Machine Learning Approaches to Advanced Outlier Detection in Psychological Datasets

Original Scientific Paper

Khoula Al. Abri
Universiti Tenaga Nasional
Department of College of Computing and Informatics
Kajang, Selangor Malaysia
Khoula.Alabri@hotmail.com

Manjit Singh Sidhu
Universiti Tenaga Nasional
Department of College of Computing and Informatics
Kajang, Selangor Malaysia
manjit@uniten.edu.my

Abstract *– The core aim of this study is to determine the most effective outlier detection methodologies for multivariate psychological datasets, particularly those derived from Omani students. Due to their complex nature, such datasets demand robust analytical methods. To this end, we employed three sophisticated algorithms: local outlier factor (LOF), one-class support vector machine (OCSVM), and isolation forest (IF). Our initial findings showed 155 outliers by both LOF and IF and 147 by OCSVM. A deeper analysis revealed that LOF detected 55 unique outliers based on differences in local density, OCSVM isolated 44 unique outliers utilizing its transformed feature space, and IF identified 76 unique outliers leveraging its tree-based mechanics. Despite these varying results, all methods had a consensus for just 44 outliers. Employing ensemble techniques, both averaging and voting methods identified 155 outliers, whereas the weighted method highlighted 151, with a consensus of 150 outliers across the board. In conclusion, while individual algorithms provide distinct perspectives, ensemble techniques enhance the accuracy and consistency of outlier detection. This underscores the necessity of using multiple algorithms with ensemble techniques in analyzing psychological datasets, facilitating a richer comprehension of inherent data structures.*

Keywords*: Outlier Detection, psychological dataset, machine learning techniques, ensemble methods*

1. INTRODUCTION

Outliers are data points that differ substantially from the overall pattern of a dataset [1] and have long been a topic of interest in various scientific fields, including psychology. In questionnaire-based psychological research, outliers pose unique challenges [2]. Considering the inherent variability in human responses and psychological traits, the presence of outliers can severely affect the accuracy and consistency of the findings. This is mainly because outliers can skew statistical measures and lead to biased or misleading conclusions. Consequently, detecting and managing outliers is critical in ensuring the integrity and credibility of research in psychology [3].

Identifying outliers in datasets presents significant challenges, especially in the context of multivariate datasets [4]. Our research is specifically designed for detecting outliers in observational data. The driving force behind our approach is to enhance its robustness, preparing our dataset for deeper, more precise subsequent analyses [5 , 6].

However, traditional techniques encounter substantial limitations when applied to multivariate datasets, especially those that lack predetermined labels for outlier identification. These constraints expose a significant gap in our methodological framework, emphasizing the immediate need for the design and implementation of advanced techniques [7]. Ideally, such techniques would be capable of effectively addressing the complex characteristics inherent in multivariate psychological datasets.

With advancements in machine learning, several cutting-edge algorithms have emerged as promising alternatives for outlier detection. In this study, we examine three algorithms: LOF, OCSVM, and IF, which have demonstrated effectiveness in various models for outlier detection [8, 9, 10]. The LOF algorithm measures how isolated a data point is from its neighbors, allowing

us to identify outliers that have a significantly different density from their neighbors [8]. The OCSVM algorithm effectively determines the decision boundary that distinguishes outliers from the remaining data points [10]. The IF algorithm employs a unique approach based on randomly partitioning the dataset and isolating outliers based on the number of partitions [9]. We use an integrative approach by combining these algorithms using ensemble methods such as averaging, voting, and weighted combining approaches. This ensures the results are more reliable and stable. Ensemble methods have the advantage of mitigating the weaknesses of individual algorithms while harnessing their strengths [11].

Outlier detection in psychological datasets has attracted significant attention due to its crucial impact on the accuracy of findings. However, there is a noticeable gap in research on unlabeled multivariate datasets [4]. Most of the existing methods focus on outlier detection in numerical data, predominantly using proximity-based techniques [7]. In their research, "Ahmad A. A. Alkhatib" and "Qusai Abed-Al" explored the use of ensemble methods for outlier detection in multivariate a forest fire environment dataset. Their findings emphasized the advantages of leveraging multiple algorithms to enhance accuracy and robustness.

The main contribution of our research is the detailed analysis of a unique dataset derived from Omani students, consequently bridging a significant geographical and cultural gap in the literature. By utilizing three advanced outlier detection algorithms—specifically LOF, OCSVM, and IF—our study highlights the data-driven aspects of identifying fabricated responses in psychological datasets. To further enhance our methodological robustness, we have integrated ensemble techniques, achieving heightened accuracy in outlier detection. This multi-layered approach, when combined with a comparative analysis of algorithmic outputs, demonstrates our work's position at the intersection of psychology and machine learning. It effectively addresses critical methodological challenges and provides a robust framework to detect anomalies in multivariate psychological datasets.

