Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2025, 19(12), 402–419
https://doi.org/10.12913/22998624/209999
ISSN 2299-8624, License CC-BY 4.0

Received: 2025.07.08
Accepted: 2025.10.01
Published: 2025.11.01

Aggregating evaluation metrics for anomaly detection: a unified scoring approach

Alicja Rachwał[1*], Łukasz Gałka[1], Albert Rachwał[1], Paweł Karczmarek[1]

[1] Department of Computational Intelligence, Lublin University of Technology, Nadbystrzycka 38D, 20-618 Lublin, Poland
* Corresponding author's e-mail:alicja.rachwal@pollub.pl

ABSTRACT

This paper introduces a procedure that transforms multiple evaluation metrics into a single aggregated score, providing a comprehensive and interpretable summary of machine learning performance. The approach is demonstrated on a set of metrics obtained from various anomaly detection algorithms based primarily on Isolation Forest. Seven relevant performance metrics are aggregated using diverse techniques, including the arithmetic mean, weighted mean, Choquet integral, the OWA operator, and several Smooth OWA variants based on different interpolation Newton-Cotes quadratures. For methods requiring them, two distinct sets of weights are used. The results show that, particularly in anomaly detection tasks where individual metrics may lead to inconsistent evaluations, the aggregated score effectively reflects metric preferences and enables quick identification of the best-performing algorithm for a given dataset.

Keywords: multi-criteria decision making, model ranking, metric aggregation, evaluation metrics, OWA operator, Choquet integral.

INTRODUCTION

The widespread usage of machine learning and artificial intelligence has made model evaluation a central concern in both academic and applied research. Whether working with classical algorithms or state-of-the-art AI systems, practitioners face the challenge of selecting evaluation strategies that meaningfully reflect model behavior across diverse datasets and tasks. Far from being a mere afterthought, evaluation metrics now play a critical role in shaping how results are interpreted, compared, and communicated, ultimately influencing which models are trusted, deployed, or discarded. Across current studies, there emerges a common thread. While a wide variety of evaluation metrics have been proposed, each with their own assumptions and domains of applicability, there is still no consensus on how to systematically compare these metrics. There is a necessity to introduce a single metric that aggregates classical measures in order to provide an unambiguous evaluation of classifier quality, especially under complex conditions such as class imbalance or temporal ambiguity. Such fusion enables more coherent model comparison, particularly when individual metrics yield conflicting or ambiguous results. In this context, techniques from multi-criteria decision making (MCDM), including the ordered weighted averaging (OWA) operator [1] and the Choquet integral [2] offer robust tools for combining performance indicators while capturing user-defined preferences or dependencies among metrics. These methods go beyond conventional arithmetic mean by incorporating notions of importance, interaction, and non-linearity, making them especially suitable for high-stakes evaluations such as anomaly detection under class imbalance.

Challenges in model evaluation and metric diversity

Although metrics are a tool rather than a topic of their own, there are some academic papers emerging on new methodologies for evaluating the quality of machine learning models. The

paper [3] introduces a benchmark-based methodology for aggregating software quality metrics into ratings. Instead of relying solely on direct numerical aggregation (such as averages or inequality indices like Gini), the authors propose a two-stage process. Firstly, raw metric values are mapped to risk profiles using thresholds derived from benchmark datasets. Secondly, these profiles are converted into interpretable ratings. The approach is demonstrated on a benchmark of 100 software systems, showcasing its applicability to real-world quality assessment scenarios and its robustness to data variations. In [4] the challenge of aggregating low-level software quality metrics into meaningful system-level indicators is considered. The Squale model is introduced, which uses normalization and weighting to unify diverse metrics. The method is validated on the Eclipse project and compared against traditional and inequality-based aggregations. The methodology proposed in [5] includes the use of the Choquet integral to aggregate multi-dimensional quality indicators in the context of data fusion from heterogeneous sources. Although its focus is on data quality dimensions such as freshness and consistency, the paper offers valuable insight into how conflicting criteria can be systematically combined into a unified score. This methodology is conceptually aligned with efforts in model evaluation that aim to reconcile diverse performance metrics into coherent, interpretable outcomes. The paper [6] introduces the unified performance measure (UPM), a modified F1-score designed to better handle imbalanced classification problems. UPM is tested on synthetic and real datasets, showing superior stability and informativeness compared to classical metrics. It offers a promising direction for standardizing binary classifier evaluation.

