# Univariate Outlier Detection: Precision-Driven Algorithm for Single-Cluster Scenarios

El hairach Mohamed Limam *, Insaf Bellamine *, Amal Tmiri

Posted Date: 30 April 2024

doi: 10.20944/preprints202404.2008.v1

Keywords: outlier detection, machine learning, univariate data analysis, data mining

 doi:10.20944/preprints202404.2008.v1



*Article*

# Univariate Outlier Detection: Precision-Driven Algorithm for Single-Cluster Scenarios

**Mohamed Limam El hairach [1,*], Amal Tmiri [2] and Insaf Bellamine [1]**

[1] Chouaïb Doukkali University, El jadida 24000, Morocco; insafbellamine20@gmail.com
[2] National School of Arts and Crafts, Mohammed V University, Rabat, Morocco; b_tmiri@yahoo.fr
* Correspondence: mohammed.limam25@gmail.com

**Abstract:** This study introduces a novel algorithm tailored for the precise detection of lower outliers in univariate datasets, particularly suited for scenarios with a single cluster. The approach leverages a combination of transformative techniques and advanced filtration methods to efficiently segregate anomalies from normal values. Notably, the algorithm emphasizes high precision, ensuring minimal false positives, and requires only a few parameters for configuration. Its unsupervised nature enables robust anomaly filtering without the need for extensive manual intervention. To validate its efficacy, the algorithm is rigorously tested using real-world data obtained from photovoltaic (PV) module strings with similar DC capacities, containing various anomalies. Results demonstrate the algorithm's capability to accurately identify lower outliers while maintaining computational efficiency and reliability in practical applications.

**Keywords:** outlier detection; machine learning; univariate data analysis; data mining

## 1. Introduction

Anomaly detection is pivotal across various sectors, such as finance, healthcare, and industrial monitoring, where accurately identifying anomalies is vital for operational efficiency and risk mitigation. In many applications, achieving high precision in anomaly detection is paramount, particularly in scenarios with a single cluster of data points where anomalies are less common but their detection holds significant consequences. In such contexts, false positives can result in costly disruptions or erroneous decisions, highlighting the importance of algorithms prioritizing precision.

Univariate outlier detection techniques, such as the three-sigma rule and boxplot, rely on statistical assumptions regarding the normal distribution of data. Additionally, alternative tests or criteria for univariate outlier detection, such as Grubbs' test or Chauvenet's criterion [1], may offer more effective solutions.

Moreover, the advent of unsupervised anomaly detection algorithms has brought about substantial advancements in anomaly detection methodologies. Unsupervised algorithms offer the distinct advantage of being able to operate without labeled training data, making them particularly well-suited for scenarios where labeled anomalies are scarce or difficult to obtain. This capability empowers these algorithms to autonomously discern anomalies from normal data points, enhancing their adaptability and scalability across various domains.

In this proposal, we address the need for high precision in anomaly detection, particularly in environments characterized by one cluster of data, where precision is of utmost importance. We propose the development and application of an unsupervised anomaly detection algorithm tailored to such scenarios, leveraging transformative techniques and advanced filtration methods to achieve precise anomaly detection. By harnessing the benefits of unsupervised learning, our algorithm aims to provide a reliable solution for accurately identifying anomalies while minimizing false positives, thereby enhancing decision-making and operational integrity in real-world applications.

 doi:10.20944/preprints202404.2008.v1



## 2. Related Works

In the late 1990s, Carling [2] proposed improvements to traditional outlier detection methods, particularly focusing on their application to non-Gaussian data and addressing the need for robust techniques in the presence of skewed distributions. Solak [3] discusses various methods available to detect outliers in univariate data sets, including Grubbs and Dixon tests, which can handle multiple outliers in some cases. Maciá-Pérez et al. [4] present an efficient algorithm grounded in Rough Set Theory for outlier detection, stressing its utility in decision-making processes. Jiang et al. [5] propose novel outlier-based initialization algorithms for K-modes clustering, emphasizing the importance of accurate cluster center selection. In [6], Sandqvist investigated non-parametric approaches for detecting outliers in survey data, highlighting the importance of addressing asymmetric distributions and heavy-tailed data prevalent in such datasets.