This paper is structured as follows: Section II introduces the materials and methodology, explaining the algorithms and ensemble strategies employed for outlier detection. Section III details the results, presenting a comparative analysis of the algorithms and their performance alongside ensemble techniques. Section IV offers a discussion, focusing on the integration between individual algorithmic strengths and their combined capabilities. Finally, Section V concludes the study by highlighting insights from the examination of outlier detection algorithms and emphasizing the complex nature of anomalies in datasets.

2. MATERIALS AND METHOD

This research utilized a dataset obtained from tenth-grade students in Oman over four months, with the assistance of the Ministry of Education and career experts. The dataset comprises 1004 observations, each corresponding to a unique student. Within the dataset, there are 142 features, encompassing both psychological measures and demographic attributes. This research primarily focused on detecting outliers in the observations and did not delve into feature extraction or selection. Fig.1 displays a selected subset of the unlabeled multivariate psychological dataset.

Gender	Region	Parents Marital Status	Result in grade 9	Weight	Height
Female	Al Batinah North	Married	80% - 90%	50 kg - 60 kg	150 cm - 155 cm
Male	Musandam	Married	80% - 90%	80 kg - 90 kg	175 cm - 180 cm
Male	Musandam	Married	80% - 90%	70 kg - 80 kg	175 cm - 180 cm
Male	Musandam	Married	80% - 90%	70 kg - 80 kg	170 cm - 175 cm
Female	Al Batinah North	Married	70% - 80%	50 kg - 60 kg	160 cm - 165 cm
Female	Al Batinah North	Married	90% - 100%	40 kg - 50 kg	160 cm - 165 cm
Female	Al Batinah North	Married	80% - 90%	40 kg - 50 kg	150 cm - 155 cm
Female	Al Batinah North	Married	60% - 70%	50 kg - 60 kg	150 cm - 155 cm
Male	Ash Sharqiyah North	Married	50% - 60%	40 kg - 50 kg	170 cm - 175 cm
Female	Al Batinah North	Married	50% - 60%	30 kg - 40 kg	145 cm - 150 cm
Female	Al Batinah North	Married	80% - 90%	40 kg - 50 kg	145 cm - 150 cm
Female	Ad Dhahirah	Married	90% - 100%	30 kg - 40 kg	145 cm - 150 cm
Female	Ad Dhahirah	Married	60% - 70%	50 kg - 60 kg	150 cm - 155 cm
Female	Al Batinah North	Married	70% - 80%	30 kg - 40 kg	160 cm - 165 cm
Female	Ad Dhahirah	Married	80% - 90%	70 kg - 80 kg	150 cm - 155 cm
Female	Al Batinah North	Married	80% - 90%	50 kg - 60 kg	150 cm - 155 cm
Female	Al Batinah North	Married	80% - 90%	70 kg - 80 kg	160 cm - 165 cm

Fig. 1. A subset of unlabeled psychological

2.1. MULTIFACETED APPROACHES: THREE ALGORITHMS IN OUTLIER DETECTION

This research aims to identify outliers using LOF, OCSVM, and IF. These different algorithms were chosen because they are known to be flexible and successful in detecting outliers in high-dimensional datasets. These algorithms also exhibit excellent outlier sensitivity, indicating a strong capacity to distinguish between outliers and inliers.

Parameters play a pivotal role in determining the efficacy of an algorithm's outlier detection capability. For this specific dataset and outlier detection task, the techniques were optimized by adjusting these parameters [12]. Specific adjustments include the number of neighbors and contamination level for the LOF algorithm, the “Nu” value and kernel type for the OCSVM, and the number of estimators and contamination for the IF algorithm. Such parameters can significantly influence the sensitivity of outlier detection. Table 1 provides a detailed breakdown of these algorithm parameters. In this section, we detail the methodologies and experimental setups for both the selected algorithms and ensemble techniques.

A. Local outlier factor

Within the scope of this study, the LOF algorithm is configured with two primary parameters. The first parameter, "n_neighbors," is adjusted to 20. This defines the number of neighboring data points that the algorithm considers when computing local density, thereby allowing each data point to be understood about its 20 closest neighbors in the dataset. In our visual analyses, n_neighbors=20 showed an optimal balance, offering a clear distinction between outliers and inliers, and ensuring consistent outlier identification across the dataset [13].

As illustrated in Fig. 2, the choice of the number of neighbors, n_neighbors, significantly impacts the LOF scores. Observing the visual patterns, it becomes evident that setting n_neighbors to 20 provides a clear boundary between inliers and outliers. This decision was largely driven by visual analyses, ensuring a balance between detection accuracy and specificity while minimizing overlaps and potential over-sensitivity. The visuals effectively validate our selection, emphasizing its capability to accurately identify genuine outliers.