Many scientific papers focus on potential problems arising from the incorrect application of some metric. The widespread use of reciever operating characteristic (ROC) curve for imbalanced binary classification tasks is criticized in [7]. The authors argue that such plots can mislead performance interpretation. They advocate for precision-recall (PR) curve, which more accurately reflect classifier behavior when positive cases are rare. Experimental and theoretical results support PR curves as a more reliable evaluation tool in real-world imbalanced scenarios. In [8] a theoretical relationship between ROC and PR curves is explored, particularly in the context of imbalanced datasets. They demonstrate that dominance in ROC space implies dominance in PR space and introduce the concept of the achievable PR curve. The study also highlights that improving the ROC AUC metrics may not result in optimal PR AUC performance, which is critical for algorithm evaluation. The paper [9] investigates how to select optimal decision thresholds for classifiers to maximize the F1 score in binary and multilabel contexts. It derives theoretical thresholds under different assumptions, such as well-calibrated probabilities or uninformative classifiers. Results reveal unintuitive behaviors and underscore the need for careful threshold selection in imbalanced classification settings. In [10] the Matthews correlation coefficient (MCC) is compared with F1 score and accuracy in binary classification, especially under class imbalance. They demonstrate that MCC offers more balanced and informative evaluations by incorporating all four elements of the confusion matrix. Experimental results across synthetic and real datasets support MCC as a superior and more reliable evaluation metric. The research [11] critically examines the ROC area under the curve (AUC) as a performance metric for classification systems, highlighting a fundamental inconsistency: AUC implicitly applies different misclassification cost assumptions across classifiers. The author argues that such a practice is incoherent since misclassification costs are inherent to the classification problem, not dependent on the classifier. As a solution, a coherent alternative metric is proposed that preserves consistency in cost assumptions, offering a more reliable basis for comparative evaluation.

Related work on metric aggregation

The article [12] presents a statistically grounded comparison of five text classification algorithms across varying category distributions and training data volumes. While it does not propose new evaluation metrics, it highlights how model performance can fluctuate significantly depending on dataset characteristics, implicitly demonstrating the limitations of fixed evaluation criteria. These observations support the case for data-sensitive model ranking, reinforcing the need for adaptable and context-aware evaluation strategies. The study [13] examines the performance of two variants of the Naive Bayes classifier, multivariate Bernoulli and multinomial, across multiple text classification tasks. Although not framed as a metric-focused study, its detailed

empirical comparisons reveal how performance is influenced by factors such as vocabulary size and feature representation. These findings underscore the importance of context-aware analysis, echoing broader challenges in constructing reliable ranking frameworks that account for data-specific behavior. The review paper [14] offers a broad overview of commonly used evaluation metrics in machine learning, particularly in classification tasks across binary, multi-class, and multi-label settings. It highlights the limitations of traditional metrics such as accuracy and F1-score when used in isolation, and emphasizes the lack of standardization in metric usage as a major obstacle for meaningful model comparison. The authors of [15] present a detailed taxonomy of 20 evaluation metrics used in time series anomaly detection. They define a set of desirable metric properties and evaluate each metric's suitability through case studies and experiments. Their analysis underscores the need for domain- and task-specific metric selection, reinforcing the importance of metric-aware model evaluation frameworks. In [16], affiliation precision/recall is proposed, which is a new class of evaluation metrics designed for anomaly detection in time series. Unlike traditional metrics, these are parameter-free, interpretable, and resilient to adversarial predictions. Their framework enables local, fine-grained evaluation of detection quality, addressing significant shortcomings in conventional approaches. The study [17] evaluates the limitations of traditional metrics, such as accuracy and F1-score, when applied to anomaly detection in time series data from industrial control systems. The authors propose an improved, range-based evaluation metric by modifying the Time-series aware precision and recall (TaPR) to account for ambiguous temporal boundaries of anomalies. Their approach highlights the importance of tailoring metrics to data characteristics and operational contexts, reinforcing the inadequacy of point-based evaluation for real-world detection tasks. The authors of [18] present a mathematical extension of classical precision and recall tailored to detect and evaluate range-based anomalies in time series data. Their model allows for customization based on domain-specific priorities, bridging the gap between generic metrics and real-world anomaly detection needs. This work offers a formal foundation for evaluating detection systems beyond point-based accuracy, aligning closely with efforts to develop composite, context-aware evaluation frameworks.