Subsequently, in [7], research on the Improved Boxplot for Univariate Data presented a modification to the traditional boxplot method to better handle skewed distributions, offering insights into alternative approaches for outlier detection in univariate data. These studies, spanning decades, collectively contribute valuable insights and methodologies for detecting anomalies in univariate datasets, providing a foundation for further research in this area.

Additional research has further enriched the field of detecting anomalies in univariate data, offering diverse perspectives and methodologies. Marsh and Seo [8] provided a comprehensive review and comparison of outlier detection methods for univariate datasets, offering insights into selecting appropriate techniques across varying data characteristics. In [9], a study on robust data mining approaches for industrial process modeling highlighted the importance of handling outliers and missing data, contributing robust techniques applicable to diverse datasets. Additionally, the introduction of novel clustering algorithms, such as DWMB in [10], presents alternative approaches for identifying patterns in data, which could complement traditional outlier detection methods. Furthermore, the proposal of reinforced Extreme Learning Machines for robust regression in the presence of outliers [11] addresses the robustness of machine learning algorithms, offering insights into enhancing outlier detection techniques in univariate data analysis. Lastly, the review on outlier and anomaly detection in time series data [12] provides a structured overview of techniques applicable to univariate datasets, offering insights into methods for detecting anomalies over time. Collectively, these studies contribute to the evolving landscape of detecting anomalies in univariate data, offering diverse methodologies and perspectives for further exploration.

## 3. Methodology

In this study, we propose an innovative algorithm designed for anomaly detection in univariate datasets, particularly when a singular cluster is expected. We assume that the data consist of positive real values.

**The Algorithm Steps:**

Step 1: Normalization

The initial step involves addressing the filtration of zero values from the dataset, considering their potential status as outliers contingent upon contextual considerations. Subsequently, the data is normalized to the range $[0, 1]$

Step 2: Lower Value Filtration

To establish a decisive boundary for anomaly identification, we filter out lower normalized values. The filtration threshold is set as the greater of either the median or the mean of the normalized values. This choice takes into account the positioning of the median and mean in positively or negatively

 doi:10.20944/preprints202404.2008.v1



skewed distributions. Values greater than the threshold are set to 0, while values below the threshold are replaced by the difference between the value and the reference value.

Figure 1 illustrates four distinct probability distributions: a normal distribution characterized by a symmetrical bell curve; a negatively skewed distribution, where the mean is typically less than the median and the tail extends towards lower values; a positively skewed distribution, where the mean is typically greater than the median and the tail extends towards higher values; and a uniform distribution, featuring constant probability density across the range of values. The same principle applies to other distributions, such as exponential, bimodal, and multimodal distributions, where the relationship between the mean and median may vary.

