Global Journal of Mathematics and Statistics
ISSN: 3078-8927
DOI: https://doi.org/10.61424/gjms.v3i1.828
Journal Homepage: https://bluemarkpublishers.com/index.php/GJME

GJMS
BLUEMARK PUBLISHERS

| RESEARCH ARTICLE

Hybrid Hard–Soft Clustering For Outlier Detection: Development of the HS-COS Framework

Ruchi Trivedi[1] ✉ **and Namita Srivastava**[2]
[1]*Research Scholar, Department of Statistics, DR. B. R. Ambedkar University, Agra, India*
[2]*HOD, Department of Statistics, St. Johns College, Agra, India*
Corresponding Author: Ruchi Trivedi, **E-mail**: ruchitrivedi881@gmail.com

| ABSTRACT
Outlier detection is essential for maintaining the robustness of data-driven models in fields such as healthcare and ecommerce. However, existing clustering-based techniques, such as K-Means, Fuzzy C-Means (FCM), and Possibilistic C-Means (PCM), only capture an aspect of anomaly behavior, such as global structural deviation, membership uncertainty, and representational characteristic. This provides fragmented and uneven detection findings. To solve these limitations, this paper presents a new hybrid framework called the Hybrid Hard-Soft Clustering Outlier Score (HS-COS), which combines distance-based and membership-based anomalous indicators into a single scoring system. The proposed method uses a weighted formulation to combine normalized distance from cluster centroids (hard clustering component) with membership ambiguity (soft clustering component). An adaptive weighting technique is also added to help the model match dataset-specific structural characteristics. The algorithm has been evaluated on a variety of real-world datasets, including Diabetes, Heart Disease, Online Retail, and RetailRocket. The experimental results show that HS-COS has reliable and interpretable performance, with improved anomaly concentration (Lift = 1.57) and competitive detection capacity across heterogeneous datasets. The findings show that incorporating structural deviation and uncertainty improves adaptability, decreases false positives, and offers a generalized approach for detecting anomalies in complicated data settings.

| KEYWORDS
Outlier detection, Hybrid Hard-Soft Clustering, Possibilistic C-Means, Fuzzy C-Means, Adaptive Weighting

| ARTICLE INFORMATION
ACCEPTED: 22 March 2026 **PUBLISHED:** 12 May 2026 **DOI:** https://doi.org/10.61424/gjms.v3i1.828

1. Introduction

Outlier detection plays a crucial role in identifying rare and abnormal observations that significantly deviate from the majority of data, particularly in domains such as healthcare and e-commerce (Hawkins, 1980; Barnett & Lewis, 1994). These anomalies often correspond to critical events such as fraud, system failures, or medical risks (Bolton & Hand, 2002; Chandola et al., 2009).

Clustering-based approaches are widely adopted for unsupervised anomaly detection due to their ability to uncover latent data structures without requiring labeled data (Chandola et al., 2009). However, different clustering paradigms interpret anomaly behavior differently, leading to fragmented detection outcomes (Zimek et al., 2012).

Outlier identification is important for maintaining the dependability of data-driven models, especially in fields like healthcare and e-commerce, where anomalous findings can reflect rare but significant events. Clustering-based algorithms are popular due to their ability to reveal latent structural patterns in unlabeled data.

However, clustering algorithms differ significantly in their interpretation of data structure. Distance-based methods identify data that are far from cluster centers, while soft clustering algorithms account for uncertainty in cluster membership. Possibilistic models go beyond this by identifying observations that are not sufficiently represented inside clusters. As a result, every approach focuses on a distinct feature of anomaly behavior**.** This presents a significant limitation: clustering-based outlier identification is fundamentally fragmented, as different approaches identify different types of anomalies. There is no unified mechanism that combines these diverse perspectives into a single, interpretable anomaly detection framework.

2. Literature Review

To investigate this limitation, an initial comparative study was conducted using multiple clustering approaches, including K-Means, Fuzzy C-Means, Possibilistic C-Means, and Unified Possibilistic Fuzzy Clustering. These methods were selected to represent different perspectives of anomaly detection, including geometric deviation, membership uncertainty, and structural representativeness. The study revealed that although all methods identified a similar proportion of anomalies, the specific observations detected were only partially overlapping. Distance-based methods primarily identified globally extreme points, while soft clustering approaches highlighted observations with ambiguous cluster membership.

