ISSN: 2093-5374 / E-ISSN: 2093-5382

Efficient Outlier Detection in High-Dimensional Data Using Unsupervised Machine Learning

Girish Reddy Ginni[1*], and Dr. Srinivasa L. Chakravarthy[2]

[1*]Department of CSE, GITAM University, Gandhi Nagar, Rushikonda, Visakhapatnam, Andhra Pradesh, India. girishloshankar@gmail.com, https://orcid.org/0009-0005-5242-8839

[2]Department of CSE, GITAM University, Gandhi Nagar, Rushikonda, Visakhapatnam, Andhra Pradesh, India. chakri.ls@gmail.com, https://orcid.org/0000-0001-9141-4863

Received: July 26, 2024; Revised: September 05, 2024; Accepted: October 03, 2024; Published: December 30, 2024

Abstract

A fundamental concept in data mining and ML is outlier detection. Outlier identification and clustering often work together, as identifying outliers can lead to better clustering. Most current research projects have focused primarily on outlier identification and clustering as separate aspects, but their close relationship needs to be explored. By considering this relationship, we can improve cluster quality while detecting outliers, providing dual benefits. We have proposed an unsupervised ML framework for efficiently detecting outliers in high-dimensional datasets. An objective function has been defined to enhance cluster compactness, which improves the outlier detection process. By improving the clustering process through problem transformation and enhanced K-Means, we can develop an integrated approach that achieves high-quality clustering and outlier identification simultaneously. We have introduced an algorithm called Learning-based Outlier Detection (LbOD), which is novel in its simultaneous approach to partition space, objective function, and cluster optimization. A prototype has been built to evaluate the proposed framework and algorithm's ability to discover outliers using multiple benchmark high-dimensional datasets. Our empirical study has shown that the LbOD algorithm outperforms many existing outlier detection methods.

Keywords: Outlier Detection, Clustering, Unsupervised Learning, Machine Learning, High Dimensional Data.

1 Introduction

In many real-world applications, large volumes of data to be processed. The dataset contains data points represented and used appropriately for deriving business intelligence. However, some data points may be abnormal compared to others. Such data points are known as outliers, and detecting them has many valuable applications. Outlet detection research has been conducted using various methods, such as heuristic and learning-based methods. With the emergence of AI, the usage of ML is increasing to solve problems in applications of different domains. In this context, outlier detection helps solve problems and improves data quality for machine learning and other data-driven applications. Literature has rich information about various heuristic and other outlier detection methods.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 15, number: 4 (December), pp. 192-212. DOI: *10.58346/JOWUA.2024.I4.013*

*Corresponding author: Department of CSE, GITAM University, Gandhi Nagar, Rushikonda, Visakhapatnam, Andhra Pradesh, India.

There is a connection between clustering and outlier identification. Enhancing cluster validity and outlier identification, the COR method effectively combines both objectives (Liu et al., 2019). Hilal et al., (2022) recently focused on anomaly detection as unsupervised models. The rise of financial fraud is challenging current mechanisms. Because financial crime presents serious risks, fraud detection technologies are constantly being improved. Machine learning approaches such as regression for detection, grouping, and classification are highlighted in recent publications (Sadgali et al., 2019). Meng et al., (2019) provided ideas for future studies by reviewing trajectory outlier identification systems based on multi-attribute representation, distance measurements, and algorithm improvements. Erhan et al., (2021) examined anomaly detection in sensor systems, classifying approaches into data-driven and traditional categories while considering topologies such as Cloud, Fog, and Edge. It draws attention to obstacles and practical solutions. Credit scoring, a hybrid ensemble model that combines balanced sampling with voting-based outlier detection, performs better. Outperforming benchmark models, the model tackles unbalanced data difficulties and outlier adaptation (Zhang et al., 2021). Dhiman et al., (2021) propose detecting anomalies in wind turbine gearboxes using SCADA data, and an adaptive threshold and TWSVM approach are suggested. Outcomes demonstrate better performance compared to baseline classifiers. Avci et al., (2021) examined and contrasted ML and DL techniques based on vibration for structural damage detection (SDD). ML techniques that concentrate on feature extraction and classification perform better than conventional ones (Thomas et al., 2024). Many existing methods dealing with outlier detection, including machine learning techniques, have shown deteriorated performance for many reasons. Moreover, there is an issue with the model's scalability besides its accuracy. Our contributions to this paper are listed below.

1. We proposed an unsupervised ML framework to detect outliers from high-dimensional datasets efficiently.
2. We proposed a learning-based Outlier Detection (LbOD) algorithm, which is novel in its simultaneous approach to partition space, objective function, and cluster optimization (Giji Kiruba et al., 2023).
3. A prototype is built to assess the suggested algorithm and framework's ability to discover outliers in multiple benchmark high-dimensional datasets.

This is the format for the rest of the paper. Section 2 examines the most current research on several techniques for detecting outliers. Section 3 presents an automated ML-based approach for identifying outliers in high-dimensional data. Conversely, Section 4 showcases the findings from our empirical investigation using many high-dimensional datasets. In addition to outlining potential directions for further study, Section 5 presents the results drawn.

2 Related Work

This section examines current research on a range of techniques of outlier detection. Yang et al., (2023) observed that group strangers are missed by conventional outlier detectors concentrating on individual items. The NR framework improves performance of current detectors by utilizing representative objects. Mensi et al., (2023) found that outliers to other data points are used to identify them. To find outliers based on pairwise distance, Proximity Isolation Forest expands upon the Isolation Forest. Enes et al., (2023) explored and said that time series are essential for many applications. This study presents a pipeline for grouping multivariate time series to facilitate anomaly identification. Qais et al., (2024) used a clustering approach using K-means and fuzzy c-means for outlier identification; induction heating

safety was improved, and 96% accuracy was attained. Chen et al., (2020) regulated Biofeedback is included in severe games to help with good player behavior regulation and mental stress reduction.

Liu et al., (2019) found a connection between clustering and outlier identification. The COR method combines both objectives to enhance cluster validity and outlier identification. Carreño et al., (2020) observed that unclosed impedes research in rare events, anomalies, novelty, and outlier identification. In this study, standardization is suggested. Hilal et al., (2022) focused on anomaly detection in recent times as unsupervised models. The rise of financial fraud is challenging current mechanisms. Brito et al., (2022) utilized unsupervised techniques and SHAP for explainability, and a novel approach to defect identification and diagnosis for rotating equipment is proposed. Sadgali et al., (2019) found that fraud detection technologies constantly improve because financial crime presents a severe risk. Machine learning approaches such as regression for detection, grouping, and classification are highlighted in recent publications.

Stetco et al., (2019) examined artificial intelligence models for monitoring wind turbine status, validation, and data source categorization and regression. Model optimization and dataset problems are areas of future development. Meng et al., (2019) provided ideas for future studies by reviewing trajectory outlier identification systems based on multi-attribute representation, distance measurements, and algorithm improvements. Raghavan & El Gayar, (2019) showed how to overcome obstacles and successfully implement a robotic instrument for mTBI patient evaluation in an urban ED. Bashar & Nayak, (2020) observed that standard and neural network models outperform TAnoGan, a GAN-based technique for anomaly identification in time series with sparse data (Khedr et al., 2020). Erhan et al., (2021) examined anomaly detection in sensor systems, classifying approaches into data-driven and traditional categories while considering topologies such as Cloud, Fog, and Edge. It draws attention to obstacles and practical solutions.

Crimi et al., (2018) compared the effectiveness of deep spatial autoencoders to patch-based approaches in the unsupervised brain MR image anomaly identification process. Latent space constraints and absence of adversarial training requirements (Mallikarjuna & Rao, 2019). Accurate and swift segmentations between slices point to potential uses as previous knowledge and in unsupervised lesion segmentation. Ruff et al., (2021) developed deep learning for anomaly detection to enhance the detection of complicated datasets, bringing methods together and examining relationships between traditional and deep techniques. Zhang et al., (2021), a hybrid ensemble model that combines balanced sampling with voting-based outlier detection, performs better for credit scoring. Outperforming benchmark models, the model tackles unbalanced data difficulties and outlier adaptation. Dhiman et al., (2021) propose detecting anomalies in wind turbine gearboxes using SCADA data, an adaptive threshold and TWSVM approach is suggested. Outcomes demonstrate better performance compared to baseline classifiers. Avci et al., (2021) examined and contrasted ML and DL techniques for structural damage detection (SDD), which was determined by vibration. ML techniques that concentrate on feature extraction and classification perform better than conventional ones.

Yang et al., (2021) addressed the issue of outlier pollution in conventional approaches, and a mean-shift outlier detector is proposed. The mean-shift approach enhances performance in outlier identification tasks by eliminating the bias associated with outliers. Alaettin & Volkan, (2021) process is gaining popularity as more and more devices are connected and produce constant data streams. Accuracy, complexity, and primary method of recent algorithms are examined; popular tools and open problems are also covered. Chakraborty et al., (2019) suggested using ensemble probabilistic neural networks and stacked autoencoders to solve situations involving numerous outliers and class imbalance. Future research aims to expand to unsupervised techniques for various kinds of outliers. Thangaramya

et al., (2020) presented a novel secure routing method, FRCSROD, for WSNs that use outlier detection and fuzzy criteria. FRDOA enhances energy efficiency, dependability, and security by identifying hostile nodes (Belhadi et al., 2021) presented for discerning anomalous human conduct from pedestrian data in smart cities. In less than 50 seconds, deep learning achieves 88% accuracy compared to data mining.