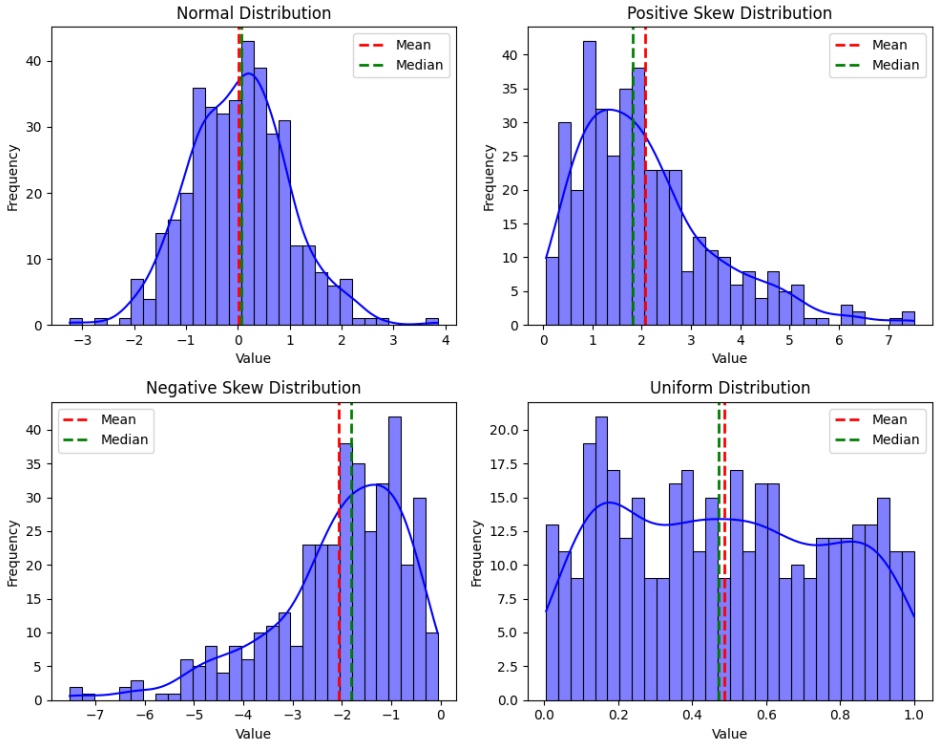


**Figure 1.** Illustrating random normal, negatively skewed, positively skewed, and uniform distributions.

**The filter used is as follows:**

$$\sigma(s^*, m, a) = \begin{cases} (s^* - \max(m, a))^2 & \text{if } s^* < \max(m, a) \\ 0 & \text{if } s^* \geq \max(m, a) \end{cases} \tag{1}$$

where:

$s^*$ is the normalized value

$m$ is the median of the normalized values

$a$ is the average of the normalized values

Given that the data is scaled within the range of 0 to 1, any value below this threshold will be substituted by the squared distance between the value and the threshold. This distance ranges from 0 to 1, with values closer to 0 indicating proximity to the threshold and values closer to 1 indicating greater distance. To differentiate between values closer to 0 and those closer to 1, the effect of the squared function is utilized.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 April 2024 doi:10.20944/preprints202404.2008.v1



When analyzing the influence of the square function within the range of 0 to 1, it's crucial to recognize its distinct impacts on numbers closer to 0 compared to those near 1. For values closer to 0, the square function significantly compresses the values, resulting in a substantial reduction in magnitude. This compression leads to a non-linear distribution where smaller values are proportionately closer together. Conversely, numbers closer to 1 experience a less drastic effect, with their magnitudes decreasing as well but not to the same extent as those near 0. This differential impact underscores the non-uniform transformational behavior of the square function, emphasizing its significance in data manipulation and analysis, particularly when dealing with ranges centered around 0.

Step 3: Hyperbolic Transformation

For a detailed exploration of our non-linear transformation strategy, Figure 2 presents the standard Hyperbolic Tangent Transformation (tanh) curve, providing a benchmark for understanding subsequent modifications.

