different datasets will be used, which are Cervical Cancer, Diabetics, Maternal Health Risk, Caesarean, and Dengue datasets. The techniques will be evaluated based on Receiver Operating Characteristics (ROC), which is used to evaluate the classifier output quality, and precision @ rank n value, which is the raw outlier scores as returned by a fitted model.

## Technique

### *Local Outlier Factor (LOF)*

LOF belongs to the family of density-oriented anomaly detection methods. It determines the degree to which a data point deviates from its surrounding data by evaluating the local density of nearby observations. If a data instance resides in a region that is significantly less dense compared to its neighbors, it is flagged as an anomaly. The method operated by comparing each point’s local density to that of its k-nearest neighbors [5]. The following steps outline how the LOF score is computed.

* + 1. *k-Distance:*

For a given point *xp ϵ X, where p = 1, 2, …, N*, the *k*-distance *dk (xp)* is defined as the smallest distance such that: at least *k* points *xq ϵ X\{xp}* are no farther than *dk (xp)*; and fewer than *k* points are strictly closer than this distance. This threshold is used to determine the neighbouring region for each point *xp*.

* + 1. *k-Distance Neighbourhood*

Based on the k-distance, the neighbourhood comprises all points  *such that their distance to is less than or equal to ,* as defined in Equation (1):

(1)

* + 1. *Reachability Distance*

The reachability distance between any two points and is calculated using Equation (2):

(2)

* + 1. *Local Reachability Distance*

Local reachability density quantifies how densely a point is surrounded by its neigbors. It is computed as the inverse of the average reachability distance within its k-distance neighborhood, as defined in Equation (3):

(3)

* + 1. *Local Outlier Factor*

The LOF score indicated how isolated a point is with respect to its neighbors. It is computed as the ratio between the average LRD of its neighbours and its own LRD, as defined in Equation. (4):

(4)

*LOFk (xp) < 1* means *xp* has higher density than its neighbours (not outlier); *LOFk (xp) > 1* means *xp* has lower density than its neighbours (outlier).

### *One Class Support Vector Machine (OCSVM)*

The OC-SVM technique operated by mapping the data such that most points are enclosed within a region separated from the origin in feature space. It aims to find a boundary that distinguished the normal observations from all others by maximizing the separation distance from the origin [27]. This is done by constructing a function , which returns +1 for observations within the majority region and -1 for points considered outside. The formulation of this decision boundary is achieved through a constrained quadratic optimization approach, as defined in Equation (5):

(5)

Where: and is a slack variable, and subject to:

### *Isolation Forest (IF)*

Isolation Forest belongs to the family of anomaly detection techniques that isolated data instances using recursive partitioning of randomly selected features. Unlike density-based approaches, this method is based on the premise that anomalies are easier to isolate than normal points, due to requiring fewer splits in the partitioning process [5]. In this algorithm, random sub-samples of the dataset are processed into trees, where points that are isolated using shorter paths are flagged as anomalies. In contrast, those requiring deeper traversal into the tree are considered normal.

Given a dataset with n instances, the anomaly score *s(xp)* for a sample point *xp* is calculated as Equation (6):

(6)

Here, *h(xp)* represents the path length for a point within a single tree, and *E (h(xp))* is the expected path length averaged across all trees. The harmonic number *H (i)* used in normalizing, is estimated by *H (i) = ln (i)* + 0.5772. A higher anomaly score corresponds to a greater likelihood of the sample being an outlier.

## Datasets

Four datasets used in this study are from UCI repository which are Cervical Cancer, Diabetics, Maternal Health Risk and Caesarean datasets while Dengue dataset are obtained from Seremban Health Organization. Each dataset has different number of records and attributes as stated in Table 1. Cervical Cancer dataset contains 20 attributes regarding on cervix behaviour risk with 72 patients’ with and without cancer instances. As for Diabetics dataset that includes 9 attributes with 768 patients’ record which contain related diabetes diseases criteria such as BMI, glucose and insulin, Maternal Health Risk dataset consist of 678 records and 7 attributes from different hospitals and clinics through the IoT based risk monitoring system. Caesarean dataset contains information about 80 pregnant women and 6 attributes that considering about the most important characteristics of delivery problem. Lastly, dengue dataset is given by Seremban Health Organization for research purposes, it contains 120 instances and 6 attributes on dengue patients’ and climate data. It is important to note that all five datasets used in this study are unlabelled and do not explicitly contain ground-truth outlier labels. Therefore, the number of actual outliers is unknown, which aligns with the nature of unsupervised outlier detection tasks. The models are evaluated based on their ability to detect anomalies inferred from patterns in the data.