Motivation for a unified aggregated metric

Across a variety of recent studies in anomaly detection, classification and image analysis, a common dependency emerges: the effectiveness of the algorithms is closely tied to the choice of evaluation metrics and similarity measures. Whether in visual imperfection detection under weak supervision [19], deep learning-based medical diagnostics [20], patient survival prediction [21] or ensemble-based intrusion detection systems [22], model performance is typically reported through standard metrics such as accuracy, precision, recall, or F1 score. Similarly, in unsupervised anomaly detection approaches, such as the modified negative selection algorithm [23], evaluation hinges on metrics capable of reflecting sensitivity to rare events, often prioritizing recall at the cost of precision. In high-dimensional industrial cybersecurity contexts [24] or general intrusion detection benchmarks involving multiple classifiers [25], the metric space directly shapes how model output is interpreted and compared. These examples illustrate the diversity of metric requirements across domains – ranging from robustness to class imbalance, to interpretability, to discriminative sensitivity in complex feature spaces.

Despite the rich spectrum of available metrics, this diversity often presents challenges rather than clarity. Researchers must navigate a wide range of evaluation criteria, each emphasizing different aspects of performance sometimes leading to contradictory conclusions and inflated analytical complexity. Furthermore, strong correlations and redundancies among metrics can distort the overall assessment of model performance and hinder efforts to construct consistent and meaningful rankings of algorithms. These issues are particularly pronounced in anomaly detection, a class of problems often treated as binary classification. Here, the task involves identifying observations that significantly deviate from the norm, and is especially sensitive to class imbalance. While accuracy may suffice for multiclass classification, binary tasks require a broader and more nuanced set of metrics, such as precision, recall, F1-score, specificity, ROC AUC, PR AUC and balanced accuracy to capture performance reliably. Faced with this multitude of options, researchers often struggle to determine which metrics to prioritize, and whether their conclusions hold consistently across different measures. To address this, this study proposes a unified evaluation framework:

a composite performance metric designed to integrate the strengths of individual criteria while reducing ambiguity and inconsistency in model assessment. This unified approach seeks to simplify the evaluation process, enhance interpretability, and support the construction of robust rankings, particularly in challenging settings such as anomaly detection, where classical metrics may yield misleading insights due to inherent data imbalances.

We also aim to use the most recent, refined aggregation methods. Recent development in metric aggregation emphasize generalized operators and smoothing techniques to enhance interpretability and robustness. [26] introduce smooth OWA operators inspired by Newton-Cotes quadratures, demonstrating notable gains in classification accuracy through pre-aggregation smoothing. Similarly, in [27] a generalized and smoothed variant of the Choquet integral are proposed, retaining its theoretical properties while improving numerical precision. These innovations reflect a broader trend toward flexible, mathematically grounded aggregation strategies suited for high-stakes, multi-metric evaluation scenarios.Unlike simple averaging or heuristic scoring, our method leverages structured aggregation and multi-criteria ranking techniques to produce a consistent evaluation output across different model types and tasks.

BACKGROUND

Anomaly detection metrics

Among the metrics that determine the quality of an anomaly detection algorithm, a division can be made into those dependent and independent of the decision threshold. To the first group belong precision, recall, their combination F1 score, specificity, accuracy. All these metrics are based on different measurements of the proportion of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) cases. Metrics independent of the decision threshold include ROC AUC and PR AUC, measuring the area under a respectively defined curve. These metrics, in addition to anomaly detection, can be used identically for binary classification algorithms, and in modified form for multiclass classification, where the metric is calculated separately for each class and the results are averaged across all classes.

Precision indicates the proportion of samples predicted as anomalies that are actually anomalous.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall (sensitivity, true positive rate) measures the proportion of actual anomalies that are correctly identified.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

F1 score is the harmonic mean of precision and Recall. This metric balances both aspects and is especially useful when a trade-off between false alarms and missed detections is required.