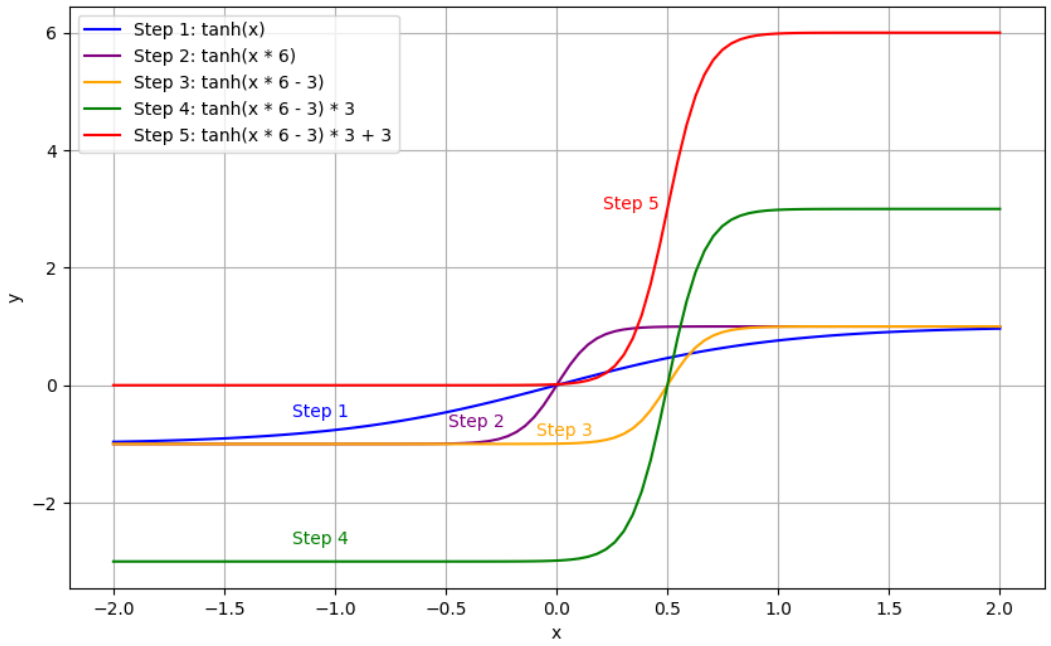


**Figure 2.** Modified Standard Hyperbolic Tangent Transformation (tanh) curve.

**The formula for the Modified Hyperbolic Tangent Transformation is:**

$$f(\sigma(s^*, m, a)) = \tanh(\alpha \cdot \sigma(s^*, m, a) - \beta) \cdot \lambda + \eta \tag{2}$$

**Parameters:**

- $\alpha$: Controls the slope of the transformation.
- $\beta$: Shifts the center of the tanh curve.
- $\lambda$: Scales the output to fit the desired range.
- $\eta$: Shifts the entire graph vertically.

**Justification for Parameter Choices:** The parameter $\alpha$ is set to 6, and the parameters $\beta$, $\lambda$, and $\eta$ are set to 3. The choice of 6 for scaling the input ($x$) influences the horizontal stretching of the sigmoid curve, making the algorithm more sensitive to variations in the data. The parameter 3, subtracted from the scaled input, shifts the center of the tanh curve horizontally, impacting where the transformation focuses attention within the data range.

 doi:10.20944/preprints202404.2008.v1



Suppose $x$ is an element in the range (0,1). The function $\tanh(\alpha \cdot x - \beta) \times \lambda + \eta$, set with parameters $\alpha = 6$, $\beta = 3$, $\lambda = 3$, and $\eta = 3$, expands the range from (0,1) to (0,6), showcasing its transformative capability in data manipulation.

**Transformation Steps:**

1. $\tanh(x)$: The initial graph has a sigmoid shape, ranging from -1, rising gradually, reaching a maximum slope around 0, and then descending back to -1.
2. $\tanh(x \cdot 6)$: scaling the input by a factor of 6 horizontally stretches the sigmoid curve, resulting in a more elongated shape compared to the original tanh function.
3. $\tanh(x \cdot 6 - 3)$: shifting the scaled input to the right by 3 units horizontally relocates the curve to the right.
4. $\tanh(x \cdot 6 - 3) \cdot 3$: Multiplying the values vertically scales the graph, increasing the amplitude threefold.
5. $\tanh(x \cdot 6 - 3) \cdot 3 + 3$: Adding 3 shifts the entire graph vertically upwards by three units.

In summary, the parameter choices of 6 and 3 reflect a well-considered balance achieved through experimentation, enhancing the algorithm's ability to highlight differences between normal and anomalous data points.