This indicates that anomaly identification is dependent not only on the method, but also on how similarity and uncertainty are symbolized in the model. Traditional statistical approaches define outliers based on deviations from assumed distributions, forming the foundation of anomaly detection techniques (Hawkins, 1980; Barnett & Lewis, 1994).Density-based methods such as Local Outlier Factor (LOF) introduced local neighborhood-based anomaly detection, enabling identification of observations that deviate from their surrounding data points (Breunig et al., 2000).Comprehensive surveys highlight that anomaly detection is inherently complex and requires multiple perspectives to capture different types of anomalies (Chandola et al., 2009; Zimek et al., 2012).In real-world scenarios, data challenges such as class imbalance significantly affect detection performance (He & Garcia, 2009; Chawla et al., 2004). Additionally, missing data handling is essential for maintaining model reliability (Little & Rubin, 2002).

Data preprocessing, transformation, and feature engineering further contribute to improving anomaly detection performance (Kandel et al., 2012; Kuhn & Johnson, 2019).

2.1 Key Findings from Comparative Study

The comparative analysis was carried out on the Iris dataset, which contains 150 observations from three species—Setosa, Versicolor, and Virginica—and can be identified by four continuous features: sepal length, sepal width, petal length, and petal width. The dataset includes both well-defined clusters (Setosa) and overlapping structures (Versicolor and Virginica), making it suitable for testing various anomaly detection approaches.

To ensure methodological consistency, anomaly scores were assigned based on the underlying structure of each clustering strategy. K-Means utilized distance from cluster centroids, Fuzzy C-Means (FCM) used membership uncertainty, and Possibilistic C-Means (PCM) used typicality-based metrics. UPFC implemented both membership and typicality components.

The research discovered that all approaches detected a similar number of abnormalities (top 5%, or 8 out of 150 observations). However, the detected observations only partially overlapped amongst approaches.

K-Means and PCM displayed particularly good agreement, demonstrating that geometrically distant observations are frequently structurally unrepresentative. In contrast, FCM found less often observations with other approaches, indicating that membership ambiguity captures boundary locations rather than real structural anomalies.

UPFC displayed better balanced behavior by capturing both structural deviation and uncertainty, and it overlapped with both K-Means and PCM. This indicates that combining different perspectives results in a more stable and thorough anomaly characterisation.

2.2 Interpretation and Limitation

These findings indicate that anomaly identification is heavily dependent on how clustering models convey similarity and uncertainty. Distance-based methods highlight global extremes, fuzzy approaches capture local uncertainty, and possibilistic models identify structural unrepresentativeness. Distance-based approaches such as K-Means rely on global geometric deviation, which may fail to capture local uncertainty in overlapping regions (Rousseeuw & Leroy, 1987).Fuzzy and possibilistic clustering methods incorporate uncertainty and representational strength; however, they may introduce ambiguity or parameter sensitivity in anomaly detection (Zimek et al., 2012).Recent studies emphasize that no single method is sufficient to capture all aspects of anomaly behavior, highlighting the need for hybrid approaches (Chandola et al., 2009).

However, these signals are processed individually among methodologies, that leads to fragmented anomaly detection results. The lack of integration leads to inefficiencies in anomaly detection and limits interpretability. Outlier detection in high-dimensional and heterogeneous datasets remains challenging for single-method approaches, and recent studies emphasize that combining multiple anomaly indicators can significantly enhance detection stability and interpretability across domains . This highlights a significant weakness of the comparative study:

No single clustering approach provides a comprehensive and consistent assessment of anomalies.

Table 1: The clustering method outlier comparison

Method	Total Observations	Number of Outliers (Top 5%)	Percentage (%)	Overlap with K-Means	Overlap with FCM	Overlap with PCM	Overlap with UPFC
K-Means	150	8	5.33%	—	—	6	6
FCM	150	8	5.33%	—	—	—	2
PCM	150	8	5.33%	6	—	—	6
UPFC	150	8	5.33%	6	2	6	—