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Specificity (true negative rate) indicates the proportion of normal samples correctly identified as normal.

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

Accuracy represents the overall proportion of correctly classified samples, including both anomalies and normal.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

In anomaly detection, accuracy can be misleading if the data is highly imbalanced, e.g. 95% normal cases.

ROC AUC (area under the receiver operating characteristic curve) represents the area under the ROC curve, which plots the true positive rate ($TPR=\frac{TP}{TP+FN}$) versus the false positive rate ($FPR=\frac{FP}{FP+TN}$) at various decision thresholds. The area is usually calculated by the trapezoidal rule, which involves approximating the area under the ROC curve by dividing it into trapezoids with vertical lines at the FPR values and horizontal lines at the TPR values. Then, the area is computed by summing the areas of the trapezoids. ROC AUC equal to 1 indicates excellent classification performance, 0.5 corresponds to random guessing.

PR AUC (area under the Precision-Recall curve) represents the area under the curve plotting Precision against Recall at various thresholds. It

is particularly suitable for evaluating models on imbalanced datasets. For various thresholds, Precision-Recall pairs are computed, thus forming the PR curve. The area is calculated usually by the trapezoidal rule, similarly as for ROC AUC. A high PR AUC score indicates that the model detects anomalies effectively without generating too many false positives.

Aggregation methods

There exists a vast number of aggregation methods, including the most basic ones like the mean and weighted average. In this section, a few more sophisticated ones are recalled, such as the OWA operator and Choquet integral, and recent modification of the OWA operator – its smoothed version.

OWA [28] is a flexible aggregation method which applies weights not to specific components, but to their ordered values.

$$OWA(x_1, x_2, \dots, x_n) = \sum_{i=1}^{n} w_i \cdot x_{(i)} \tag{6}$$

where: $\sum_{i=1}^{n} w_i = 1$ and $x_{(i)}$ is the i-th largest value in the vector $(x_1, x_2, \dots, x_n)$.

Smooth OWA operator [26] is a modification of the OWA operator, associated additionally with smoothing method denoted as Q:

$$SmoothOWA(x_1, x_2, \dots, x_n) = \\ = \sum_{i=1}^{n} w_i \cdot Q(x_{(i)}) \tag{7}$$

where: $Q(x_{(i)})$ means an application of a chosen Newton-Cotes formula to the element $x_{(i)}$. Let us recall few Newton-Cotes quadratures that can be used for such smoothing:

$$Q_S(x_{(i)}) = \frac{1}{6}x_{(i-1)} + \frac{2}{3}x_{(i)} + \frac{1}{6}x_{(i+1)} \\ (Simpson's\ quadrature) \tag{8}$$

$$Q_{\frac{3}{8}}(x_{(i)}) = \frac{1}{8}x_{(i-1)} + \frac{3}{8}x_{(i)} \\ + \frac{3}{8}x_{(i+1)} + \frac{1}{8}x_{(i+2)}\ \ (^3/_8\ quadrature) \tag{9}$$

$$Q_T(x_{(i)}) = \frac{1}{2}x_{(i)} + \frac{1}{2}x_{(i+1)} \\ (trapezoidal\ quadrature) \tag{10}$$

$$Q_{ONC3}(x_{(i)}) = \frac{2}{3}x_{(i-1)} - \frac{1}{3}x_{(i)} + \frac{2}{3}x_{(i+1)} \\ (3 - point\ Open\ NC\ quadrature) \tag{11}$$

With such operation of applying the Newton-Cotes quadratures, each element $x_{(i)}$ is smoothed by its neighboring elements in the vector of sorted input. Note that if the index of an element is less than 1, we take the value of $x_{(1)}$ instead of that element, and if the index is greater than n, we take $x_{(n)}$.