Landauer et al., (2020) observed that, though their massive, unstructured data makes them challenging to analyze, log files are essential for cyber security. Logging techniques and goals are reviewed in this paper's classification of log clustering approaches. Djenouri et al., (2019) examined the detection of outliers in urban traffic, classifying techniques into flow and trajectory detection. Different strategies are spoken about, emphasizing patterns. Tang et al., (2019) approach to multi-kernel SVM with K-means clustering for large-scale data categorization is presented (Santhosh & Prasad, 2023). The process chooses typical examples, decreases the amount of human labeling, and greatly increases accuracy and efficiency. Munoz-Organero, (2019) stated that a unique method integrates Human Activity classification (HAR) with DRNN for outlier identification and sub-activity classification. The approach is tested in many settings and yields encouraging outcomes. Chen et al., (2021) experimented on actual datasets to demonstrate that LRTG outperforms existing techniques. Adaptive neighbors, l2, 1-norm, and Tucker decomposition are integrated in a unique multi-view clustering technique called LRTG.

Fitriyani et al., (2020) used the Cleveland and Statlog datasets; the HDPM achieved accuracy rates of 98.40% and 95.90%. The goal of HDCDSS is to enhance early detection of cardiac disease. Rogers et al., (2019) offered a framework for structural health monitoring (SHM) using non-parametric grouping based on Bayesian principles. This method adjusts live, provides excellent accuracy, and does not require pre-collected training data (Zigui et al., 2024). Population-level applicability will be included in future development. Thöle et al., (2019) indicated dust and export production flows close to the Kerguelen Plateau, reducing export production during glaciers. During interglacials, the Antarctic Zone shows increased export production, highlighting Fe fertilization and changes in the Southern Ocean's upwelling. Fitriyani et al., (2019) suggested using ensemble learning, iForest, and SMOTETomek to create a DPM regarding hypertension and type 2. On four datasets, the DPM achieved good accuracy. Deepak et al., (2021) provided an autoencoder variant that outperforms current techniques for identifying surveillance footage abnormalities.

Mishra & Pandya, (2021) expanded of IoT presents security threats because of power and cost limitations, particularly the potential for DDoS assaults. Future security initiatives and intrusion detection models are examined in this study. Kraus et al., (2019) examined the detection of clusters in scatterplots on 2D, 3D, and virtual reality displays. Better overview was achieved with restricted VR regions, while scatterplot representations benefited from 3D VR's increased memory and orientation for cluster recognition. Liu et al., (2019) presented SO-GAAL, a method for detecting outliers that directly generates prospective outliers to overcome high-dimensional data sparsity. Performance is further improved by expanding to MO-GAAL with numerous generators, especially on various datasets—subsequent research endeavors to incorporate group learning for stability and investigate distinct network configurations for various data kinds. Wang et al., (2019) approach to detecting outliers highlights their advantages and disadvantages and suggests areas for further study to be improved. Usama et al., (2019) examined the growing field of unsupervised ML in networking and described its uses, including anomaly detection and traffic engineering. It highlights difficulties and potential research paths while offering insights into current advances. Many existing methods dealing with outlier detection, including machine learning techniques, have shown deteriorated performance for many reasons. Moreover, there is an issue with the model's scalability besides its accuracy.

3 Preliminaries

An overview of K-means and entropy is given in this section. A variant known as K-means (Ruff et al., 2021) was created to address K-means' susceptibility to outliers. Few outliers are known to diverge the centroids from their inherent locations. To address this, specific data points distant from their centroids are considered outlier candidates. These data points are not given a cluster name or updated centroidally. K-means assigning data points and updating the centroid are two similar iterative phases. We determine the separations between every data point and the closest centroid throughout the data point assignment process. We then rank the distances, identifying potential outliers as the data points with the highest or lowest distances. Since these outlier candidates aren't given cluster names, K-means updates the centroid similarly. It is noteworthy that during the iteration, the outlier candidates are evolving. In contrast, K-means requires the number of clusters as well as the K and o outliers as input parameters. Regarding its clean mathematical formulation, convergence, and model effectiveness, it has many qualities that are similar to K-means. Dhiman et al., (2021) clarifies that entropy or total correlation alone is insufficient for outlier spotting. They put forth the following new measure of holoentropy. In holoentropy, the overall relationship between the random vector Y and its entropy are added together to form the holoentropy HL(Y), which may be stated as the sum of the entropies for all characteristics. Based on information theory, holoentropy is an outlier identification measure that handles categorical data and accounts for total correlation and entropy (Chawla & Gionis, 2013). For a tidy and effective solution, we derive our suggested goal function, which is based on holoentropy associated with K-Means algorithm.

4 Proposed Methodology

The suggested approach for effectively identifying outliers in high-dimensional data is presented in this section. The underlying algorithm, processes, and recommended outlier identification framework are all part of the technique.

4.1. Problem Definition

Providing a high dimensional dataset proposing a machine learning-based framework to detect outliers efficiently is the challenging problem.

4.2. The Proposed Outlier Detection Framework

The unsupervised machine learning model is the foundation for the suggested outlier identification approach. The framework's architecture allows it to process high-dimensional data as input and produce output as identified outliers. A particular data city is put through an audition process to use clustering to exploit a methodology. The clustering method divides data points into a number of categories. Strange values among data points that are dissimilar to other values are known as outliers. There are many applications linked to outlier detection. The outlier detection methods help solve many real-time problems.

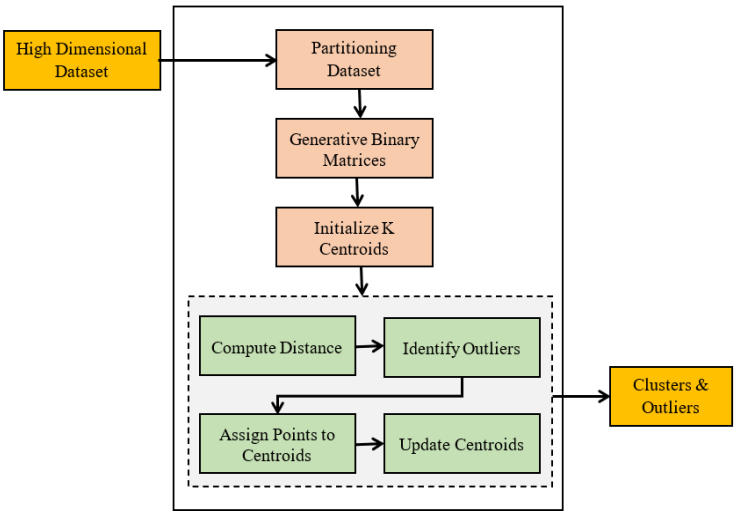

Figure 1: Overview of the Proposed Outlier Detection of Framework

Sometimes, outlier detection itself is useful in applications like credit code fraud detection. Figure 1 shows the proposed framework designed for automatically finding outliers in large-scale data collections.

4.3. Objective Function

Outlier identification and cluster analysis are closely related activities. A few outlier points may quickly destroy a cluster's structure; On the other hand, outliers are defined by the cluster concept and are points that do not belong to any cluster. To address this difficulty, we concentrate on the proposed approach's clustering-based outlier detection. In particular, after identifying o points as outliers, the remaining occurrences are split into K clusters by parallel performing outlier identification and clustering tasks. Corresponding symbols used in the next sections are displayed in Table 1. A small number of outliers might undermine the cluster structure, and these outliers want to be recognized by the cluster boundary. It is similar to a chicken-and-egg dilemma due to the coupling relationship between outlier identification and cluster analysis. We draw inspiration from consensus clustering (Chakraborty et al., 2019), which combines many fundamental divisions created to prevent the circular dependency problem in a combined approach considering outlier detection and clustering; the data should be thoroughly fused to minimize the bad effects of outliers. Furthermore, the clusters are necessary for the definition of outliers. The two considerations above encourage us to create several simple partitions to convert the information into partition space from the original feature space. This procedure is comparable to consensus clustering's fundamental partition generation approach (Wu & Wang, 2011; Strehl & Ghosh, 2002). With n points and d features, let X be the data matrix. When X is divided into K distinct clusters, it may be shown as a set of K object subsets, including a vector label $\pi = (L_\pi(x_1),\cdots, L_\pi(x_n))$, where x_1 is mapped by $L_\pi(x_1)$ a label in K. The r fundamental partitions $\Pi = \{\pi_i\}, 1 \le i \le r$, maybe obtained by applying certain basic partition generation strategies, including K-means clustering with varied cluster counts. For π_i, let K_i denotes cluster number and let $R = \sum_{i=1}^{r} K_i$. Then, using Π, the following B = {b_l},1 < l ≤ n, binary matrix, may be obtained in Eq. (1):

$$b_l = \left(b_{l,1}, \dots, b_{l,i}, \dots, b_{l,r}\right), \quad \text{with}$$
$$b_{l,i} = \left(b_{l,i1}, \dots, b_{l,ij}, \dots, b_{l,iKi}\right), \quad \text{and} \qquad (1)$$

$$b_{l,ij} = \begin{cases} 1, & if\ L_{\pi_i}(x_l) = j \\ 0, & otherwise \end{cases}$$

It is important to remember that creating fundamental divisions does not need a particular method. To construct primary partitions, K-means with varying cluster counts are advised for efficiency and simplicity. Based on the fundamental divisions produced by K-means, our proposed approach nonetheless yields encouraging results despite K-means' susceptibility to outliers. The binary value represents the information unique to cluster membership created by the definition of outliers. Because of these two characteristics, the binary space is preferable to the continuous space because it facilitates the identification of outliers due to its categorical traits. For instance, Holoentropy is a widely used metric for identifying outliers in categorical data (Dhiman et al., 2021).