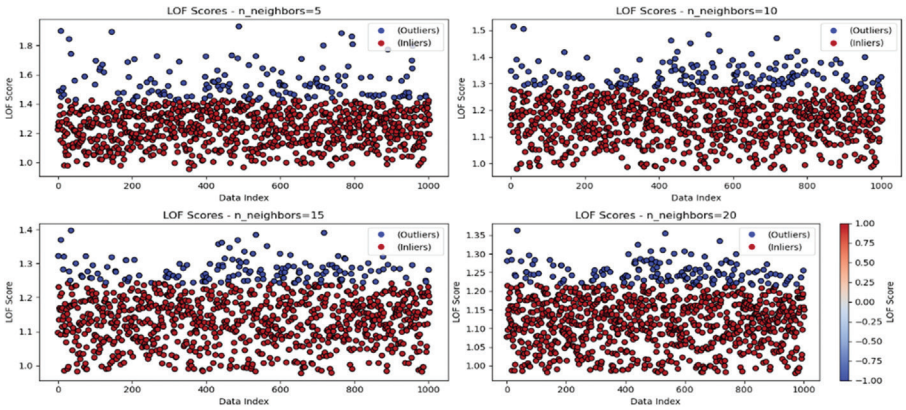

Fig. 2. Comparison of LOF Scores for Various Numbers of Neighbors

The second parameter, "contamination," is utilized to determine the threshold for outlier identification. By setting the "contamination" parameter to 0.15%, we define the acceptable ratio of observations that can be classified as outliers without raising false positives.

In the LOF algorithm, a key computation is the "reachability distance" (RD) between two data points, represented as y_1 and y_2 [14]. The RD incorporates both the local density around y_2 and the Euclidean distance separating y_1 and y_2. Mathematically, RD is the greater of two distances: the distance from y_2 to its k-th closest data point and the direct distance between y_1 and y_2.This is formally captured in equation (1) as:

$$\mathrm{RD}(\mathrm{y}_1, \mathrm{y}_2) = \max\{\mathrm{k} - \mathrm{dist}(\mathrm{y}_2), \mathrm{dist}(\mathrm{y}_1, \mathrm{y}_2)\} \quad (1)$$

Here, " *k-dist* (y_2)" represents the distance from y_2 to its k-th nearest neighbor, and " dist (y_1,y_2)" denotes the Euclidean distance between y_1 and y_2 [8].

Subsequently, we calculate the Local Reachability Density (LRD) for each data point, specifically y_1 [15]. LRD is essentially the inverse of the average RD of y_1 based on its k nearest neighbors. It involves summing the RD to each of y_1's k neighbors and computing the inverse of the average. LRD provides a quantitative measure of the density of data points in the vicinity of y_1. LRD is determined using the equation (2):

$$\mathrm{LRD}(\mathrm{y}_1) = \frac{1}{\frac{\Sigma \mathrm{RD}(\mathrm{y}_1, \mathrm{y}_2)}{\mathrm{K}}} \quad (2)$$

LOF evaluates the local density of y_1 in comparison to that of its neighbors. Data points with densities comparable to their neighbors possess LOF values around 1, while outliers with significantly lower local densities have LOF values much higher [8]. The equation (3) represents the LOF:

$$LOF(y_1) = \Sigma \frac{\left(\frac{\mathrm{LRD}(y_2)}{\mathrm{LRD}(y_1)}\right)}{\mathrm{k}} \quad (3)$$

Utilizing the LOF algorithm in this manner isolates outliers by analyzing variations in local density relative to neighboring data points. This ensures a robust and accurate distinction between outliers and inliers based on their surrounding data context.

B. One class support vector machine

The OCSVM algorithm was utilized with a specific configuration comprising a Radial Basis Function (RBF) kernel, a 'nu' parameter set at 15%, and a 'gamma' parameter defined as 'auto', which corresponds to 1 divided by the number of observations. The core functionality of OCSVM involves the identification of a hyperplane within a transformed feature space, which effectively segregates the dataset from the origin [16]. The fundamental objective is to ascertain the optimal hyperplane by minimizing the function using equation (4):

$$\frac{1}{2} * \left|\left|\mathrm{w}\right|\right|^2 + \left(\frac{1}{\nu * \mathrm{n}}\right) * \Sigma\xi_\mathrm{i} - \rho \quad (4)$$

subject to the constraints using equation 5:

$$\mathrm{w}^\mathrm{T}\varphi(\mathrm{y}_\mathrm{i}) \geq \rho - \xi_\mathrm{i}, \text{and } \xi_\mathrm{i} \geq 0 \quad (5)$$

The weight vector, represented by 'w', is fundamental in defining the decision hyperplane. This vector directly impacts the alignment of the hyperplane in the feature space. In contrast, φ represents the feature mapping function responsible for transforming input data into a different space. This transformation is driven by the selected kernel, which in our setup is the Radial Basis Function (RBF). The parameter 'ρ' is vital since it sets the decision boundary and is determined based on the dataset. Meanwhile, 'ξ' denotes the slack variables which grant the model some tolerance by allowing certain data points to reside on the undesirable side of the decision boundary. This flexibility is essential to strike a balance between optimizing the margin and managing outliers in the training set. The decision function is formulated in equation 6 as follows:

$$F(y) = \text{sgn}(w^T \varphi(y) - \rho) \quad (6)$$

Data points adhering to the criterion $w^T \varphi(y) - \rho > 0$ are located within the decision boundary and are consequently classified as inliers. Alternatively, instances for which $w^T \varphi(y) - \rho < 0$ are designated as outliers.