The Choquet integral [29] is a nonlinear aggregation function that not only considers the magnitude of input values but also the interactions between them, as defined by a fuzzy measure g. It is especially useful when the importance of a group of criteria depends on their combination, not just individual relevance. Let (X, Ω) be a measurable space and h: X → [0,1] be an Ω -measurable function. The Choquet integral of function h with respect to a fuzzy measure g [30] is expressed as

$$\mathrm{Ch} \int h \circ g = \sum_{\substack{i=1 \\ g(A_i), x_{(i+1)} = 0}}^{n} \left(h(x_{(i)}) - h(x_{(i+1)})\right) \tag{12}$$

where: $x_{(i)}$ are sorted inputs, $A_i = \{x_{(i)}, \dots x_{(n)}\}$ and g is a monotonic set function – Sugeno λ-fuzzy measure. The fuzzy measure g of any subset $\cup_{i \in I}\{x_{(i)}, \dots x_{(n)}\}$, where I ∈ {1, 2, ..., n} is calculated as

$$g\left(\bigcup_{i \in \mathbf{I}} \{x_{(i)}\}\right) = \frac{1}{\lambda}\left[\prod_{i=1}^{n}(1 + \lambda g_i) - 1\right] \tag{13}$$

which can be also presented in the form of

$$g(A_{i+1}) = g(A_i) + g(\{x_{(i+1)}\}) + \\ + \lambda g(A_i) g(\{x_{(i+1)}\}) \tag{14}$$

The value of λ ∈ (–1, 0) ∪ (0, ∞) is obtained from a polynomial eqaution solved for λ:

$$1 + \lambda = \prod_{i=1}^{n}(1 + \lambda) g_i \tag{15}$$

where: $g_1 = g(\{x_{(1)}\}, g_2 = g(\{x_{(2)}\}, g_n = g(\{x_{(n)}\}$.

The Choquet integral generalizes the weighted mean and can model redundancy or synergy between features. It is widely used in multicriteria decision making and information fusion. In a practical application, the measure g can be used

to model the relevance of various information sources, while *h* may denote results obtained from those sources. The fuzzy integral is then used to combine the outcomes nonlinearly.

Quality of ranking metrics

Two metrics for comparing rankings are presented: NDCG which derives from cumulative gain (CG) – metric often used for search engines; and MAP which originates in the precision metric for classification.

Discounted cumulative gain (DCG) is a measure of ranking quality. It is often normalized so that it is easier comparable, giving normalized discounted cumulative gain (NDCG) [31]. DCG is a refinement of a simpler metric (cumulative gain), which is a sum of the relevance values of all elements in the ranking:

$$CG_k = \sum_{i=1}^{k} rel_i \tag{16}$$

where: rel_i is the relevance score (e.g. the real algorithm quality) and k is the number of elements considered.

CG does not take into account the position in ranking, only the relevance. To include the position in ranking, DCG is used. There are two versions of this metric, namely standard:

$$rel_1 + \sum_{i=2}^{k} \frac{rel_i}{log_2(i+1)} \tag{17}$$

or with logarithm starting from the beginning:

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)} \tag{18}$$

The logarithm causes that the lower positions, e.g $i=5, 6, ...$, have less impact on the score. This means that DCG rewards ranking which puts relevant elements higher in order.

To enable comparison between different rankings and scenarios, DCG is normalized with respect to ideal ranking (IDCG) as follows:

$$NDCG_k = \frac{DCG_k}{IDCG_k} \tag{19}$$

where: $IDCG_k$ is a DCG value computed for an ideally sorted ranking. The value of $NDCG_k$ belongs to range [0, 1], and the value of 1 corresponds to an ideal ranking.

Another measure which can be used to assess the quality of ranking is mean average precision (MAP). MAP measures how well relevant elements are ranked, rewarding those ranked higher. Unlike NDCG, MAP treats the ranking task as a set of binary decisions: whether an element is relevant or not – and checks the precision with which all relevant elements are found in subsequent positions. Let us first define average precision

$$AP = \frac{1}{R} \sum_{i=1}^{k} P(i) \cdot rel_i \tag{20}$$

where: k is the number of elements in ranking, R is the number of actually relevant elements in ranking, $P(i)$ is the precision for position i, computed as the proportion of relevant items at i to the number i and $rel_i \in \{0,1\}$ is the relevance of the i-th position (1 for relevant or 0 for non-relevant).