Step 4: Detection of Anomalies

The final result of the transformation splits outliers from the data. Outliers will have the greatest transformed values. By visualizing the final result, the threshold can be selected based on the range where data become split between anomalies and normal data.

Figure 3 provides a visual representation of the algorithm's flowchart, illustrating the interconnected steps involved in processing the data and identifying lower outliers.

 doi:10.20944/preprints202404.2008.v1



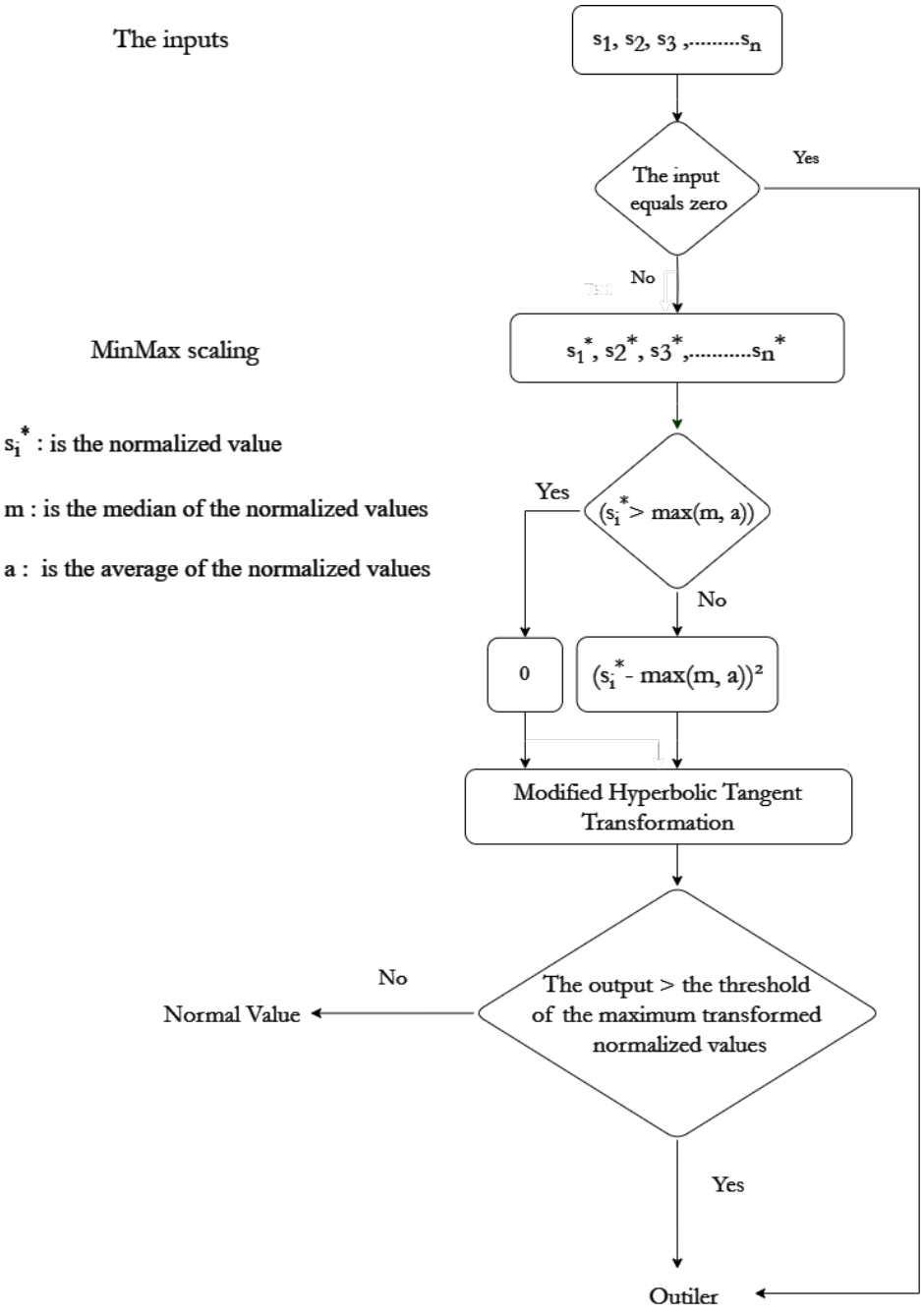


**Figure 3.** Flowchart of the Proposed Algorithm

## 4. Result and Discussion

In our study, we applied the proposed algorithm to real-world data collected from a Photovoltaic (PV) plant located in Morocco, specifically the Noor PV Ouarzazate (55 MW), which has been in operation for six years. Our focus was on the strings' current data of PV modules recorded over the course of a single day, where drops in current serve as indicators of anomalies. These anomalies encompass various faults such as hotspots, defective diodes, and degraded PV modules, all of which have been traditionally detected using tools like thermal imaging and I-V curve tests. Our objective was to leverage the proposed algorithm to precisely detect strings containing anomalies, thereby enhancing the efficiency of anomaly detection processes. Figure 4 provides a visual representation of the raw data before any processing, while Figure 5 illustrates the data after the application of our algorithm. Notably, the figure encapsulates the final stage of our algorithm, displaying the transformed

 doi:10.20944/preprints202404.2008.v1



data after the application of the customized tanh function. The plot effectively demonstrates how the non-linear transformation enhances the contrast between normal and anomalous data points, providing practical interpretability.