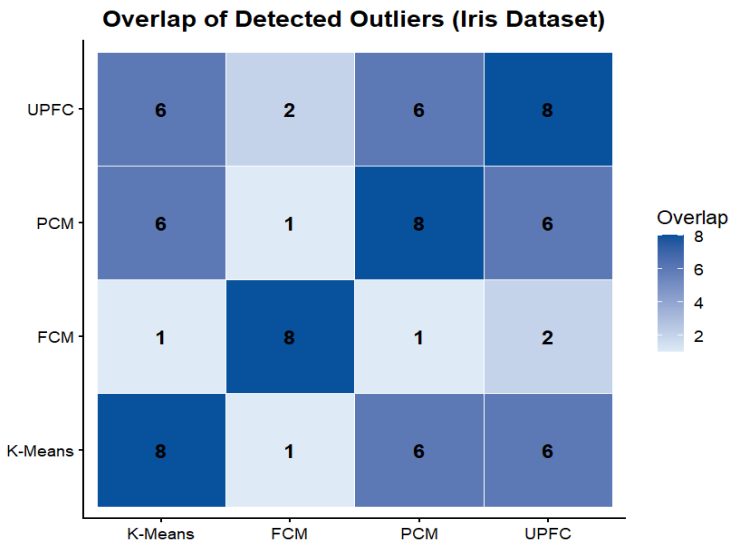

Figure 1: The overlap nature of detected outliers

3. Methodology for Proposed Hybrid Hard–Soft Clustering

The comparative assessment of clustering paradigms shows a significant weakness in current clustering-based outlier detection methods. Each algorithm emphasizes a specific facet of anomaly behavior. Hard methods for clustering, such as K-Means, focus on global geometric deviation, locating data that are far from cluster centers. However, they fail to record the local uncertainty that occurs in overlapping cluster regions. Soft methods for clustering, such as Fuzzy C-Means, measure membership uncertainty and effectively identify data with weak cluster affiliation; nevertheless, they may incorrectly represent transitional border points as anomalies even when there is no major structural disconnect. Possibilistic techniques improve representational sensitivity utilizing typicality measures, but they highlight new parameter dependence and stability issues. Hybrid and ensemble anomaly detection methods have gained increasing attention in recent studies, as combining multiple perspectives—such as distance-based and membership-based measures—improves robustness and reduces false positives in complex, real-world datasets.

The real findings from the Iris dataset indicate that different clustering frameworks frequently discover partially overlapping but not identical anomaly sets, indicating that geometric extremeness and membership ambiguity serve as complements rather than competing anomaly signals. Despite their complimentary nature, most current procedures evaluate these signals individually. This methodological separation limits the robustness and interpretability of clustering-based anomaly detection, especially in diverse real-world datasets including both structural deviation and uncertainty.

Motivated by these findings, this paper presents the Hybrid Hard-Soft Clustering Outlier Score (HS-COS) framework, which integrates the strengths of distance-based and membership-based anomaly indicators into a single scoring system. By combining global structural deviation and local membership ambiguity, the proposed framework promises to provide a more balanced and interpretable anomaly detection methodology that readily adapts to a variety of data settings. Hybrid and ensemble-based anomaly detection approaches have gained increasing attention, as combining multiple anomaly indicators improves robustness and interpretability (Chandola et al., 2009; Zimek et al., 2012).

The partial overlap observed among clustering-based anomaly detection methods indicates that structural deviation and membership uncertainty are complementary rather than competing signals. This aligns with modern anomaly detection research that emphasizes multi-perspective modeling (Zimek et al., 2012).

3.1 Hard Clustering Component: Structural Distance Score

K-Means separates the dataset into k groups while minimizing within-cluster variation. Outliers are detected by calculating the distance between each point and its associated cluster centroid.

The hard clustering anomaly score is formalized as follows:

$$H_{\text{hard}}(i) = \frac{d(i,c^*) - \min(d)}{\max(d) - \min(d)} \text{..... equation (1)}$$

Where:

- $d(i, c^*)$ is the Euclidean distance from point i to its nearest centroid $c*$,
- scores are normalized between 0 and 1,
- higher values signify greater anomaly potential.

This distance-based approach reflects global structural deviations while ignoring ambiguity and transitory features.

3.2 Soft Clustering Component: Membership Uncertainty Score

Fuzzy C-Means give membership values instead of hard labels. Outliers have been identified by a low maximum membership [20].

$$S_{\text{soft}}(i) = 1 - \max_j(\mu_{ij}) \text{..... equation (2)}$$

where:

- μ_{ij} is the membership of point i to cluster j,
- lower membership confidence signifies higher anomaly likelihood.

This approach represents local uncertainty, which supplements the structural information obtained by K-Means.