|  |  |  |
| --- | --- | --- |
| **TABLE 1**. Detail description of datasets used | | |
| **Dataset** | **Records** | **Attributes** |
| Cervical Cancer | 72 | 20 |
| Diabetics | 768 | 9 |
| Maternal Health Risk | 678 | 7 |
| Caesarean | 80 | 6 |
| Dengue | 120 | 6 |

## Evaluation Metrics

In order to measure the quality of outlier detection techniques that are used, this research used ROC and Precision @ rank n value.

### *Receiver Operating Characteristics (ROC)*

The use of ROC has become increasingly ubiquitous in a huge set of application areas, including machine learning [6]. Therefore, this experiment will use ROC as the evaluation metric as ROC evaluates the trade-off between true positive rate and false positive rate, helping to measure the model’s ability to correctly distinguish outliers from normal points as calculated using Equations (7) – (11). A higher ROC value indicates better separation.

(7)

(8)

(9)

(10)

(11)

### *Precision @ Rank n*

Precision @ Rank-n evaluates how many true outliers are ranked among the top-n predictions. This is especially useful when the number of outliers is small or unknown, as it reflects the accuracy of the algorithm in identifying the most anomalous points.

In pyOD, it used pyod.utils.utility.precision\_n\_scores (y, y\_pred, n = None) to calculate precision @ rank n where parameters as listed below [28].

Parameters:

* Y (list or numpy array of shape (n\_samples,)) – The ground truth. Binary (0: inliers, 1: outliers).
* Y\_pred (list or numpy array of shape (n\_samples,)) – The raw outlier scores as returned by a fitted model.
* N(int, optional (default = None)) – The number of outliers. If not defined, infer using ground truth.

## Evaluation Platform

In this study, we used Unified Python Library for Outlier Detection (pyOD). PyOD library is one of the Python libraries that comes with a set of scalable, state-of-the-art algorithms to detect outliers in a dataset [28]. It contains more than 30 outlier detection algorithms which are very useful and made it highly cited in the literature from other researchers. Besides that, it was extremely straightforward to use as it comes with aunified API for all algorithms, technical documentation and examples. In pyOD, all detectors are initialized with a contamination parameter which by default, it is set with 0.1. Contamination is the expected proportion in the dataset and used to set the outlier score threshold on model fitting. Contamination must be set between 0 to 0.5 only. In this study, we used default contamination value which is 0.1 for all 3 outlier detection technique on all 5 datasets.

# RESULTS AND DISCUSSION

In this study, LOF, OC-SVM, and IF were selected as representatives of three major categories of outlier detection: density-based, distance-based, and cluster-based. These techniques are widely used in literature and supported by the PyOD library. Although no universal benchmark exists for unsupervised outlier detection due to the absence of ground truth, the use of ROC and P@N has been adopted in prior works [5], [6], [28]. The comparison offers insight into how these representative methods perform across diverse small-scale health-related datasets, under a unified experimental setting.

Table 2shows the comparison of the ROC and P@N for the LOF, OC-SVM and IF techniques using different dataset. As a final result, it shows that IF gives the best ROC and precision @ rank n value as the score is the highest for 3 different datasets with an average of 0.7297 and 0.5975 for ROC and P@N score respectively compared to LOF and OC-SVM which is 0.5473, 0.3916 and 0.4118. 0.4236 respectively. As all of our dataset are unsupervised and small dataset, it proven that IF is the best outlier detection technique for these comparative experiments as it provides a high accuracy for abnormal detection under unlabelled data and small dataset as supported by [10] and IF works well on multi-dimensional data which it has a linear time complexity and fast outlier detection [11].