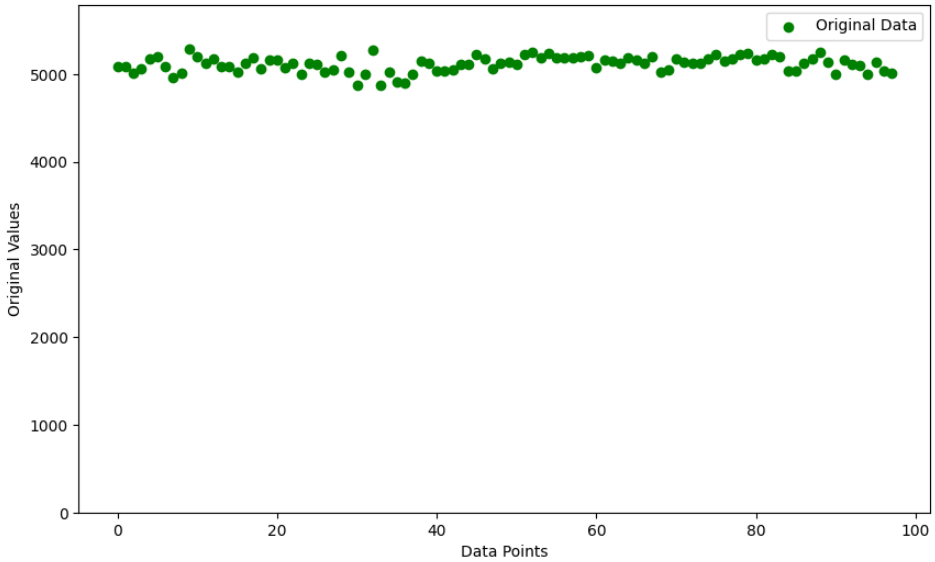


**Figure 4.** Initial state of the dataset before preprocessing.

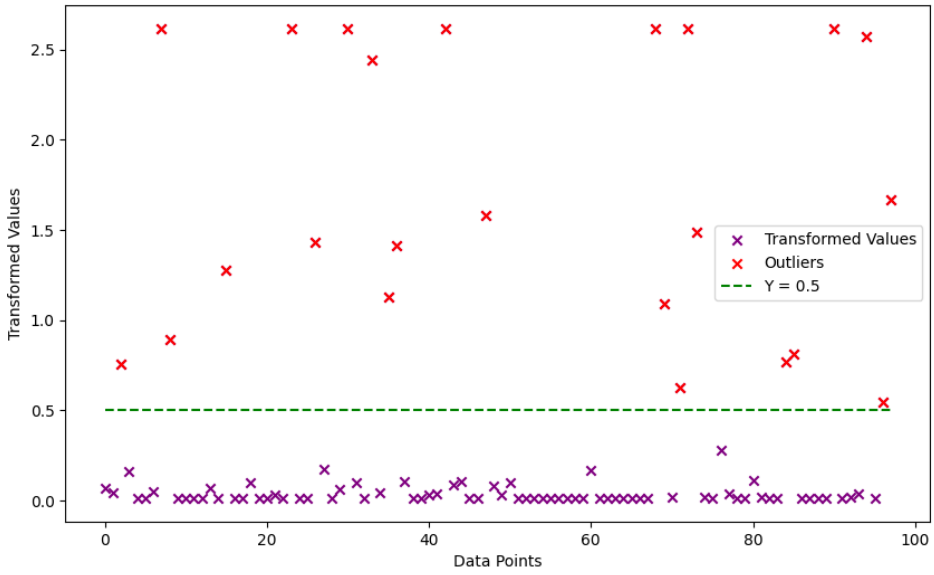


**Figure 5.** Transformed data after the application of the customized tanh function.

Our findings indicate promising results, with the algorithm achieving near-perfect precision across various threshold values as depicted in Figure 6. This high precision underscores the algorithm's efficacy in accurately identifying true anomalies, rendering it suitable for scenarios where precision is paramount. However, we observed a sensitivity in recall and accuracy to threshold adjustments, emphasizing the inherent trade-off between precision and other metrics. This flexibility in prioritizing precision underscores the adaptability of the algorithm to accommodate different application requirements and preferences.

The ability to achieve near-perfect precision renders the algorithm valuable in critical infrastructure monitoring and safety-critical systems. By minimizing false positives while maintaining high confidence in identifying true anomalies, the algorithm enhances operational efficiency and

 doi:10.20944/preprints202404.2008.v1



enables timely intervention in potential fault scenarios. However, it is crucial to acknowledge the trade-offs associated with threshold selection. Careful consideration of application-specific requirements is essential to optimize performance based on the desired balance between precision and other metrics.