The authors of (Dhiman et al., 2021) sought to reduce the dataset's Holoentropy after removing the outliers. Here, we're assuming that the entire dataset has a cluster structure. As a result, minimizing each cluster's Holoentropy makes more sense. In this method, instead of the complete dataset becoming compact once the outliers are eliminated, the clusters do. Therefore, we provide our intended purpose for the proposed approach, depending on the Holoentropy of each cluster.

$$\min_{\pi} \sum_{k=1}^{K} p_k HL(C_k), \tag{2}$$

Where π is the cluster indicator and $p_{k+} = |C_k|/(n - o)$. HL(•) referred in a definition earlier covers K clusters C1∪•••∪C_K = X\O, with $C_k \cap C_{k'} = \emptyset$ if $k \neq k'$. In Eq. 2, the objective function is the weighted Holoentropy linked to every cluster and p_k is based on size of the cluster. Beyond this work, we think of discovering K and o is associated with an orthogonal issue. Here, the variables of our suggested method—which uses the same setup as K-means—are the K and o denoting the number of clusters and outliers, respectively (Ruff et al., 2021). The following section provides an efficient solution by addressing the issue expressed in Eq. (2) by providing an objective function based on the binary matrix as described below.

$$\sum_{k=1}^{K} p_k HL(C_k) \propto \sum_{k=1}^{K} \sum_{b_l \in C_k} \sum_{i=1}^{r} \sum_{j=1}^{K_i} H\left(C_{k,ij}\right), \text{ and}$$

$$H\left(C_{k,ij}\right) = -\left(1 - p_{k,ij}\right) log\left(1 - p_{k,ij}\right) - p_{k,ij} \log p_{k,ij}, \tag{3}$$

Where the Shannon entropy is denoted by H and the probability that $b_{l,ij} = 1$ in the ij-th column of C_k is represented by $p_{k,ij}$. We offer the following lemma to clarify the meaning of $p_{k,ij}$ in Eq. (3).

Lemma 1. In K-Means clustering applied on the binary dataset, k-th centroid is computed to satisfy:

$$m_k = \left(m_{k,1}, \ldots, m_{k,i}, \ldots, m_{k,r}\right), \text{ with}$$

$$m_{k,i} = \left(m_{k,i1}, \ldots, m_{k,ij}, \ldots, m_{k,iK_i}\right), \text{ and} \tag{4}$$

$$m_{k,ij} = \sum_{b_{l,ij} \in C_k} \frac{b_{l,ij}}{|C_k|} = p_{k,ij}, \forall\ k, i, j.$$

Lemma 1's proof is clear from the centroid's arithmetic mean in the clustering process. Thus, Lemma 1 leads us to conclude that the issue linked to Eq. (3) is related to the clustering process.

Theorem 1. When applying K-means to B's $n - o$ inliers, we obtain

$$\max \sum_{k=1}^{K} p_k \sum_{i=1}^{r} \sum_{j=1}^{K_i} p_{k,ij} \log p_{k,ij} \Leftrightarrow min \sum_{k=1}^{K} \sum_{b_l \in C_k} f(b_l, m_k), \tag{5}$$

Where m_k is the $k-$th centroid as expressed in Eq. (4) while the distance function $f(b_l, m_k) = \sum_{i=1}^{r} \sum_{j=1}^{K_i} D_{KL}\left(b_{l,ij} \| m_{k,ij}\right)$, here $D_{KL}(\cdot \,||\, \cdot)$ is the KL-divergence.

Proof. The Bregman divergence (Liu et al., 2016) indicates that we have $D_{KL}(s||t) = H(t) - H(s) + (s - t)^T \nabla H(t)$, where two vectors of the same length are denoted by s and t. Next, we begin with the right side of Eq. (5).

$$\sum_{k=1}^{K} \sum_{b_l \in C_k} f(b_l, m_k) = \sum_{k=1}^{K} \sum_{b_l \in C_k} \sum_{i=1}^{r} \sum_{j=1}^{K_i} \left(H(m_{k,ij}) - H(b_{l,ij}) + (b_{l,ij} - m_{k,ij})^T \nabla H(m_{k,ij}) \right) =$$
$$\sum_{k=1}^{K} |C_k| \sum_{i=1}^{r} \sum_{j=1}^{K_i} H(m_{k,ij}) - \sum_{k=1}^{K} \sum_{b_l \in C_k} \sum_{i=1}^{r} \sum_{j=1}^{K_i} H(m_{k,ij}) \quad (6)$$

The above Eq. (6) holds due to $\sum_{b_l \in C_k} (b_{l,ij} - m_{k,ij})$, When considering the binary matrix B, the second term is a constant. Lemma 1 leads us to the conclusion of the proof.

Remark 1. *Theorem 1 reveals how K-means on B and the second component of Eq. (3) are equal. This suggests that the straightforward K-means with KL divergence on every dimension may effectively tackle a portion of this complicated problem.*

4.4. Proposed Algorithm

We proposed an algorithm known as learning-based Outlier Detection (LbOD). Our algorithm's novelty lies in its simultaneous approach to partition space, objective function, and cluster optimization.

Algorithm 1: Learning based Outlier Detection (LbOD)
Input: Data x, partition r, number of clusters K and outliers o
Output: cluster *K* and outliers O

1. Begin
2. r← Create partitions(X)
3. (B,$\tilde{B}$)←Creation Binary Matrices()
4. K← Initialize Centroids (B,$\tilde{B}$)
5. While objective value is unchanged
6. M← Compute distance between data points and nearest centroid
7. outliers← Find points of Highest Distance ()
8. Assign other points to centroids nearest
9. Compute arithmetic mean to update clusters
10. End While
11. End

Algorithm1: Learning based Outlier Detection (LbOD)

As presented in the algorithm, it takes the data number of partitions, clusters, and outliers as input. It produces clusters and the identification of outliers through the clustering process. The algorithm is based on means optimization process. The algorithm also addresses two challenges associated with the optimization problem. Since K-means clustering can resolve the second half of Eq. (3), we should focus on turning the issue towards a solution with K-means. Given that Theorem 1 indicates that $1 - p_{k,ij}$ tough to incorporate into K-means clustering, we intend to represent $1 - p_{k,ij}$ by inserting another binary matrix $\tilde{B} = \{\tilde{b_l}\}, 1 \leq l \leq n$ in the following manner.

$$\tilde{b}_l = (\tilde{b}_{l,1}, \dots, \tilde{b}_{l,i}, \dots, \tilde{b}_{l,r}), \quad \text{with}$$
$$\tilde{b}_{l,i} = (\tilde{b}_{l,i1}, \dots, \tilde{b}_{l,ij}, \dots, \tilde{b}_{l,iKi}), \text{ and} \quad (7)$$
$$\tilde{b}_{l,ij} = \begin{cases} 0, if\ L_{\pi_i}(x_l) = j \\ 1, otherwise \end{cases}$$

Eq. (7) is also used to construct $\tilde{B}$ from Π. $\tilde{B}$ is equivalent to flipping B when compared to the binary matrix B in Eq. (1). The variables like $(K_i - 1)$-of-K_i and 1-of-K_i associated with the initial data

are represented by $\tilde{B}$ and B, respectively. By using Eq. (4) to define m̃_kk in terms of $\tilde{B}$, we are able to derive

$$\tilde{m}_{k,iij} = 1 - m_{k,iij} = 1 - p_{k,iij} .$$

The issue in Eq. (3) is transformed into clustering optimization based on $\tilde{B}$ and B.

Theorem2. When $n - o$ inliers associated with $[B\ \tilde{B}]$ are subjected to K-Means, we have

$$\min_{\pi} \sum_{k=1}^{K} p_k HL(C_k) \Leftrightarrow \min \sum_{k=1}^{K} \sum_{b_l \in C_k} \left(f(b_l, m_k) + f\left(\tilde{b}_l, \tilde{m}_k\right) \right),$$

Where m_k, $\tilde{m}_k$ are the distance function and the k-th centroid determined by Eq. (4).

$f(b_l, m_k) = \sum_{i=1}^{r} \sum_{j=1}^{K_i} D_{KL}\left(b_{l,ij} \| m_{k,ij}\right)$, $f\left(\tilde{b}_l, \tilde{m}_k\right) = \sum_{i=1}^{r} \sum_{j=1}^{K_i} D_{KL}\left(\tilde{b}_{l,ij} \| \tilde{m}_{k,ij}\right)$, and $D_{KL}(\cdot \,||\, \cdot)$ is theKL-divergence.

Remark 2. K-means cannot be used to solve the issue in Eq. (3) regarding the binary matrix B. To represent $1 - p_{k,ij}$, We nontrivially add binary matrix B̃, which is B flipped. This allows clustering on [B̃] to construct the entire problem, as shown in Theorem 2. Benefits include inheriting the K-means algorithm's efficiency and suitability for scalable clustering and detection of outliers and simplifying the issue with a clean mathematical formulation.