Moreover, the decision hyperplane is typified by the collection of points x in the feature space that satisfies the equation (7):

$$w^T \varphi(y) - \rho = 0 \quad (7)$$

This hyperplane acts as the decision boundary in the transformed feature space, with the points lying on one side classified as inliers and those on the opposite side classified as outliers [16].

C. Isolation forest

In this study, the IF algorithm is employed with a specific configuration that includes 100 trees (estimators) and a contamination level of 0.15%. The algorithm calculates an anomaly score, $S(y, x)$, for each data point y. This score is calculated using the equation (8):

$$S 2^{-\frac{E(h(y))}{c(x)}} \quad (8)$$

where $E(h(y))$ represents the average path length from the root to the terminal node for instance x across all trees, and $c(x)$ is the anticipated average path distance of a failed search in a Binary Search Tree [17]. To separate outliers from inliers, a threshold T is calculated based on the contamination level. This is done using the equation (9):

$$T = quantile\ (S,\ contamination\_Level\ ) \quad (9)$$

Data points with an anomaly score less than 0 are marked as outliers in equation (10), while those with an anomaly score equal to or greater than 0 are marked as inliers in equation (11). In mathematical terms:

$$x \text{ is an outlier if } S(x, n) < 0 \quad (10)$$

$$x \text{ is an inlier if } S(x, n) >= 0 \quad (11)$$

This method enables the algorithm to distinguish between inliers and outliers' data points effectively by isolating the points that have shorter paths in the trees, which usually signifies that they are less like the other data [18].

2.2. SYNERGY IN DETECTION: AN EXPLORATION OF ENSEMBLE METHODS

The objective of ensemble methods in outlier detection is to harness the strengths of multiple algorithms and enhance the overall performance and reliability of the outlier detection process. Rather than depending on a singular or single algorithm, ensemble methods combine predictions or outlier scores from multiple algorithms to make more robust decisions about the outlier status of data points. Ensemble methods in outlier detection employ different techniques to integrate the predictions or outlier scores from multiple algorithms [19].

A. Averaging Approach

This method calculates the average outlier scores provided by different algorithms for each data point. By taking the means of these scores, this approach seeks to provide a more consistent and balanced outlier score that reduces the individual biases inherent in any single algorithm [19]. Let S_{LOF}, S_{OCSVM}, and S_{IF} represent the outlier scores from LOF, OCSVM, and IF algorithms respectively for a particular data point [20]. The combined score $S_{Average}$ for that data point using the averaging approach is given by Equation 12.

$$S_{Average}=(S\_LOF+S\_OCSVM+S\_IF)/3 \quad (12)$$

B. Voting Approach

The voting approach in outlier detection determines the status of a data point based on a majority vote from multiple algorithms. Each of these algorithms produces a binary vote, classifying the data point as either an outlier (1) or an inlier (0) using a predetermined threshold [21]. S_{LOF}, S_{OCSVM}, and S_{IF} represent the binary classifications from the LOF, OCSVM, and IF algorithms, respectively, for a particular data point.

A data point is classified as an outlier if the sum of these classifications is 2 or 3, yielding an overall score of vote is 1. Conversely, if the sum is 0 or 1, the data point is deemed an inlier, giving the vote's score a value of 0. Formally, the voting can be represented in equation 13 as:

$$\begin{cases} 1 \ if\ (S_{LOF} + S_{OCSVM} + S_{IF}) \geq 2 \\ 0 \quad Othewise \end{cases} \quad (13)$$

C. Weighted Sum Approach

The weighted sum approach assigns variable weights to the outlier scores based on criteria such as the performance or reliability of each algorithm. By assigning different weights, it allows for giving more importance to the outlier scores from algorithms that have demonstrated better performance [22]. They $S_{weighted}$ are computed by multiplying the outlier scores S_{LOF}, S_{OCSVM}, and S_{IF} from LOF, OCSVM, and IF by their respective weights (W_{lof}, W_{OVSVM}, W_{IF}), and then summing them as shown in equation 14.

$$Sweighted = (SLOF \text{ x } Wlof + SOCSVM \text{ x } WOVSVM + SIF \text{ x } WIF) \quad (14)$$

3. RESULTS

In the results framework of our research, we investigate a methodical assessment of algorithmic benchmarks in outlier detection, comparing singular methodologies with the robustness of ensemble techniques. Our objective is to provide an analytical perspective on the effectiveness of individual algorithms and the synergistic capabilities of aggregated ensemble systems in detecting outliers (fabricated responses).

3.1. ALGORITHM PERFORMANCE IN OUTLIER DETECTION

Upon examining the outcomes from the LOF, OCSVM, and IF algorithms, we identified several critical

insights related to their effectiveness in detecting outliers. Firstly, the LOF algorithm demonstrated its capability by successfully identifying a total of 151 outliers from the dataset. Significantly, these outliers displayed an average outlier score of 0.6992, emphasizing their distinct characteristics when compared to the remaining data points. Furthermore, the standard deviation of the outlier scores was determined to be 0.7149, indicating a moderate level of variability in the identified outliers' scores.