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TABLE 2**. Comparison table | | | | | | |
| **Dataset** | **LOF** | | **OCSVM** | | **IF** | |
| **ROC** | **P@N** | **ROC** | **P@N** | **ROC** | **P@N** |
| Cervical Cancer | 0.6818 | 0.25 | 0.8864 | 0.75 | 0.8182 | 0.5 |
| Diabetics | 0.5052 | 0.3148 | 0.2881 | 0.1667 | 0.5944 | 0.4259 |
| Maternal Health Risk | 0.3717 | 0.2545 | 0.1194 | 0.1455 | 0.8254 | 0.6727 |
| Caesarean | 0.5873 | 0.5556 | 0.3968 | 0.5556 | 0.9524 | 0.8889 |
| Dengue | 0.5903 | 0.5833 | 0.3681 | 0.5 | 0.4583 | 0.5 |
| Average | 0.54726 | 0.39164 | 0.41176 | 0.42356 | 0.72974 | 0.5975 |

# CONCLUSION

This study performs a comparison between LOF, OC-SVM and IF for 5 different datasets, which are cervical cancer, diabetics, maternal health risk, caesarean and dengue dataset to identify which outlier detection technique gives the best result. This study installed pyOD library and run in Python using Google Colab platform. In pyOD, ROC and precision @ rank n value are used for performance evaluation. As the result, this study has proved that IF technique gives the best result for 3 different datasets out of 5 with the highest average values for ROC and precision @ rank n value. For future work, we are working on the integration of IF with clustering method in order to improve the clustering method by removing outliers using IF.

# ACKNOWLEDGEMENTS

This research was supported by Faculty of Information Technology and Communication and Universiti Teknikal Malaysia Melaka. Special thanks to Dr. Noraswaliza Abdullah from UTeM for advice on the study areas. Special thanks to data donator from UCI Repository for the free access dataset and Seremban Health Organization for the dengue dataset.

# REFERENCES

1. J. Zhu, Y. Wang, D. Zhou, and F. Gao, “Batch Process Modeling and Monitoring With Local Outlier Factor,” IEEE Trans. Contr. Syst. Technol. **27**(4), 1552–1565 (2019).

2. Y. Wu, X. Li, F. Luan, and Y. He, “A Novel GPR-Based Prediction Model for Strip Crown in Hot Rolling by Using the Improved Local Outlier Factor,” IEEE Access **9**, 458–469 (2021).

3. et al D. Motwani, “Enhancing Intrusion Detection Systems with a Hybrid Deep Learning Model and Optimized Feature Composition,” International Journal on Recent and Innovation Trends in Computing and Communication **11**(10), 369–380 (2023).

4. J. Qin, Y. Yang, C. Gu, Z. Hong, and H. Du, “Outlier Detection for Transformer’s Oil Chromatographic Data Based on Metric Learning and the Weighted Local Outlier Factor,” in *2019 6th International Conference on Systems and Informatics (ICSAI)*, (IEEE, Shanghai, China, 2019), pp. 370–375.

5. S. Luan, Z. Gu, L.B. Freidovich, L. Jiang, and Q. Zhao, “Out-of-Distribution Detection for Deep Neural Networks With Isolation Forest and Local Outlier Factor,” IEEE Access **9**, 132980–132989 (2021).

6. S. Urbano, E. Chaumette, P. Goupil, and J.-Y. Tourneretr, “On the High-Snr Receiver Operating Characteristic of Glrt for The Conditional Signal Model,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, Calgary, AB, 2018), pp. 3939–3943.

7. H. Chen, S. Zhao, H. Bao, and H. Kang, “Research of local outlier mining algorithm based on spark,” in *2017 First International Conference on Electronics Instrumentation & Information Systems (EIIS)*, (IEEE, Harbin, 2017), pp. 1–4.

8.  H.C. Mandhare, and S.R. Idate, “A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques,” in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, (IEEE, Madurai, 2017), pp. 931–935.9. Y. Liu, R. Yuan, S. Zheng, K. Yan, and H. Miao, “An Abnormal Detection of Positive Active Total Power Based on Local Outlier Factor,” in *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*, (IEEE, Shenyang, China, 2021), pp. 180–183.

10.  S. Zhong, S. Fu, L. Lin, X. Fu, Z. Cui, and R. Wang, “A novel unsupervised anomaly detection for gas turbine using Isolation Forest,” in *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, (IEEE, San Francisco, CA, USA, 2019), pp. 1–6.