The first difficulty, which is the issue with inliers in Eq. (2) by employing the auxiliary matrix B, is fully resolved by Theorem 2. Doing this transforms a simple K-means solution into a whole one. In the subsequent section, we address the second problem, which operates on all data points instead of n-o inliers.

In this work, we investigate the proposed clustering approach, which performs data splits in parallel and identifies outliers. As a result, clustering and outlier identification operations are carried out using the same framework. As centroids associated with K-means have a probability of being outliers, they shouldn't contribute to them. Driven by K-means (Ruff et al., 2021), outliers are defined as places that deviate significantly from the nearest centroid. The problem considered in the solution is linked to partition space' Holoentropy while the K-means focused on feature space in general. Then, utilizing the auxiliary matrix B̃, we formulate the issue as an optimization of K-means. K-means is a method used to solve the Eq. (2) problem, which yields K clusters and outlier set O after careful modification and derivation. Algorithm 1 provides a summary of our suggested clustering procedure with outlier elimination. We then examine the time complexity and convergence of Algorithm 1's property. First, we create R primary partitions in Line 1. These are typically completed by K-means clustering with various cluster numbers. The processing time for this phase is $O(rt'\overline{K}nd)$, where t' and $\overline{K}$ stand for the average number of iterations and clusters, respectively. A comparable temporal complexity, O(tKnR), is indicated by lines 5-8 for the typical K-means-- method, where binary matrices' dimension is expressed as $R = \sum_{i=1}^{r} K_i$ i. Algorithm fines longest distances to find outliers. It is important to remember that parallel computing may be used to create R basic partitions, significantly the execution duration. Furthermore, when compared to the total number of points (n), t', t, r, and R is pretty small. Thus, our technique is easily scalable in clustering to discover outliers, with its time complexity effectively linear in the number of points. Furthermore, Algorithm 1 ensures local optimum with optimized convergence in clustering.

4.5. Dataset Details

The datasets utilized in (Liu et al., 2015; Liu et al., 2018; UCI Datasets) can be obtained. With varying numbers of examples, features, outliers, and clusters, every dataset belongs to a certain kind. Research on outlier detection frequently makes use of these high-dimensional datasets.

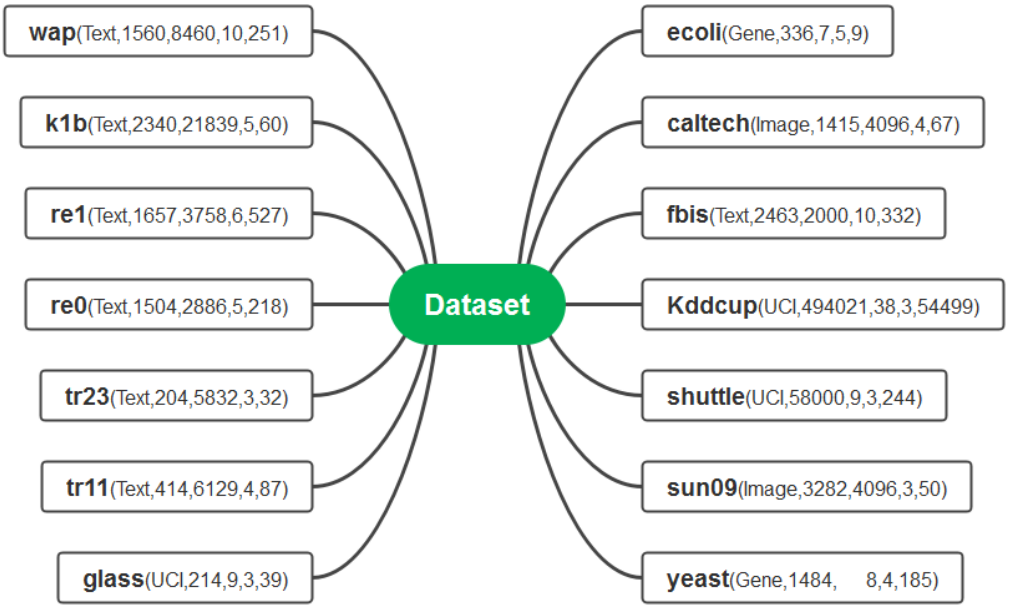

Figure 2: Shows the Data Distribution Dynamics of Different Benchmark Datasets

Figure 2 displays the specifics of each dataset, including its type, number of occurrences, number of features, number of clusters, and number of outliers.

4.6. Evaluation Metrics

The performance of the suggested outlay detection approach is assessed using a variety of metrics, as stated in Eq. (8, 9, 10, 11), respectively, including Normalized Mutual Information (NMI), Rand Index (Rn), Jaccard index, and F-measure.

$$NMI = \frac{\sum_{i,j} n_{ij} log \frac{n.n_{ij}}{n_{i+} \cdot n_{+j}}}{\sqrt{\left(\sum_i n_{i+} log \frac{n_{i+}}{n}\right)\left(\sum_j n_{j+} log \frac{n_{+j}}{n}\right)}} \quad (8)$$

$$R_n = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_i \binom{n_{i+}}{2} \cdot \sum_j \binom{n_{+j}}{2} \Big/ \binom{n}{2}}{\sum_i \binom{n_{i+}}{2} \Big/ 2 + \sum_j \binom{n_{+j}}{2} \Big/ 2 - \sum_i \binom{n_{i+}}{2} \cdot \sum_j \binom{n_{+j}}{2} \Big/ \binom{n}{2}} \quad (9)$$

$$Jaccard = \frac{|O \cap O*|}{|O \cup O*|} \quad (10)$$

$$F - measure = 2 * \frac{precition \cdot recall}{precition + recall} \quad (11)$$

Cluster creation is validated by NMI and Rn as a component of the suggested outlier identification process. The F-measure and Jaccard are used to gauge how accurate the outlier identification is.

5 Experimental Results

The experiments we conducted using the benchmark datasets outlined in section 4.5 are presented in this section. The performance of the suggested outlier identification approach is assessed using a variety of indicators listed in section 4.6. Additionally, it is contrasted with other cutting-edge techniques, and it is discovered that the suggested strategy significantly outperforms them in outlier detection.

Table 1: Performance Comparison in Terms of NMI

Dataset	NMI		
	K-means	K-means--	LbOD
Ecoli	65.05	64.18	64.92
Yeast	20.68	17.33	21.49
Caltech	79.05	77.1	89.73
Sun09	20.18	12.17	22.67
Fbis	12.18	33.7	54.98
k1b	52.95	50.17	55.15
re0	20.2	18.06	34.88
re1	19.66	15.49	38.15
re1t	10.29	21.84	62.63
tr11	7.89	12.68	26.03
Wap	43.36	33.17	50.78
Glass	37.25	37.26	39.82
Shuttle	23.55	26.16	36.15
kddcup	1.46	72.22	86.72

The performance of outlier identification techniques in terms of NMI is shown versus the number of datasets, as shown in Table 1.

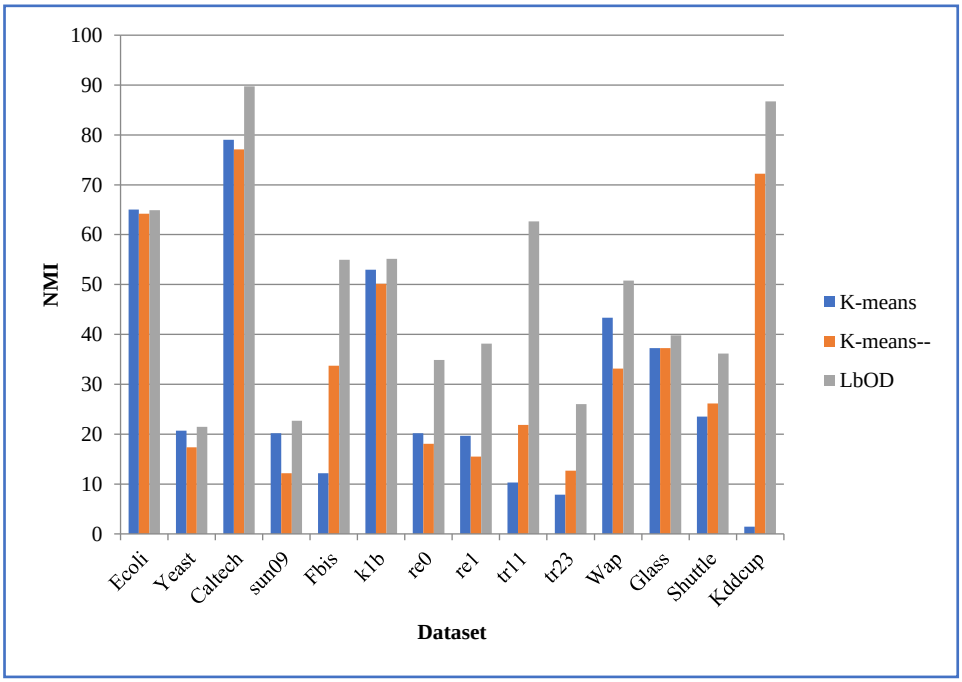

Figure 3: Performance Comparison among Outlier Detection Methods in Terms of NMI

A comparison of the performance of several outlier identification techniques in terms of NMI is shown in Figure 3. The findings demonstrate that, across all datasets utilized in the studies, the suggested outlier identification strategy outperformed other approaches.