In the provided scatter plot (Fig. 3), the Local Outlier Factor (LOF) scores depict data point anomalies based on local density. Most data points, represented as yellow dots, fall within the LOF score range for inliers: (0.9797, 1.2154), signifying they are inliers in compact regions. Conversely, the purple dots, scattered higher on the chart with scores ranging from 1.2161 to 1.3629, denote outliers in sparser areas. This distinction highlights the core concept of LOF: measuring data point deviation from a typical distribution. Thus, the plot provides a clear visual contrast between the inliers and the outliers.

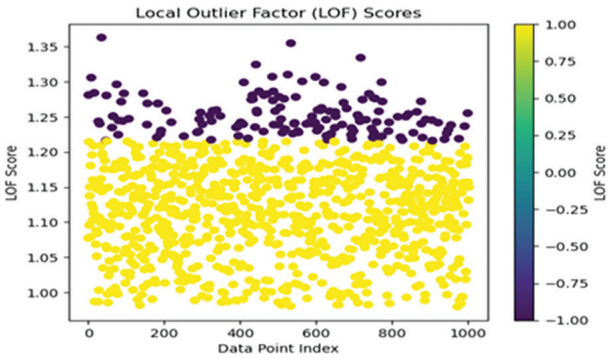

Fig. 3. Outlier detection for LOF algorithm

Subsequently, the OCSVM algorithm identified 147 outliers, which is slightly lower than anticipated. These outliers are depicted as red points in Fig. 4 and have an impressive average score of 1.840. The significant standard deviation of 2.271 for these scores indicates a broad range, bearing witness to OCSVM's sensitivity. Specifically, the OCSVM score range for inliers was (0.0000, 22.8469), and for outliers was (-9.8256, -0.0000).

Fig. 4 visually presents the difference between inliers and outliers. The dense cluster of blue points represents inliers, but they don't congregate close to zero. Some blue inliers, especially those around the 15 mark, illustrate the unique nature of the data. Despite their distance from the main group, they're still categorized as normal. On the other hand, the scattered red points show the outliers' positions, demonstrating OCSVM's ability to detect points that deviate significantly from standard patterns.

Finally, the Isolation Forest (IF) algorithm identified an equivalent number of outliers as the LOF algorithm. The IF algorithm, however, displayed a lower average outlier score of 0.0263, signifying a notable deviation from the mean score. Furthermore, the standard deviation for these scores was significantly reduced, indicating a more concentrated distribution of anomaly scores among the detected outliers, suggesting that the IF algorithm can identify outliers that closely conform to the standard deviation, differentiating it from the LOF and OCSVM algorithms.

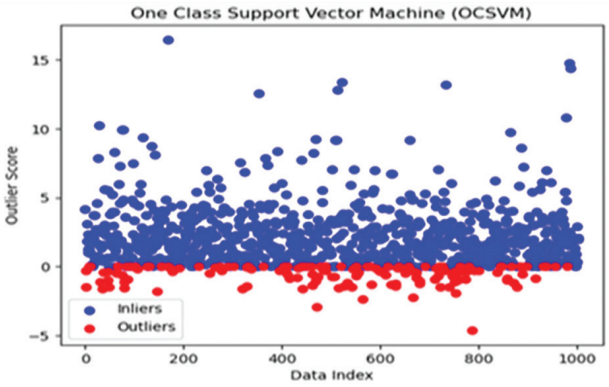

Fig. 4. Outlier detection of OCSVM algorithm

Fig. 5 presents the results obtained from the IF algorithm. The blue points represent inliers, clustering mostly around the higher regions of the graph. Specifically, the score range for inliers was between (0.0000, and 0.0946). Conversely, the red points, dispersed mostly in the lower part, highlight the outliers. The score range for these outliers was between (-0.0883, -0.0000).

This distinction emphasizes the efficiency of the IF algorithm in distinguishing between regular data points and those that deviate from the norm. The concentration of red points near the -0.000 to -0.040 range further highlights the close alignment of detected outliers with the mentioned standard deviation. IF algorithm demonstrates a precise and efficient approach to outlier detection.

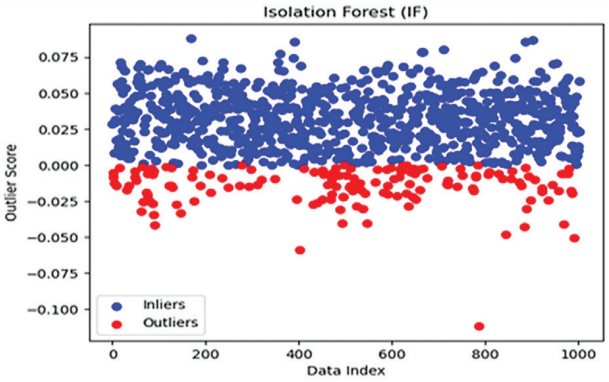

Fig. 5. Outlier detection for IF

Table. 1. Outlier Detection Overview

	LOF	OCSVM	IF	Common
Total Outliers	155	147	155	46
Unique Outliers	55	44	76	

Table 1 provides a comparative analysis of outliers detected using the three algorithms: LOF, OCSVM, and IF. Both the LOF and IF algorithms pinpointed the same number of outliers, amounting to 155, while the OCSVM identified a slightly fewer number, 147. On initial observation, this similarity in count might indicate consistent and standardized results across the three models. However, a deeper investigation of the data reveals nuanced variations.