11. X. Chun-Hui, S. Chen, B. Cong-Xiao, and L. Xing, “Anomaly Detection in Network Management System Based on Isolation Forest,” in *2018 4th Annual International Conference on Network and Information Systems for Computers (ICNISC)*, (IEEE, Wuhan, China, 2018), pp. 56–60.

12. S. Liu, Z. Ji, and Y. Wang, “Improving Anomaly Detection Fusion Method of Rotating Machinery Based on ANN and Isolation Forest,” in *2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, (IEEE, Chongqing, China, 2020), pp. 581–584.

13. E.H. Budiarto, A. Erna Permanasari, and S. Fauziati, “Unsupervised Anomaly Detection Using K-Means, Local Outlier Factor and One Class SVM,” in *2019 5th International Conference on Science and Technology (ICST)*, (2019), pp. 1–5.

14. S. Behera, and R. Rani, “Comparative analysis of density based outlier detection techniques on breast cancer data using hadoop and map reduce,” in *2016 International Conference on Inventive Computation Technologies (ICICT)*, (IEEE, Coimbatore, India, 2016), pp. 1–4.

15. A. Belhadi, Y. Djenouri, and J.C.-W. Lin, “Comparative Study on Trajectory Outlier Detection Algorithms,” in *2019 International Conference on Data Mining Workshops (ICDMW)*, (IEEE, Beijing, China, 2019), pp. 415–423.

16. W.S. Al Farizi, I. Hidayah, and M.N. Rizal, “Isolation Forest Based Anomaly Detection: A Systematic Literature Review,” in *2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, (IEEE, Semarang, Indonesia, 2021), pp. 118–122.

17. Z. Ding, and M. Fei, “An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window,” IFAC Proceedings Volumes **46**(20), 12–17 (2013).

18. D. Xu, Y. Wang, Y. Meng, and Z. Zhang, “An Improved Data Anomaly Detection Method Based on Isolation Forest,” in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, (IEEE, Hangzhou, 2017), pp. 287–291.

19. H. Xiang, J. Wang, K. Ramamohanarao, Z. Salcic, W. Dou, and X. Zhang, “Isolation Forest Based Anomaly Detection Framework on Non-IID Data,” IEEE Intell. Syst. **36**(3), 31–40 (2021).

20. R. Alsini, A. Almakrab, A. Ibrahim, and X. Ma, “Improving the outlier detection method in concrete mix design by combining the isolation forest and local outlier factor,” Construction and Building Materials **270**, 121396 (2021).

21. S. Liu, Z. Ji, and Y. Wang, “Improving Anomaly Detection Fusion Method of Rotating Machinery Based on ANN and Isolation Forest,” in *2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, (IEEE, Chongqing, China, 2020), pp. 581–584.

22. A. Mensi, and M. Bicego, “Enhanced anomaly scores for isolation forests,” Pattern Recognition **120**, 108115 (2021).

23. Y. Huang, Y. Xue, Y. Su, and S. Han, “Hyperspectral Anomaly Detection Based on Isolation Forest with Band Clustering,” in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, (IEEE, Waikoloa, HI, USA, 2020), pp. 2416–2419.

24. S. Buschjager, P.-J. Honysz, and K. Morik, “Generalized Isolation Forest: Some Theory and More Applications Extended Abstract,” in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, (IEEE, Sydney, Australia, 2020), pp. 793–794.

25. C. Tao, T. Li, and J. Huang, “Kernel Choice in One-Class Support Vector Machines for Novelty and Outlier Detection,” in *2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, (IEEE, Taiyuan, China, 2020), pp. 116–120.

26. M.E. Villa-Pérez, M.Á. Álvarez-Carmona, O. Loyola-González, M.A. Medina-Pérez, J.C. Velazco-Rossell, and K.-K.R. Choo, “Semi-supervised anomaly detection algorithms: A comparative summary and future research directions,” Knowledge-Based Systems **218**, 106878 (2021).

27. C. Tao, T. Li, and J. Huang, “Kernel Choice in One-Class Support Vector Machines for Novelty and Outlier Detection,” in *2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, (IEEE, Taiyuan, China, 2020), pp. 116–120.

28. H. Alimohammadi, and S. Nancy Chen, “Performance evaluation of outlier detection techniques in production timeseries: A systematic review and meta-analysis,” Expert Systems with Applications **191**, 116371 (2022).