Table 2: Performance Comparison in Terms of Rn

Dataset	Rn		
	K-means	K-means--	LbOD
Ecoli	67.83	62.95	70.42
Yeast	15.12	13.78	20.11
caltech	63.13	78.2	89.43
sun09	18.81	10.80	22.2
fbis	-0.67	12.65	40.68
k1b	43.99	44.22	42.01
re0	11.66	13.28	25.59
re1	4.15	5.4	23.3
tr11	0.52	8.63	59.5
tr23	-0.3	4.33	22.5
wap	14.34	12.66	36.64
Glass	23.53	25.56	26.58
Shuttle	40.85	33.44	60.29
kddcup	0.04	81.21	94.76

The performance of outlier identification techniques in terms of Rn is shown versus the number of datasets, as shown in Table 2.

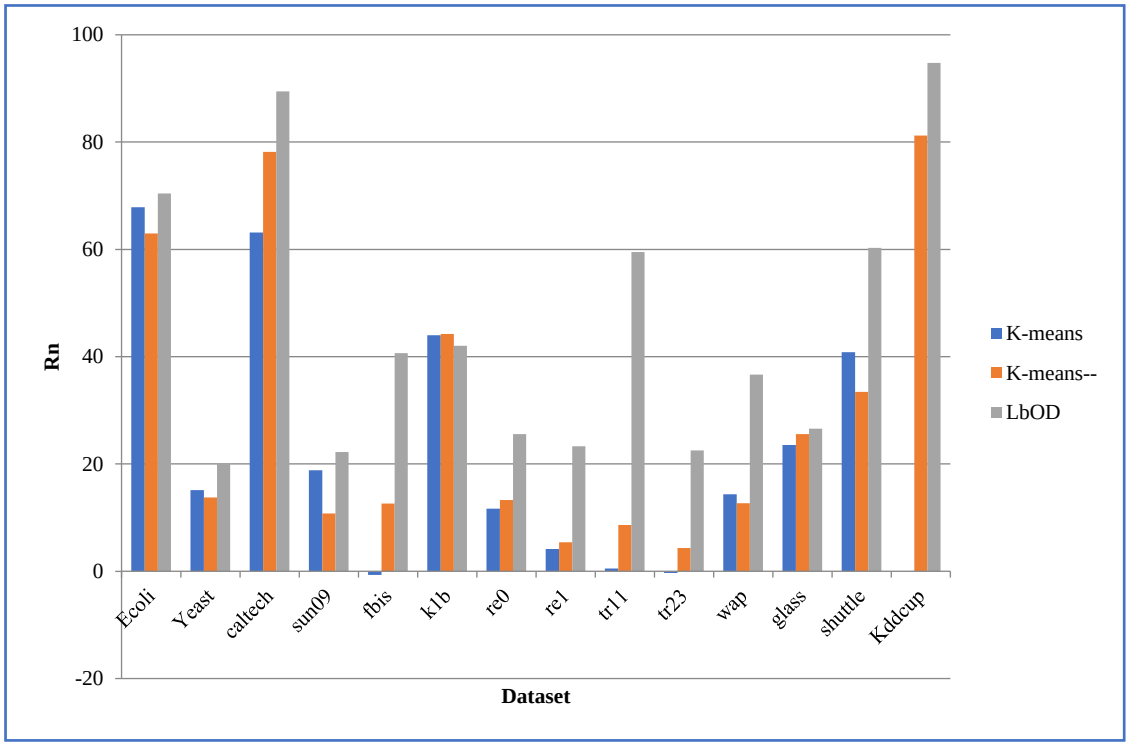

Figure 4: Performance Comparison Among Outlier Detection Methods in Terms of Rn

The performance comparison of the various outlier detection techniques in terms of Rn is given, as depicted in Figure 4. The results indicate that, when compared to current approaches, the suggested outlier identification method performed better across all experimental datasets.

Table 3: Performance Comparison in Terms of Jaccard

Dataset	Jaccard		
	K-means	K-means--	LbOD
ecoli	4.36	58.54	51.12
Yeast	6.25	20.52	51.92
caltech	19.68	45.81	98.58
sun09	1.93	3.71	2.49
fbis	0.09	5.36	26.01
k1b	0	0	21.35
re0	5.56	9.5	29.52
re1	0.56	17.09	29.70
tr11	0	10.35	29.52
tr23	0	6.89	15.01
wap	1.11	11.29	23.31
glass	13.64	32.28	35.54
shuttle	0	5.39	6.51
kddcup	0.01	18.32	16.61

The performance of outlier identification techniques in terms of Jaccard is shown versus the number of datasets, as seen in Table 3.

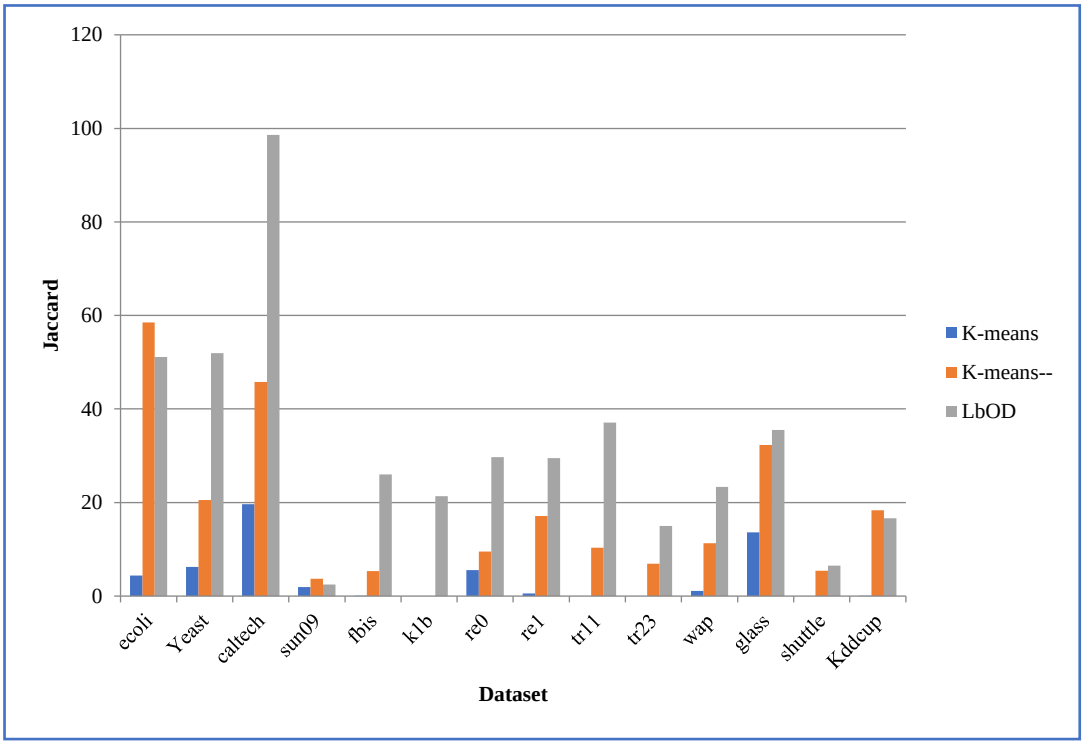

Figure 5: Performance Comparison among Outlier Detection Methods in Terms of Jaccard

The performance comparison of the various outlier detection techniques in terms of Jaccard is given as illustrated in Figure 5. The results indicate that, when compared to current approaches, the suggested outlier identification method performed better across all experimental datasets.

Table 4: Performance Comparison in Terms of F- measure

Dataset	F-measure		
	K-means	K-means--	LbOD
ecoli	8.21	76.18	67.45
Yeast	11.79	33.61	68.36
caltech	31.47	64.21	99.29
sun09	3.78	7.15	4.86
fbis	0.17	10.18	41.3
k1b	0	0	35.16
re0	10.52	17.35	45.78
re1	1.09	29.21	45.58
tr11	0	18.76	54.18
tr23	0	12.92	26.19
wap	2.17	20.28	37.80
glass	23.64	49.56	52.42
shuttle	0	10.22	12.29
kddcup	0.02	31.59	28.51

As presented in Table 4, the performance of outlier detection methods in terms of F-measure is provided against number of datasets.