When examining the unique outliers detected by each algorithm, we observe variations that suggest different underlying algorithms and sensitivities in anomaly detection. LOF, which utilizes the method of measuring local deviation of density relative to its neighbors, identified 55 unique outliers. OCSVM, using the SVM technique to separate normal and outlier data points in a transformed feature space, detected 44 unique outliers. The IF algorithm, employing tree-based partitioning strategies, identified 76 unique outliers. This difference in the unique anomalies found by each algorithm highlights the variations in their detection mechanisms.

Furthermore, even with the proximity in the overall outlier counts, the actual number common across all three algorithms is only limited to 41 outliers. This signifies that while the models have some overlapping observations, they also bring their unique perspectives in anomaly detection.

In essence, while the three algorithms display similar overall counts of outliers, their observations diverge considerably. Such variations underline the importance of employing multiple algorithms in anomaly detection tasks to capture a broader and more varied understanding of the underlying data.

3.2. OUTLIER DETECTION THROUGH ENSEMBLE TECHNIQUE

The Averaging algorithm consistently detected 155 outliers across various iterations. This consistency in identifying outliers illustrates both the robustness of the method and the effective coordination among the integrated algorithms. Fig. 6 provides an in-depth visualization of the Averaging approach. On the x-axis, data points range from 0 to 1000, while the y-axis presents combined scores, which extend from -1 to 1. The continuous blue line signifies the averaged combined scores for each data point. The distribution of these scores appears quite consistent, indicating a steady data trend.

The red dots concentrated mainly in the lower section of the graph represent the detected outliers. Their position suggests that these points deviate from the average scores by a significant margin. The clustering of the outliers around specific score values emphasizes the effectiveness of the Averaging approach in detecting anomalous data points that deviate from the standard.

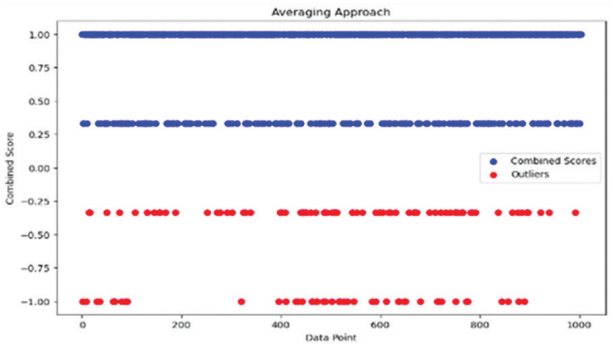

Fig. 6. Averaging approach

Similarly, the Voting method consistently detected 155 outliers across multiple iterations. The coherence in output illustrates the compatibility among the involved algorithms and confirms the efficiency of the Voting technique. In Fig. 7, we observe the Voting method in detail. As in the previous graph, the horizontal axis plots the data points, while the vertical axis illustrates voting scores, which range between -3 and 3.

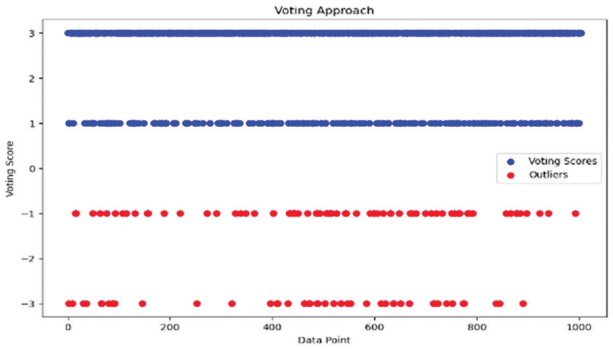

Fig. 7. Voting approach

Blue dots indicate the inliers, providing a visualization of how each data point was interpreted across the ensemble. Most of these scores are clustered within a narrow range, demonstrating a consensus among the involved algorithms. In contrast, the red dots, which represent the outliers, differ significantly from most of the blue dots. This points out the consistent identification of these points as anomalies across multiple iterations.

Subsequently, the Weighted approach, demonstrating performance metrics comparable to the Averaging and Voting methods, detected 151 outliers across multiple iterations. This uniformity strengthens the reliability and precision of the Weighted system. Fig. 8 presents the results of the Weighted approach. The graph's design aligns with Fig. 6 and 7, featuring data points on the horizontal axis and weighted combined scores on the vertical, varying from -1 to 5.