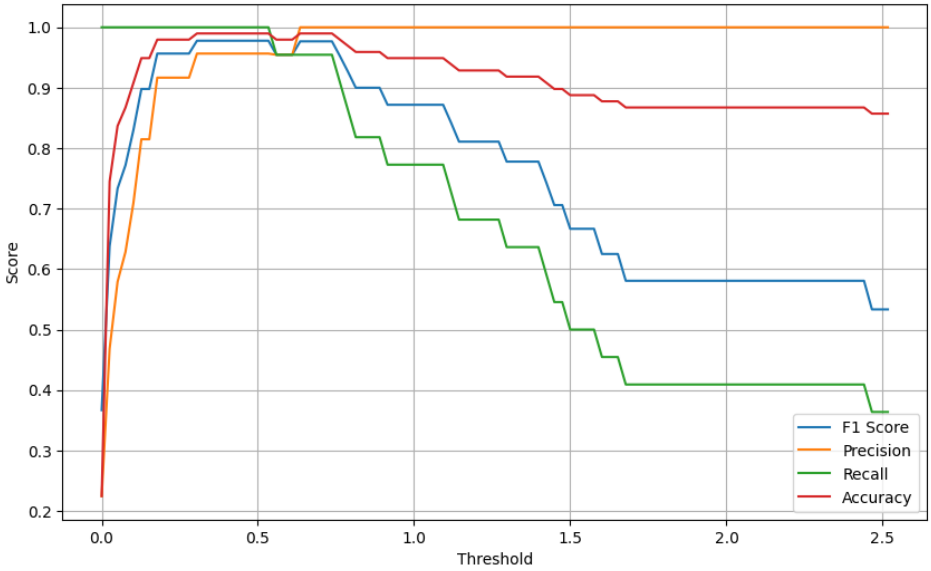


**Figure 6.** Variation of F1 score, recall, accuracy, and precision against changing threshold values.

## 5. Conclusions

In summary, our study introduces a novel algorithm designed for precise anomaly detection in univariate datasets, particularly in scenarios with a single cluster of data points. Applied to real-world data from the Noor PV Ouarzazate 55 MW plant in Morocco, our algorithm demonstrated near-perfect precision in identifying anomalies in photovoltaic module strings.

Through a combination of transformative techniques and advanced filtration methods, our algorithm effectively distinguishes anomalies from normal data points. This precision is crucial for applications where false positives can have significant consequences, highlighting the algorithm's efficacy in scenarios where precision is paramount.

Despite achieving near-perfect precision, we acknowledge the trade-offs associated with threshold selection, emphasizing the need for careful consideration of application-specific requirements. Nevertheless, the algorithm's ability to enhance operational efficiency and enable timely intervention in potential fault scenarios makes it invaluable for critical infrastructure monitoring and safety-critical systems.

In conclusion, our proposed algorithm represents a significant advancement in anomaly detection methodologies, offering a reliable and efficient solution for accurately identifying anomalies in various datasets. By prioritizing precision and adaptability, our algorithm holds promise for improving decision-making processes and ensuring operational integrity across different sectors.

**Funding:** This research received no external funding.

**Data Availability Statement:** The data used for this research paper are not publicly available due to restrictions imposed by the company.

**Conflicts of Interest:** The authors declare no conflicts of interest.

 doi:10.20944/preprints202404.2008.v1



## References


1. Ross, S. M., et al. "Peirce's criterion for the elimination of suspect experimental data." *Journal of Engineering Technology* 20.2 (2003): 38–41.
2. Carling, K. "Resistant outlier rules and the non-Gaussian case." *Computational Statistics & Data Analysis* 33.3 (2000): 249–258.
3. Solak, M. K. (2009). Detection of multiple outliers in univariate data sets. *Paper SP06-2009, Schering*.
4. Maciá-Pérez, F., Berna-Martinez, J. V., Oliva, A. F., & Ortega, M. A. A. (2015). Algorithm for the detection of outliers based on the theory of rough sets. *Decision Support Systems*, 75, 63–75. Elsevier.
5. Jiang, F., Liu, G., Du, J., & Sui, Y. (2016). Initialization of K-modes clustering using outlier detection techniques. *Information Sciences*, 332, 167–183. Elsevier.
6. Sandqvist, A. P., & KOFETH Zurich. "Detecting outliers in weighted univariate survey data." (2015).
7. Walker, M. L., Dovoedo, Y. H., Chakraborti, S., & Hilton, C. W. "An improved boxplot for univariate data." *The American Statistician* 72.4 (2018): 348–353.
8. Seo, S. "A review and comparison of methods for detecting outliers in univariate data sets." PhD thesis, University of Pittsburgh (2006).
9. Zhu, J., Ge, Z., Song, Z., & Gao, F. "Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data." *Annual Reviews in Control* 46 (2018): 107–133.
10. Rehman, A. U., & Belhaouari, S. B. "Divide well to merge better: A novel clustering algorithm." *Pattern Recognition* 122 (2022): 108305.
11. Frenay, B., & Verleysen, M. "Reinforced extreme learning machines for fast robust regression in the presence of outliers." *IEEE Transactions on Cybernetics* 46.12 (2015): 3351–3363.
12. Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. "A review on outlier/anomaly detection in time series data." *ACM Computing Surveys (CSUR)* 54.3 (2021): 1–33.