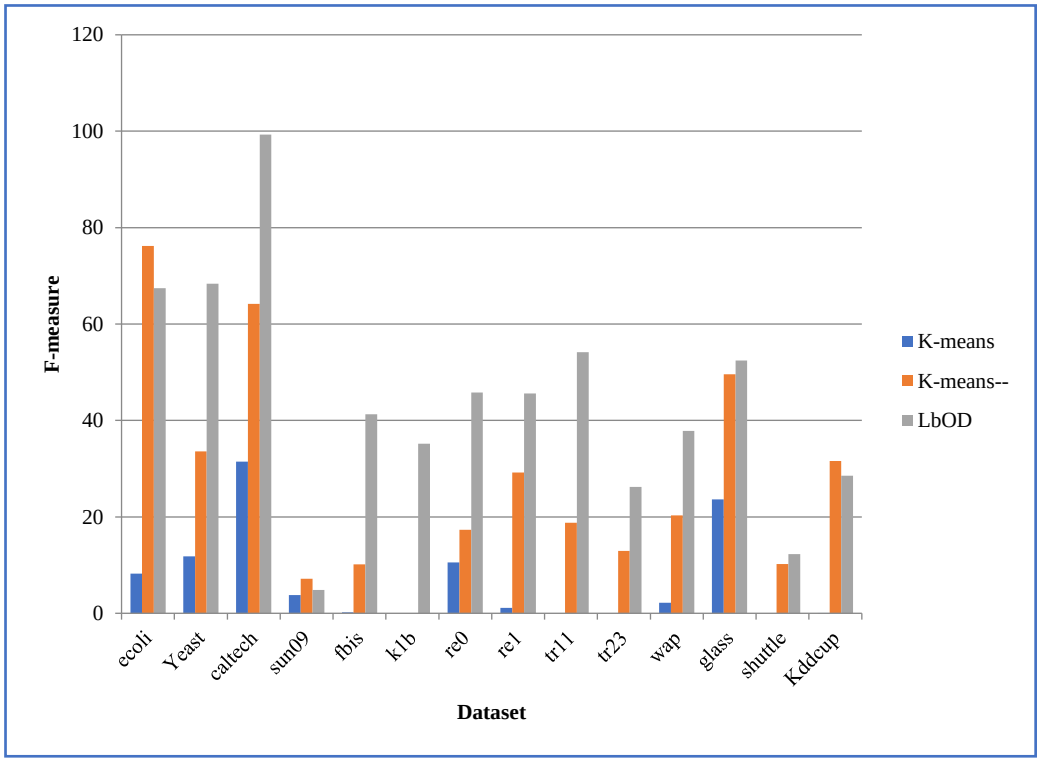

Figure 6: Performance Comparison among Outlier Detection Methods in Terms of F-measure

As president in Figure 6, the performance comparison among outlier detection methods in terms of F-measure is provided. It is observed from the results that the proposed outlier detection method showed better performance over existing methods against all the datasets used in experiments.

Table 5: Performance Comparison Among Many Outlier Detection Methods in Terms of Jaccard Index

Dataset	Jaccard								
	LOF	COF	LDOF	FABOD	iForest	OPCA	TONMF	K-means--	LbOD
ecoli	20.00	38.46	5.88	20.00	38.28	5.88	0.00	45.76	47.37
yeast	11.45	11.45	5.11	13.85	23.75	26.71	8.66	14.38	50.47
caltech	2.29	0.75	1.52	8.06	27.62	0.00	0.00	30.36	97.19
sun09	1.01	2.04	0.00	2.04	2.04	0.00	0.00	3.27	2.27
fbis	8.32	5.56	4.90	6.41	5.40	4.40	8.32	5.21	23.77
k1b	0.00	0.00	0.00	0.84	0.00	0.00	1.69	0.00	20.53
re0	2.59	5.31	3.07	6.34	2.83	11.79	7.13	8.82	28.50
re1	21.85	15.44	15.19	18.83	16.85	17.77	16.98	16.75	27.64
tr11	10.13	8.75	19.18	10.83	8.75	8.75	12.99	9.93	34.06
tr23	4.92	4.92	6.67	10.34	6.67	1.59	10.34	5.87	12.35
wap	10.82	12.30	6.36	12.81	11.31	6.58	7.49	10.98	22.01
glass	16.42	36.84	4.00	25.81	13.04	14.71	0.00	24.00	32.67
shuttle	12.44	12.96	0.21	7.25	1.46	3.61	0.00	5.39	5.58
Kddcup	11.54	15.26	3.40	8.50	21.22	15.66	8.66	15.06	15.98

The performance of additional outlier identification techniques in terms of Jaccard is given versus the number of datasets, as seen in Table 5.

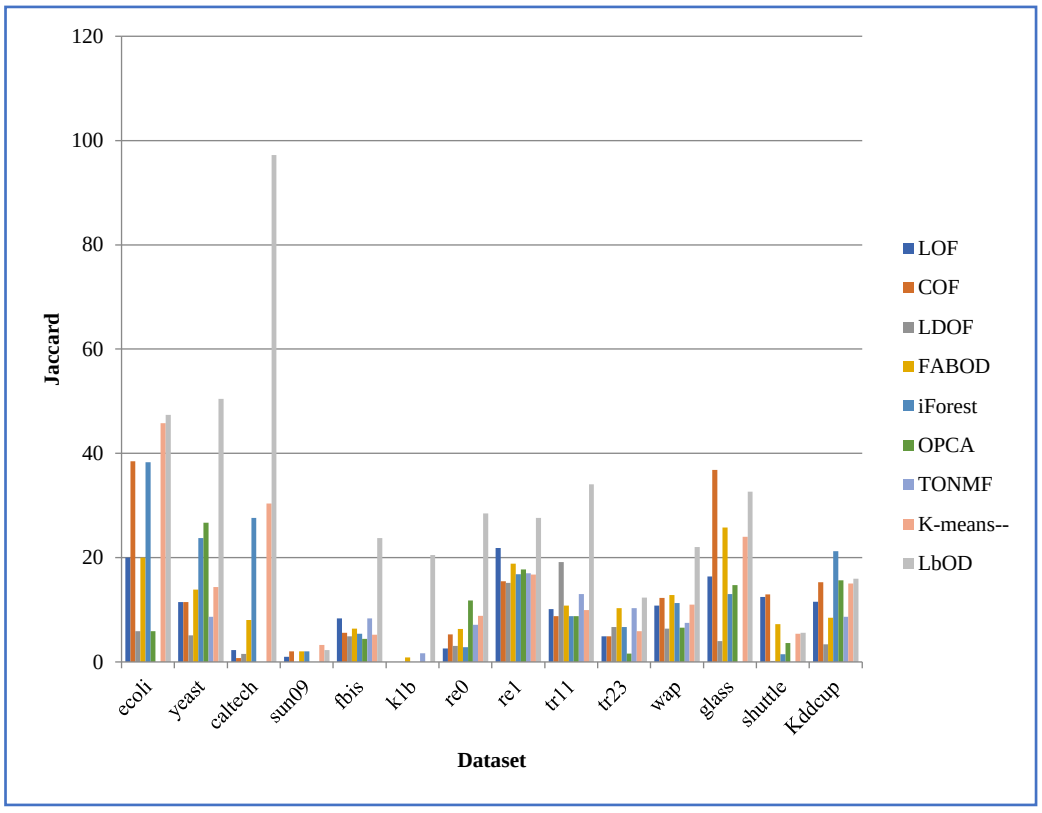

Figure 7: Performance Comparison among more Outlier Detection Methods in Terms of Jaccard

As presented in Figure 7, the performance comparison among more outlier detection methods in terms of Jaccard is provided. It is observed from the results that the proposed outlier detection method showed better performance over existing methods against all the datasets used in experiments.

Table 6: Performance Comparison Among Many Outlier Detection Methods in Terms of F-measure Index

Dataset	F-measure								
	LOF	COF	LDOF	FABOD	iForest	OPCA	TONMF	K-means--	LbOD
ecoli	33.33	55.56	11.11	33.33	55.56	11.11	0.00	61.58	64.21
yeast	20.54	20.54	9.73	24.32	38.38	11.11	8.11	24.69	67.07
caltech	4.48	1.49	2.99	14.93	43.28	0.00	1.49	44.37	98.57
sun09	2.00	4.00	0.00	4.00	4.00	0.00	6.00	6.34	4.44
fbis	15.36	10.54	9.34	12.05	43.28	8.43	15.36	9.91	38.35
k1b	0.00	0.00	0.00	1.67	0.00	0.00	3.33	0.00	34.06
re0	5.05	10.09	5.96	11.93	5.50	21.10	13.30	16.20	44.34
re1	35.86	26.76	26.38	31.69	28.84	30.17	29.03	28.70	43.28
tr11	18.39	16.09	32.18	19.54	16.09	16.09	22.99	18.06	50.74
tr23	9.37	9.37	12.50	18.75	12.50	3.12	18.75	11.08	21.88
wap	19.52	21.91	11.95	22.71	23.75	12.35	13.94	19.78	36.06
glass	28.21	53.85	76.90	22.71	23.08	25.64	0.00	37.97	49.18
shuttle	22.13	22.95	0.41	13.52	2.87	6.97	0.00	10.22	10.56
Kddcup	20.53	18.65	0.31	11.62	35.01	27.08	15.94	26.03	27.55

The performance of additional outlier identification techniques in terms of F-measure is given versus the amount of datasets, as shown in Table 6.