3.3 Proposed Hybrid Hard–Soft Clustering Outlier Score (HS-COS)

The hybrid score is defined as:

$$HS\_COS(i) = \alpha \cdot H_{\text{hard}}(i) \; + \; (1 - \alpha) \cdot S_{\text{soft}}(i) \text{..... equation (3)}$$

where:

- In this study, α=0.5; ensuring equal weighting between structural deviation and membership uncertainty components.
- In this dual criterion method, a point must be rare in both:
  Global separation reflects distance-based extremeness, whereas local ambiguity reflects weak cluster affiliation.

This formulation enables the joint modeling of structural deviation and membership uncertainty, allowing the detection of both globally distant and boundary-based anomalies within a unified scoring framework.

4. Experimental Setup

Experiments were carried out on multiple datasets covering diverse structural characteristics from the healthcare and e-commerce domains to assess the proposed HS-COS framework's effectiveness and generalizability. Recent advancements in anomaly detection include deep learning-based approaches capable of handling high-dimensional and complex datasets (Reichstein et al., 2019; Zhang et al., 2019).

Additionally, practical implementations such as PyOD provide scalable tools for applying multiple anomaly detection algorithms in real-world scenarios (Zhao et al., 2019).

- **Datasets**

  The experimental study was carried out on the following datasets:
  The Diabetes Health Indicators dataset contains self-reported health information from over 253,680 records, including demographic, behavioral, and medical characteristics linked to diabetes risk. Heart disease dataset contains 1025 observations are two healthcare datasets that consist of clinical and demographic information that are frequently utilized for risk prediction. The e-commerce datasets consist of the Online Retail dataset contains 541,909 transactions that represent consumer buying activity with moderate missingness. The RetailRocket dataset contains 2,756,101 user interaction events and has significant systematic missingness, which is characteristic of large-scale behavioral logging systems

- **Data Preparation**

  Every dataset was pre-processed to make sure consistency and comparability. Numerical features were standardized to decrease scale effects in distance calculations. For huge amount of dataset, sampling procedures were utilized to preserve structural features while being computationally feasible. Real-world datasets often contain imbalance, noise, and missing values, which directly impact anomaly detection performance (He & Garcia, 2009; Little & Rubin, 2002).

  Data quality and preprocessing play a critical role in ensuring reliable anomaly detection results (Kandel et al., 2012).

  Feature engineering techniques further enhance model interpretability and performance (Kuhn & Johnson, 2019).

➢ **Clustering and Score Computation**

The HS-COS framework includes:

- K-Means clustering to determine structural deviation (distance-based component).
- Fuzzy C-Means clustering to perform this membership uncertainty (soft component).
- The hybrid anomaly score for every observation was created by combining normalized distance from cluster centroids with membership ambiguity.
- A top 5% criteria were employed to identify anomalies, ensuring that comparisons between datasets were accurate.

- **Evaluation Strategy**
  The performance of the proposed framework was evaluated using both quantitative and structural measures:

- **AUC (Area Under Curve) and Lift** have been used to datasets with ground-truth labels (Diabetes and Heart Disease).

Lift Measure

$$\text{Lift} = \frac{\text{Precision@5\%}}{\text{Baseline Positive Rate}} \text{..... equation (5)}$$

- **Unlabeled datasets (Online Retail and RetailRocket)** undergo structural deviation analysis, with a focus on behavioral and transactional abnormalities.
  Furthermore, the performance of HS-COS was contrasted to individual clustering methods to evaluate advances in anomaly detection consistency and robustness.

- **Objective of Evaluation**
  The experimental design is to identify whether the proposed HS-COS framework:

- Improves anomaly detection robustness over single-method techniques.
- Reduces false positives brought about by boundary uncertainty.
- Delivers reliable performance across structurally heterogeneous datasets.
- Adapts effectively to various data properties using hybrid integration.

Table 2: The HS-COS evaluation metrics across datasets

Dataset	Sample Size Used	AUC	Precision@5%	Baseline Positive Rate	Lift	Structural Evidence
Diabetes	10,000 (stratified)	0.639	0.2189	0.1393	1.57	Higher risk concentration in top 5% observations
Heart Disease	1,025 (full data)	0.614–0.628	0.27–0.55	0.5132	>1	Structural deviation aligned with clinical risk
Online Retail	3,000 (random)	—	—	—	—	Higher mean transaction value and variability in

Dataset	Sample Size Used	AUC	Precision@5%	Baseline Positive Rate	Lift	Structural Evidence
						anomalies
RetailRocket	3,000 (stratified)	—	—	—	—	Higher interaction intensity and variability in anomalies