Next, MAP is an average of AP for multiple rankings

$$\text{MAP} = \frac{1}{Q} \sum_{\text{q}=1}^{\text{Q}} AP_q \tag{21}$$

where Q is the number of queries or test cases (e.g. different datasets that algorithms were tested on) and AP_q is the average Precision for the q-th case.

In the context of evaluating methods for aggregating the rankings of algorithms (e.g., anomaly detection classifiers), MAP can be used to see how well a ranking method positions algorithms that are among the "relevant" (e.g., top-3 best). What matters is not only their presence in the ranking, but also their positions – the higher they are, the greater the contribution to the MAP score. MAP is a stricter metric than NDCG because it only rewards matches of relevant objects.

PROPOSED METHODOLOGY

In this section, an overview of our proposed ranking method is presented. Its computational complexity is discussed.

Aggregation algorithm

Let N algorithms be given (e.g., anomaly detection, as is the case in this research). For each of them, the values of n metrics (e.g., recall, accuracy) have been calculated. An aggregation method is chosen and, if necessary, appropriate weights must be assigned (e.g., for weighted average, OWA operator, for calculation of fuzzy measures for Choquet integral). Using the chosen method, aggregation of n metrics for each of the N algorithms is performed. In this way, N values of the aggregated metric are obtained. They can be compared between algorithms, assuming that they speak in the most broad way about the quality of models' performance. Based on the aggregated results, we rank the algorithms. Thus, the aggregated metric provided good interpretability and ease of comparison between algorithms. The process of ranking creation with the use of aggregation method is presented in Figure 1.

If several different aggregation methods were used for testing purposes, or if the real ranking order of the algorithms is known, metrics for evaluating the quality of ranking (such as NDCG or MAP) can be applied to the obtained rankings to compare them. Based on this, the weights of the metrics and the aggregation method most suitable for the considered issue could be selected.

Computational complexity

To consider computational complexity of our method, few components are analyzed. The computation of anomaly detection metrics, the complexity of PCA algorithm (if it is used for weight acquirement) and finally the chosen aggregation method.

The basic binary metrics (precision, recall, F1 score, accuracy, specificity) rely on the counting of TP, TN, FP, FN values, resulting in very low complexity of $O(n)$, where n is the number of observations in the dataset. The threshold-independent metrics: ROC AUC and PR AUC have higher complexity, because they require sorting of predictions and calculating the area under the respective curve. The complexity of such calculations equals $O(n\log n)$.

Computational complexity of PCA algorithm [32], which in some variants of our proposal is used to obtain aggregation weights, depends on the number n of observations in dataset and the number of features d. The value of this complexity is $0(\min(n^2d, nd^2))$.

Basic aggregations like average or weighted average have the complexity of $O(n)$. Both OWA operator and Choquet integral are characterized by $O(n\log n)$ complexity, because they require sorting of input data.

NUMERICAL EXPERIMENTS

This section includes the description of the dataset used for experiments, two methods of weight selection and two variants of metric aggregation: with and without analysis of their correlation and removal of redundant information. In the end, the rankings of the algorithms are obtained and compared, using NDCG and MAP metrics.

Dataset description

The experimental dataset contains 260 records corresponding to the application of 10 anomaly detection algorithms evaluated on 26 distinct