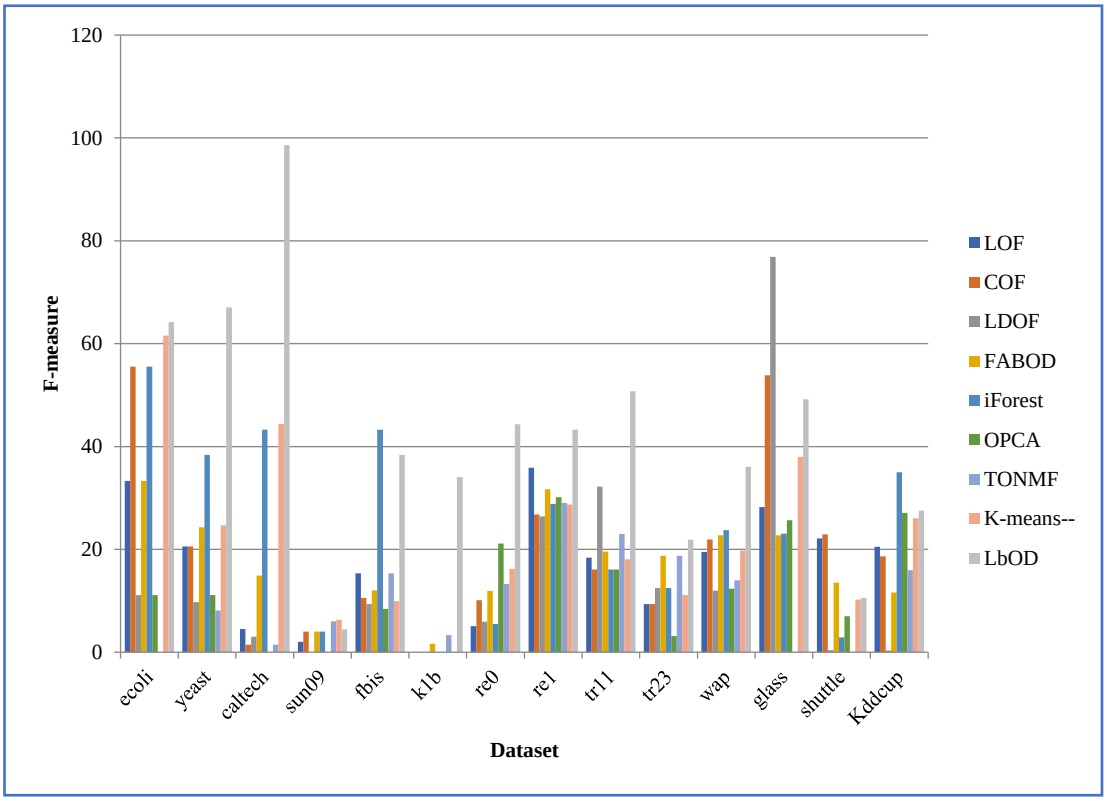

Figure 8: Performance Comparison among more Outlier Detection Methods in Terms of F-measure

A comparison of the performance of additional outlier identification techniques with respect to Jaccard is given in Figure 8. The outcomes demonstrate that, when applied to all of the datasets used in the studies, the suggested outlier identification strategy outperformed current approaches.

6 Discussion

We cover a wide range of clustering with outlier reduction issues in this area. Traditionally, clustering separates many points into discrete groups according to the similarity between the points inside the cluster. A firm or soft label is applied to each spot. Even though robust clustering is intended to lessen the influence of outliers, the cluster label is applied to every point, even outliers. In contrast, the issue we tackle in this work uses clustering to identify the outlier set and only assign labels to inliers. Technically speaking, our approach is a subset of non-exhaustive clustering, wherein individual data points may be labeled or belong to many clusters. Our approach differs in space between K-means and features. Apart from effectively satisfying the notions of outliers and holoentropy, the partition space also enables the spherical structure assumption of the K-means optimization process.

Extensive attempts have been made to flourish the hot study field of outlier detection from several angles. Very few of them do outlier identification and cluster analysis simultaneously. Except for K-means, grouping with outliers is expressed as an integer programming challenge using Langrangian Relaxation (LP) (Zhang et al., 2021), where the input parameter is the cluster building costs. In addition to having highly sophisticated algorithms, LP has difficulty setting this parameter in real-world situations, which causes LP to produce impractical answers. For this reason, we are unable to disclose LP's performance within the part dedicated to experimentation. Our approach begins with the outlier detection objective function and uses a clustering tool to solve the problem. This shows how closely related the domains of cluster analysis and outlier identification are.

Several fundamental divisions are intended to be combined into one cohesive one via consensus clustering. An adaptive KCC utility function for a K-means system is provided for the problematic consensus clustering issue by our earlier work, Consensus Clustering (KCC) (Banerjee et al., 2005; Qais et al., 2024). An analogous collection of fundamental partitions serves as the input for our method, which uses K-means to produce the partition containing outliers. Combining primary partitions to establish consensus clustering and identifying outliers is made possible by the proposed partition space, which is formed from primary partitions. This perspective views holoentropy as the utility function that quantifies the degree of similarity between the final partition and the fundamental partition in B or $\tilde{B}$. The absence of values in fundamental partitions inside the KCC framework does not contribute to the centroid update and is impractical. For the proposed method, we can outliers automatically.

7 Conclusion and Future Work

Our solution for effective outlier detection involved using unsupervised machine learning (ML) of outliers from high-dimensional datasets. An objective function is defined to improve cluster compactness, leading to efficiency in the outlier detection process. Further improvement of clustering process with problem transformation and usage of enhanced K-Means could result in an integrated approach that jointly archives quality clustering and outlier identification. We proposed an algorithm known as Learning based Outlier Detection (LbOD). The novelty of our algorithm lies in the simultaneous approach in partition space, objective function, and cluster optimization. To assess the suggested framework and algorithm's capacity to find outliers while taking into account several benchmark high-dimensional datasets, a prototype is constructed. Our empirical study has revealed that

the LbOD algorithm outperforms many existing outlier detection techniques. We want to improve our framework's future by exploiting the ensemble of multiple best-performing unsupervised learning models with a novel selection strategy.

References

[1] Alaettin, Z., & Volkan, A. (2021). Data stream clustering: a review. *The Artificial Intelligence Review*, *54*(2), 1201-1236. https://doi.org/10.1007/s10462-020-09874-x

[2] Avci, O., Abdeljaber, O., Kiranyaz, S., Hussein, M., Gabbouj, M., & Inman, D. J. (2021). A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications. *Mechanical systems and signal processing*, *147*, 107077. https://doi.org/10.1016/j.ymssp.2020.107077

[3] Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., & Lafferty, J. (2005). Clustering with Bregman divergences. *Journal of machine learning research*, *6*(10), 1705-1749.

[4] Bashar, M. A., & Nayak, R. (2020, December). TAnoGAN: Time series anomaly detection with generative adversarial networks. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1778-1785). IEEE. https://doi.org/10.1109/ssci47803.2020.9308512

[5] Belhadi, A., Djenouri, Y., Srivastava, G., Djenouri, D., Lin, J. C. W., & Fortino, G. (2021). Deep learning for pedestrian collective behavior analysis in smart cities: A model of group trajectory outlier detection. *Information Fusion*, *65*, 13-20. https://doi.org/10.1016/j.inffus.2020.08.003

[6] Brito, L. C., Susto, G. A., Brito, J. N., & Duarte, M. A. (2022). An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery. *Mechanical Systems and Signal Processing, 163,* 108105. https://doi.org/10.1016/j.ymssp.2021.108105

[7] Carreño, A., Inza, I., & Lozano, J. A. (2020). Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. *Artificial Intelligence Review*, *53*, 3575-3594. https://doi.org/10.1007/s10462-019-09771-y

[8] Chakraborty, D., Narayanan, V., & Ghosh, A. (2019). Integration of deep feature extraction and ensemble learning for outlier detection. *Pattern Recognition*, *89*, 161-171. https://doi.org/10.1016/j.patcog.2019.01.002

[9] Chawla, S., & Gionis, A. (2013, May). k-means–: A unified approach to clustering and outlier detection. In *Proceedings of the 2013 SIAM international conference on data mining* (pp. 189-197). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972832.21

[10] Chen, T., Liu, X., Xia, B., Wang, W., & Lai, Y. (2020). Unsupervised anomaly detection of industrial robots using sliding-window convolutional variational autoencoder. *IEEE Access*, *8*, 47072-47081. https://doi.org/10.1109/access.2020.2977892.

[11] Chen, Y., Xiao, X., Peng, C., Lu, G., & Zhou, Y. (2021). Low-rank tensor graph learning for multi-view subspace clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(1), 92-104. https://doi.org/10.1109/TCSVT.2021.3055625

[12] Crimi, A., Bakas, S., Kuijf, H., Menze, B., & Reyes, M. (Eds.). (2018). *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers* (Vol. 10670). Springer. https://doi.org/10.1007/978-3-030-11723-8_16

[13] Deepak, K., Chandrakala, S., & Mohan, C. K. (2021). Residual spatiotemporal autoencoder for unsupervised video anomaly detection. *Signal, Image and Video Processing*, *15*(1), 215-222. https://doi.org/10.1007/s11760-020-01740-1

[14] Dhiman, H. S., Deb, D., Muyeen, S. M., & Kamwa, I. (2021). Wind turbine gearbox anomaly detection based on adaptive threshold and twin support vector machines. *IEEE Transactions on Energy Conversion*, *36*(4), 3462-3469. https://doi.org/10.1109/TEC.2021.3075897

[15] Djenouri, Y., Belhadi, A., Lin, J. C. W., Djenouri, D., & Cano, A. (2019). A survey on urban traffic anomalies detection algorithms. *IEEE Access, 7,* 12192-12205. https://doi.org/10.1109/ACCESS.2019.2893124

[16] Enes, J., Expósito, R. R., Fuentes, J., Cacheiro, J. L., & Touriño, J. (2023). A pipeline architecture for feature-based unsupervised clustering using multivariate time series from HPC jobs. *Information Fusion*, *93*, 1-20. https://doi.org/10.1016/j.inffus.2022.12.017

[17] Erhan, L., Ndubuaku, M., Di Mauro, M., Song, W., Chen, M., Fortino, G., ... & Liotta, A. (2021). Smart anomaly detection in sensor systems: A multi-perspective review. *Information Fusion*, *67*, 64-79. https://doi.org/10.1016/j.inffus.2020.10.001

[18] Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2019). Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. *IEEE Access, 7,* 144777-144789. https://doi.org/10.1109/ACCESS.2019.2945129