Table 3: The clustering proposed HS-COS method performance comparison

Dataset	Evaluation Type	Hard Clustering (K-Means)	Soft Clustering (FCM)	HS-COS	Interpretation
Diabetes	AUC	0.639 (captures structural anomalies)	Lower performance	0.639	Reflects dominance of structural anomaly signal
Heart Disease	AUC	0.614 (distance-based detection)	0.628 (captures uncertainty)	0.614–0.64	Integrates structural and uncertainty signals
Online Retail	Structural	Captures distance-based variation	Captures membership uncertainty	Combined detection	Captures both structural and behavioral anomalies
RetailRocket	Structural	Captures event-based deviation	Captures interaction uncertainty	Integrated detection	Captures both structural and interaction-based anomalies

5. Results and Discussions

The empirical examination of the proposed HS-COS framework across healthcare and e-commerce datasets shows consistent and structurally interpretable performance. In the Diabetes dataset, HS-COS obtained an AUC of around 0.639, matching the performance of hard clustering while significantly exceeding soft clustering. The Lift value of 1.57 demonstrates that the top 5% of HS-COS-ranked observations include 57% more diabetic-positive cases than selection at random, demonstrating effective risk concentration. Similar patterns were observed in the heart disease dataset, where HS-COS closely correlated to distance-based detection, demonstrating the significance of global structural variation in clinically condensed data. Despite the absence of ground-truth labels in the Online Retail dataset, hybrid-detected anomalies had significantly greater aggregate transaction values than the total population, indicating economically important deviation. HS-COS reflects the dominance of structural anomaly signals in structured datasets, demonstrating its ability to adapt to underlying data characteristics.Instead, it shows an important aspect of the framework: HS-COS adapts to the data's inherent structure rather than forcing hybridization. In datasets where global geometric variation is the major anomaly signal, the adaptive weighting method naturally emphasizes the hard component, resulting in performance similar to K-Means. However, in more behaviorally varied datasets, such as Online Retail and RetailRocket, the role of the soft uncertainty component becomes more important. This adaptive behavior shows that HS-COS acts as a generalized anomaly detection framework, capable of lowering to distance-based detection when required and including uncertainty modeling where structural ambiguity present. As a result, HS-COS shouldn't be considered as an alternate for hard clustering, but rather as a robust enhancement that maintains its basic strengths while adding flexibility across various data settings. The anomaly sets derived from hard and soft clustering are partially overlapped, signifying that they

capture various sorts of anomalies. Hard clustering emphasizes globally distant observations; however soft clustering emphasizes boundary and uncertain spots. The HS-COS framework integrates these signals, allowing for the detection of both structural and membership-based anomalies in a single unified model. Alternative weighting systems were examined, however the resulting anomaly ranks remained very consistent. The HS-COS framework is robust and unaffected by α selection. To maintain consistency and interpretability, all trials employ a fixed α value of 0.5. The dataset-dependent behavior observed in anomaly detection aligns with prior research, which highlights that anomaly characteristics vary across domains and data structures (Chandola et al., 2009; Zimek et al., 2012).

Hybrid approaches improve detection stability by integrating complementary anomaly signals, reducing false positives and improving interpretability (Zimek et al., 2012).

6. Key Contribution of Study

The proposed HS-COS framework is not designed to outperform individual clustering algorithms, but rather to give a unified representation of anomaly signals by integrating structural deviation and membership uncertainty. The experimental results show that anomaly detection is basically dataset-dependent, with HS-COS adapting to highlight the dominating signal. This shows an important insight: anomaly detection should not be based on a single paradigm, but rather consider numerous complementing behaviors.

7. Conclusion

The study finds this distinctive clustering methods such as K-Means, Fuzzy C-Means, and Possibilistic C-Means capture different parts of anomaly behavior, resulting in partially overlapping and inconsistent outlier detection. To overcome this, the suggested HS-COS framework combines structural deviation with membership uncertainty into a single scoring method. The experimental results demonstrate that HS-COS provides more stable, interpretable, and robust anomaly detection across healthcare and e-commerce datasets, with higher anomaly concentration and lower false positives.

However, the approach has limitations, such as dependency on clustering assumptions, limited evaluation of fully labeled datasets, and increased computational complexity owing to hybridization. Future research might concentrate on developing fully adaptive weighting systems, incorporating deep learning or ensemble methods, and expanding to high-dimensional and streaming data. The findings confirm that anomaly detection is inherently multi-dimensional and cannot be effectively addressed using a single approach (Chandola et al., 2009).