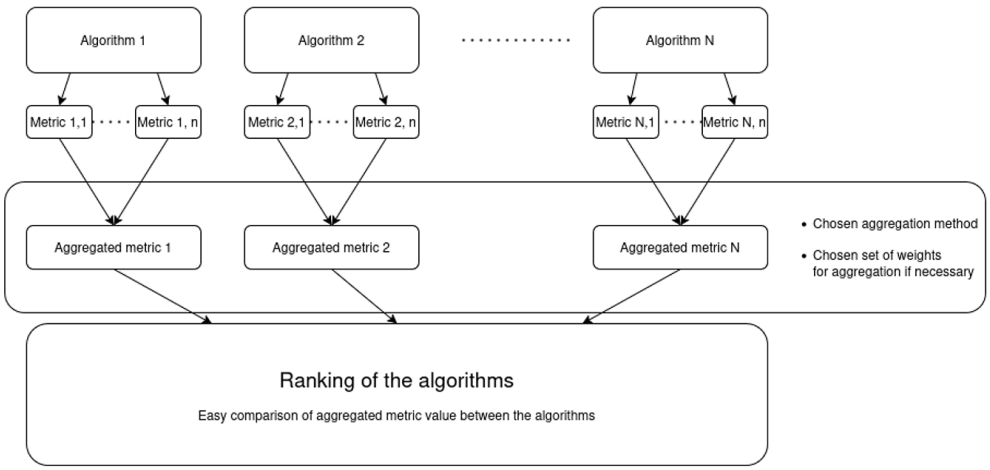

Figure 1. The process of algorithms ranking with aggregated metric

benchmark datasets. For metric computation, the following algorithms were tested: dimensionality reduction-based isolation forest (DRIF), fuzzy dimensionality reduction-based isolation forest (FDRIF), isolation forest (IF), *k*-nearest neighbors detector (KNN), local outlier factor (LOF), minimal spanning tree clustering for isolation forest (MSTCIF), minimal spanning tree-based isolation forest with evaluation function built on the fuzzy rules in Takagi-Sugeno model (MSTIF-TS), minimal spanning tree-based isolation forest (MSTIF), one-class support vector machine detector (OCSVM), principal component analysis outlier detector (PCA), while the datasets were: Annthyroid, Arrhytmia, Breast, Cardio, Forest Cover, Glass, Http, Ionosphere, Letter, Lympho, Mammography, Mnist, Musk, Optdigits, Pendigits, Pima, Satellite, Satimage-2, Shuttle, Smtp, Speech, Thyroid, Vertebral, Vowels, WBC, Wine – all are publicly available and widely recognized datasets for anomaly detection. For each dataset – algorithm combination, seven evaluation metrics were computed to evaluate the models' performance. The following metrics were selected: accuracy (overall effectiveness of algorithm), precision (correctness of positive predictions), recall (detection of positive cases), F1 score (balance of precision and recall), Specificity (recognition of negative cases), ROC AUC (ability to distinguish between classes regardless of threshold), and PR AUC (evaluation of effectiveness with unbalanced data). Since this article aims not to compare algorithms per se, but to present a universal method for ranking models, thenames of the algorithms have been disguised as Algorithm A, B, and so on.

To analyze general differences between the metric values, a boxplot is created, as shown on Figure 2. Among the evaluated metrics, precision generally yielded the lowest values. The F1 and PR AUC metrics followed, both characterized by rather similar distributions. Conversely, the metrics accuracy, ROC AUC, recall, and specificity generally exhibited higher scores, albeit with the presence of some lower-valued outliers in each case.The exact statistics values are presented in Table 1 and discussed below.

For Precision, the mean value is only 28.64%, with the median even lower. The F1 and PR AUC metrics had slightly higher means, just above 35%. In contrast, the remaining metrics demonstrated average values ranging between 74% and 78%. The standard deviations across all metrics ranged from 20% to 30%, indicating moderate variability. Minimum values are typically close to 0%, while maximum values are most often at or near 100%.

The correlation analysis of the evaluation metrics reveals several noteworthy patterns. As one can observe from correlation matrix visible on Figure 3, strong positive correlations areobserved between F1 score and precision, as well as between F1 score and PR AUC, indicating that these metrics convey highly similar information. A similarly high correlation is noted between accuracy and specificity, and to a slightly lesser extent between accuracy and ROC AUC, suggesting that accuracy is largely driven by correct classification of negative cases. Additionally, ROC AUC showed substantial correlation with specificity, while PR AUC is highly correlated with Precision. On the other hand, recall exhibited low or negligible correlations with most other metrics (e.g. -0.03 with accuracy, -0.1 with specificity), highlighting its distinct role in capturing the model's ability to detect positive (anomalous) instances. These findings suggest potential redundancy among several metrics, particularly within the groups F1–precision–PR AUC and accuracy–specificity. Therefore, when designing aggregation strategies or selecting key metrics for evaluation, it may be worth considering to retain only one representative from each highly correlated group to avoid overemphasizing overlapping aspects of model performance.