Blue dots represent the weighted scores, indicating the variations in weights allocated to various data points. A large portion of these points have scores centered around the middle, implying a consistent distribution of weights from different algorithms. Conversely, the red dots, which denote outliers, are more spread

out compared to the Averaging and Voting methods, suggesting that the Weighted method provides a more detailed identification of outliers based on the ensemble algorithms' weighted scores.

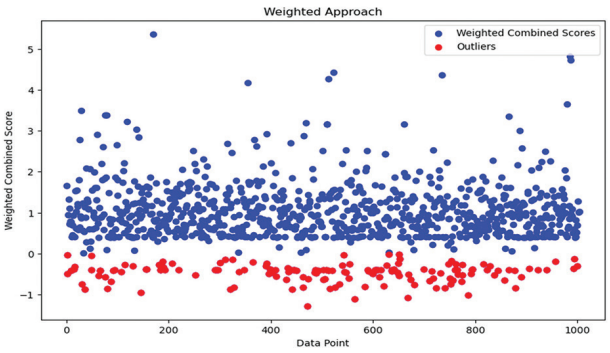

Fig. 8. Weighted approach

After a comprehensive evaluation of the three algorithms, it became evident that 150 outliers, interpreted as fabricated responses or imprecisions, were consistently identified, highlighting a robust consistency across the combined approaches.

4. DISCUSSION

In the domain of current algorithmic research, the synergy between the capabilities of individual algorithms and their combined implementations in outlier detection is crucial and remains a focal point in contemporary technology research. The findings presented in this study highlight the deep complexities embedded in these integrations.

From an individual algorithm perspective, LOF, OC-SVM, and IF each demonstrated unique strengths in detecting outliers, with their underlying principles and methodologies varying significantly. While there was a similarity in the overall counts, the number of outliers identified consistently by all three algorithms was relatively few, pointing to the distinct observation patterns intrinsic to each method. Such diversity indicates the notion that relying on a single detection technique might not capture all facets of outliers present in a dataset.

The table on outlier detection provides an insightful overview of the strengths and weaknesses of various algorithms. While the overall data indicates consistency in their results, a detailed analysis shows that each algorithm identifies distinct data points that might be missed by others. It's crucial to observe that even an algorithm with a high success rate might sometimes fail to detect certain outliers that other algorithms capture. This element emphasizes the significance of comprehending the functionality of each algorithm, especially when there's a requirement to optimize or adjust them for specific datasets.

A core finding from our research highlights the advantages of using integrated methods for outlier detection. The consistent results from techniques such as Averaging, Voting, and Weighted methods demonstrate the benefits of combining the capabilities of different algorithms. Each of these combined approaches identified outliers in the range of 151-155, showcasing their stable performance. Moreover, the identification of 150 common outliers among these methods shows a significant overlap, strengthening confidence in the accuracy of these aggregated results.

It's crucial to emphasize that the effectiveness of using combined methods is not only evident in their results but also in their ability to detect outliers potentially overlooked by individual algorithms. Striking a balance between comprehensive analysis and detailed outlier identification becomes essential for developers and researchers in determining the best detection method.

5. CONCLUSION

From the comprehensive exploration of outlier detection algorithms, it's evident that individual algorithms offer specific strengths in the field of data analysis. Their diverse observational patterns and the distinct outliers they detect illustrate the multifaceted nature of anomalies in datasets. In this study, the use of a combined approach proves invaluable. Techniques like averaging, weighted sum, and voting have been more consistent and trustworthy. When compared, single algorithms offer insights into some aspects of the data, but when combined, they present a more comprehensive view.

Thus, while individual algorithms provide unique perspectives on outlier detection, combined methods offer a holistic view. However, as with all computational methods, tailoring the approach to the specific needs of the dataset and research question is paramount. The results of this study pave the way for more in-depth discussions and research in the realm of outlier detection in computer science.

Given the complexity of contemporary datasets, it's essential for data analysts to continually refine and explore these combined methods. By doing so, they can more effectively interpret intricate data, thereby facilitating the identification of anomalous patterns. This, in turn, assists various industries in making informed decisions. Future studies should concentrate on optimizing the amalgamation of these techniques to ensure that data analysis remains robust and precise.

6. REFERENCES

[1] V. N. Tran, J. Kim, "Robust and efficient uncertainty quantification for extreme events that deviate significantly from the training dataset using polynomial chaos-kriging", Hydrology, Vol. 609, No. 6, 2022, p. 127716.

[2] R. Cao, X. Liu, J. Zhou, D. Chen, D. Peng, T. Chen, "Outlier Detection for Spotting Micro-expressions", Proceedings of the IEEE International Conference on

Bioinformatics and Biomedicine, Houston, TX, USA, 9-12 December 2021, pp. 3006-3011.

[3] X. Zhang, H. Ren, L. Gao, B. Shia, M. Chen, L. Ye, R. Wang, L. Qin, "Identifying the predictors of severe psychological distress by auto-machine learning methods", Informatics in Medicine Unlocked, Vol. 39, No. 4, 2023, p. 101258.