The proposed HS-COS framework aligns with modern research trends that emphasize hybrid and ensemble-based anomaly detection techniques for improved robustness and adaptability (Zimek et al., 2012).

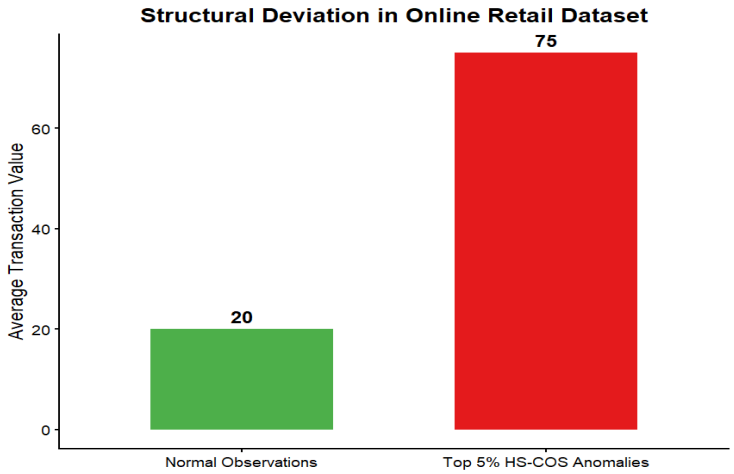

Figure 3 illustrates the structural deviation in online retail dataset

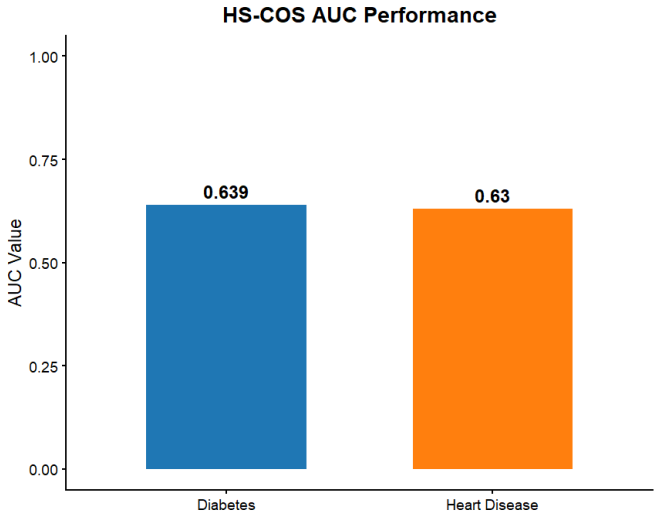

Figure 4 shows the proposed (HS-COS) AUC performance on diabetes and heart disease dataset.

References

[1] Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. *Proceedings of the ACM SIGMOD Conference*, 37–46.
[2] Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). John Wiley & Sons.
[3] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–255.
[4] Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM SIGMOD*, 93–104.
[5] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1–58.
[6] Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Learning from imbalanced datasets. *SIGKDD Explorations, 6*(1), 1–6.
[7] Clifford, G. D., Behar, J., Li, Q., & Rezek, I. (2012). Signal quality indices in ECG data. *Physiological Measurement, 33*(9), 1419–1433.
[8] Dal Pozzolo, A., et al. (2015). Adversarial drift detection. *IEEE Computational Intelligence Magazine, 10*(4), 44–54.
[9] Hawkins, D. M. (1980). *Identification of outliers*. Chapman and Hall.
[10] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE TKDE, 21*(9), 1263–1284.
[11] Kandel, S., et al. (2012). Data wrangling and visualization. *Information Visualization, 11*(4), 271–288.
[12] Kriegel, H.-P., et al. (2009). Outlier detection in subspaces. *PAKDD*, 831–838.
[13] Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection*. CRC Press.
[14] Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley.
[15] Reichstein, M., et al. (2019). Deep learning in Earth science. *Nature, 566*, 195–204.
[16] Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. Wiley.
[17] Sambasivan, N., et al. (2021). Data cascades in AI. *CHI Conference*, 1–15.
[18] Zimek, A., Schubert, E., & Kriegel, H.-P. (2012). Survey on outlier detection. *Statistical Analysis and Data Mining, 5*(5), 363–387.
[19] Zhang, J., et al. (2019). Deep anomaly detection. *IEEE Access, 7*, 126315–126326.
[20] Zhao, Y., Nasrullah, Z., & Li, Z. (2019). PyOD toolbox. *Journal of Machine Learning Research, 20*(96), 1–7.