[19] Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). HDPM: an effective heart disease prediction model for a clinical decision support system. *IEEE Access*, *8*, 133034-133050. https://doi.org/10.1109/ACCESS.2020.3010511

[20] Giji Kiruba, D., Benita, J., & Rajesh, D. (2023). A Proficient Obtrusion Recognition Clustered Mechanism for Malicious Sensor Nodes in a Mobile Wireless Sensor Network. *Indian Journal of Information Sources and Services, 13*(2), 53–63. https://doi.org/10.51983/ijiss-2023.13.2.3793

[21] Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: a review of anomaly detection techniques and recent advances. *Expert systems with applications*, *193*, 116429. https://doi.org/10.1016/j.eswa.2021.116429

[22] Khedr, A. M., Raj, P. P., & Al Ali, A. (2020). An Energy-Efficient Data Acquisition Technique for Hierarchical Cluster-Based Wireless Sensor Networks. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *11*(3), 70-86. https://doi.org/10.22667/JOWUA.2020.09.30.070

[23] Kraus, M., Weiler, N., Oelke, D., Kehrer, J., Keim, D. A., & Fuchs, J. (2019). The impact of immersion on cluster identification tasks. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 525-535. https://doi.org/10.1109/TVCG.2019.2934395

[24] Landauer, M., Skopik, F., Wurzenberger, M., & Rauber, A. (2020). System log clustering approaches for cyber security applications: A survey. *Computers & Security*, *92*, 101739. https://doi.org/10.1016/j.cose.2020.101739

[25] Liu, H., Li, J., Wu, Y., & Fu, Y. (2019). Clustering with outlier removal. *IEEE transactions on knowledge and data engineering, 33*(6), 2369-2379. https://doi.org/10.1109/TKDE.2019.2954317

[26] Liu, H., Liu, T., Wu, J., Tao, D., & Fu, Y. (2015, August). Spectral ensemble clustering. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 715-724). https://doi.org/10.1145/2783258.2783287

[27] Liu, H., Shao, M., Li, S., & Fu, Y. (2016, August). Infinite ensemble for image clustering. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1745-1754). https://doi.org/10.1145/2939672.2939813

[28] Liu, H., Shao, M., Li, S., & Fu, Y. (2018). Infinite ensemble clustering. *Data Mining and Knowledge Discovery*, *32*, 385-416. https://doi.org/10.1007/s10618-017-0539-5

[29] Liu, H., Wu, J., Tao, D., Zhang, Y., & Fu, Y. (2015, June). Dias: A disassemble-assemble framework for highly sparse text clustering. In *Proceedings of the 2015 SIAM International Conference on Data Mining* (pp. 766-774). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611974010.86

[30] 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*, *32*(8), 1517-1528. https://doi.org/10.1109/TKDE.2019.2905606

[31] Mallikarjuna, M., & Rao, R. P. (2019). Classification of Capital Markets by Using Cluster Analysis. *International Academic Journal of Accounting and Financial Management, 6*(1), 11–22. https://doi.org/10.9756/IAJAFM/V6I1/1910002

[32] Meng, F., Yuan, G., Lv, S., Wang, Z., & Xia, S. (2019). An overview on trajectory outlier detection. *Artificial Intelligence Review*, *52*, 2437-2456. https://doi.org/10.1007/s10462-018-9619-1

[33] Mensi, A., Tax, D. M., & Bicego, M. (2023). Detecting outliers from pairwise proximities: Proximity isolation forests. *Pattern Recognition*, *138*, 109334. https://doi.org/10.1016/j.patcog.2023.109334.

[34] Mishra, N., & Pandya, S. (2021). Internet of things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review. *IEEE Access*, *9*, 59353-59377. https://doi.org/10.1109/ACCESS.2021.3073408

[35] Munoz-Organero, M. (2019). Outlier detection in wearable sensor data for human activity recognition (HAR) based on DRNNs. *IEEE Access*, *7*, 74422-74436. https://doi.org/10.1109/ACCESS.2019.2921096

[36] Qais, M. H., Kewat, S., Loo, K. H., & Lai, C. M. (2024). Early outlier detection in three-phase induction heating systems using clustering algorithms. *Ain Shams Engineering Journal*, *15*(3), 102467. https://doi.org/10.1016/j.asej.2023.102467

[37] Raghavan, P., & El Gayar, N. (2019, December). Fraud detection using machine learning and deep learning. In *2019 international conference on computational intelligence and knowledge economy (ICCIKE)* (pp. 334-339). IEEE. https://doi.org/10.1109/ICCIKE47802.2019.9004231

[38] Rogers, T. J., Worden, K., Fuentes, R., Dervilis, N., Tygesen, U. T., & Cross, E. J. (2019). A Bayesian non-parametric clustering approach for semi-supervised structural health monitoring. *Mechanical Systems and Signal Processing, 119,* 100-119. https://doi.org/10.1016/j.ymssp.2018.09.013

[39] Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., ... & Müller, K. R. (2021). A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, *109*(5), 756-795. https://doi.org/10.1109/JPROC.2021.3052449

[40] Sadgali, I., Sael, N., & Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. *Procedia computer science, 148,* 45-54. https://doi.org/10.1016/j.procs.2019.01.007

[41] Santhosh, G., & Prasad, K. V. (2023). Energy Saving Scheme for Compressed Data Sensing Towards Improving Network Lifetime for Cluster based WSN. *Journal of Internet Services and Information Security, 13*(1), 64-77. https://doi.org/10.58346/JISIS.2023.I1.007

[42] Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., ... & Nenadic, G. (2019). Machine learning methods for wind turbine condition monitoring: A review. *Renewable energy*, *133*, 620-635. https://doi.org/10.1016/j.renene.2018.10.047

[43] Strehl, A., & Ghosh, J. (2002). Cluster ensembles---a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, *3*(Dec), 583-617.

[44] Tang, T., Chen, S., Zhao, M., Huang, W., & Luo, J. (2019). Very large-scale data classification based on K-means clustering and multi-kernel SVM. *Soft Computing*, *23*, 3793-3801. https://doi.org/10.1007/s00500-018-3041-0

[45] Thangaramya, K., Kulothungan, K., Indira Gandhi, S., Selvi, M., Santhosh Kumar, S. V. N., & Arputharaj, K. (2020). Intelligent fuzzy rule-based approach with outlier detection for secured routing in WSN. *Soft Computing*, *24*, 16483-16497. https://doi.org/10.1007/s00500-020-04955-z

[46] Thöle, L. M., Amsler, H. E., Moretti, S., Auderset, A., Gilgannon, J., Lippold, J., ... & Jaccard, S. L. (2019). Glacial-interglacial dust and export production records from the Southern Indian

Ocean. *Earth and planetary science letters*, *525*, 115716. https://doi.org/10.1016/j.epsl.2019.115716

[47] Thomas, M., Balamurugan, P. Real-Time Violence Detection and Alert System using MobileNetV2 and Cloud Firestore Proceedings of the 2nd IEEE International Conference on Networking and Communications 2024, ICNWC 2024, 2024

[48] UCI Datasets. https://archive.ics.uci.edu/datasets

[49] Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. L. A., Elkhatib, Y., ... & Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE access*, *7*, 65579-65615. https://doi.org/10.1109/ACCESS.2019.2916648

[50] Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. *IEEE Access*, *7*, 107964-108000. https://doi.org/10.1109/ACCESS.2019.2932769

[51] Wu, S., & Wang, S. (2011). Information-theoretic outlier detection for large-scale categorical data. *IEEE transactions on knowledge and data engineering*, *25*(3), 589-602. https://doi.org/10.1109/TKDE.2011.261

[52] Yang, J., Chen, Y., & Rahardja, S. (2023). Neighborhood representative for improving outlier detectors. *Information Sciences*, *625*, 192-205. https://doi.org/10.1016/j.ins.2022.12.041.

[53] Yang, J., Rahardja, S., & Fränti, P. (2021). Mean-shift outlier detection and filtering. *Pattern Recognition*, *115*, 107874. https://doi.org/10.1016/j.patcog.2021.107874

[54] Zhang, W., Yang, D., & Zhang, S. (2021). A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring. *Expert Systems with Applications*, *174*, 114744. https://doi.org/10.1016/j.eswa.2021.114744

[55] Zigui, L., Caluyo, F., Hernandez, R., Sarmiento, J., & Rosales, C. A. (2024). Improving Communication Networks to Transfer Data in Real Time for Environmental Monitoring and Data Collection. *Natural and Engineering Sciences, 9*(2), 198-212. https://doi.org/10.28978/nesciences.1569561

Authors Biography

Girish Reddy Ginni, is a dedicated researcher and PhD student in the Department of Computer Science at GITAM University, Visakhapatnam, Andhra Pradesh, India. He received his master's degree in Computer Science Engineering at Pydah College of Engineering, Affiliated to Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India. His research interests are Artificial Intelligence and Machine Learning applications and optimization methods.

Dr. Srinivasa L. Chakravarthy, is an Educator & researcher of Department of Computer Science at GITAM University, Visakhapatnam, Andhra Pradesh, India. He received his Phd. from Andhra University India. He has 25 years of academic experience. His research interests are Artificial Intelligence and Machine Learning, Quantum Computing, Game theory. He is working as reviewer for noteworthy journals that are SCOPUS, SCI indexed journal. He has published 25 articles in reputed journals.