[4] K. Matsue, M. Sugiyama, "Unsupervised feature extraction from multivariate time series for outlier detection", Intelligent Data Analysis, Vol. 26, No. 6, 2022, pp. 1451-1467.

[5] S. Li, Y. Wang, X. Xie, "Prediction of uniaxial compression strength of limestone based on the point load strength and SVM model", Minerals, Vol. 11, No. 12, 2021, p. 1387.

[6] B. Dastjerdy, A. Saeidi, S. Heidarzadeh, "Review of Applicable Outlier Detection Methods to Treat Geomechanical Data", Geotechnics, Vol. 3, No. 2, 2023, pp. 375-396.

[7] H. M. Campbell, A. E. Murata, J. T. Mao, B. McMahon, G. H. Murata, "A novel method for handling pre-existing conditions in multivariate prediction model development for COVID-19 death in the Department of Veterans Affairs", Biological Methods and Protocols, Vol. 7, No. 1, 2022, pp. 1-18.

[8] O. Alghushairy, R. Alsini, T. Soule, X. Ma, "A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams", Big Data and Cognitive Computing, Vol. 5, No. 1, 2020, pp. 1-24.

[9] M. Heigl, K. A. Anand, A. Urmann, D. Fiala, M. Schramm, R. Hable, "On the Improvement of the Isolation Forest Algorithm for Outlier Detection with Streaming Data", Electronics, Vol. 10, No. 13, 2021, pp. 15-34.

[10] E. H. Budiarto, A. E. Permanasari, S. Fauziati, "Unsupervised Anomaly Detection Using K-Means, Local Outlier Factor and One-Class SVM", Proceedings of the 5th International Conference on Science and Technology, Yogyakarta, Indonesia, 30-31 July 2019, pp. 1-5.

[11] A. M. Jabbar, "Local and Global Outlier Detection Algorithms in Unsupervised Approach: A Review", Iraqi Journal for Electrical and Electronic Engineering, Vol. 17, No. 1, 2021, pp. 76-87.

[12] H. Alazzam, A. Sharieh, K. E. Sabri, "A lightweight intelligent network intrusion detection system using OCSVM and pigeon inspired optimizer", Applied Intelligence, Vol. 52, No. 4, 2022, pp. 3527-3544.

[13] S. Zhong, S. Fu, L. Lin, X. Fu, Z. Cui, R. Wang, "A novel unsupervised anomaly detection for gas turbine using Isolation Forest", Proceedings of the IEEE International Conference on Prognostics and Health Management, San Francisco, CA, USA, 17-20 June 2019, pp. 1-6.

[14] Z. Chen, K. Xu, J. Wei, G. Dong, "Voltage fault detection for lithium-ion battery pack using local outlier factor", Measurement, Vol. 146, No. 16, 2019, pp. 544-556.

[15] P. R. M. Júnior, T. Boult, J. Wainer, A. Rocha, "Open-Set Support Vector Machines", IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 52, No. 6, 2022, pp. 3785-3798.

[16] W. B. Richardson, J. Meyer, S. V. Solms, "Towards Machine Learning and Low Data Rate IoT for Fault Detection in Data-Driven Predictive Maintenance", Proceedings of the IEEE World AI IoT Congress, Seattle, WA, USA, 10-13 May 2021, pp. 202-208.

[17] W. Zhang, Z. Lin, X. Liu, "Short-term offshore wind power forecasting - A hybrid model based on Discrete Wavelet Transform (DWT), Seasonal Autoregressive Integrated Moving Average (SARIMA), and deep-learning-based Long Short-Term Memory (LSTM)", Renewable Energy, Vol. 185, No. 5, 2022, pp. 611-628.

[18] S. Borrohou, R. Fissoune, H. Badir, "Data cleaning survey and challenges – improving outlier detection algorithm in machine learning", Journal of Smart Cities and Society, Vol. 2, No. 3, 2023, pp. 125-140.

[19] S. Ardabili, A. Mosavi, A. R. Várkonyi-Kóczy, "Advances in machine learning modeling: Reviewing hybrid and ensemble methods", Proceedings of the International Conference on Global Research and Education: Engineering for Sustainable Future, Balatonfüred, Hungary, 4-7 September 2019, pp. 215-227.

[20] C. Wang, X. Li, J. Sun, X. Zhang, Y. Li, X. Peng, J. Liu, Z. Jiao, "Ensemble Learning Model of Power System Transient Stability Assessment Based on Bayesian Model Averaging Method", Proceedings of 6th IEEE Conference on Energy Internet and Energy System Integration, Changsha, China, 8-10 November 2022, pp. 1467-1471.

[21] R. Atallah, A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method", Proceedings of the 2nd International Conference on New Trends in Computing Sciences, Amman, Jordan, 9-11 October 2019, pp. 1-6.

[22] S. Cai, R. Sun, S. Hao, S. Li, G. Yuan, "An efficient outlier detection approach on weighted data stream based on minimal rare pattern mining", China Communications, Vol. 16, No. 10, 2019, pp. 83-99.