Aggregation of all metrics

Firstly, we present more automated variant of our methodology, where all available metrics are aggregated, regardless of their correlation. Two approaches to weight selection have been made to allow better comparison of aggregation methods. One way of selecting the weights is based on the relationships found in our dataset, and the other is expert selection based on knowledge of the metrics. The selected sets of weights were used for weighted average, OWA and smooth OWA operators, and they served as the g_i values for calculating the fuzzy measures in Choquet integral.

In the first weighting strategy, the columns with metrics' values are treated as input features for principal component analysis. The contribution of each metric to the first (and therefore most significant) of the principal components (which explains about 66% of the variance in the original

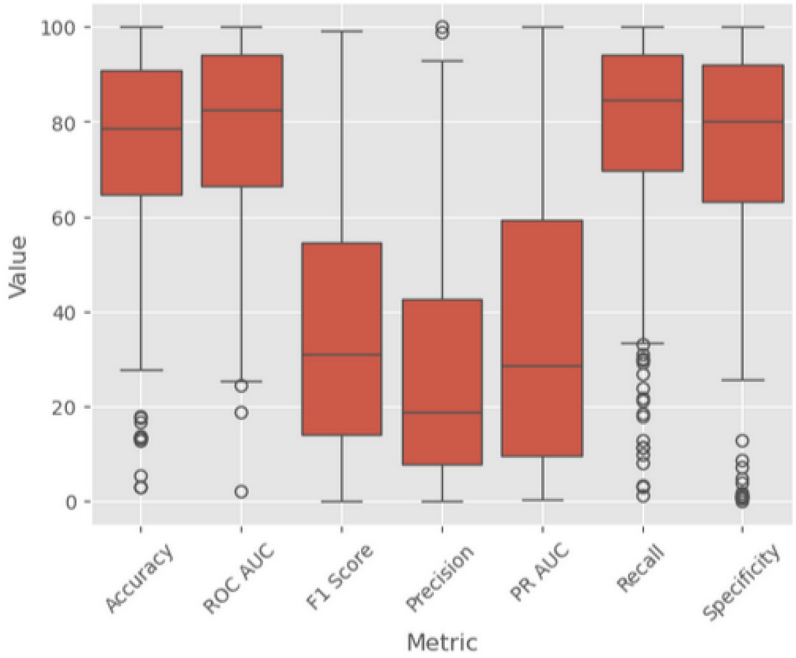

Figure 2. Boxplot of metric values

Table 1. The statistics for the metrics results for the analyzed algorithms and datasets

Metric	Min	Mean	Median	Max	Std
ROC AUC	2.2	77.99	82.40	100.0	20.06
Accuracy	3.1	74.59	78.70	100.0	21.39
F1 score	0.2	35.94	30.90	99.1	27.03
PR AUC	0.4	37.17	25.55	100.0	30.65
Precision	0.1	28.64	18.75	100.0	26.85
Recall	1.3	77.97	84.70	100.0	22.37
Specificity	0.0	74.25	80.00	100.0	23.47

data) is examined. In the Figure 4 the weights are shown after normalization to a sum of 1.

The largest weight, about 0.22, was assigned to the PR AUC metric. Slightly smaller weights (around 0.18-0.19) belong to the F1 score and precision metrics. This is followed by weights of about 0.11-0.12 for the specificity, accuracy and AUC metrics. The smallest weight of about 0.05 is assigned to the recall metric.

Based on expert knowledge, the metrics are assigned the weights visible on the barplot in Figure 5, with following explanations:

- Recall – crucial metric, because missed anomalies (false negatives) are oftentimes the most expensive.
- F1 score – a compromise between recall and precision
- Precision – reduces the number of false alerts
- Specificity – ensures that normal cases don't fall into the bag of anomalies.
- PR AUC – especially informative with strongly unbalanced classes (typical in anomalies)
- ROC AUC – describes ability to distinguish between classes, but in extremely unbalanced data is sometimes less informative than PR AUC.
- Accuracy – easy to interpret, but in anomaly detection often overestimated by the large number of normal class examples.

It is worth noting that expert weights can be modified in practical applications if, for example, we are primarily concerned with reducing false alarms, or with eliminating errors resulting from classifying anomalies as normal observations. The cost of such errors could be considered in practice. One can also run the ranking for

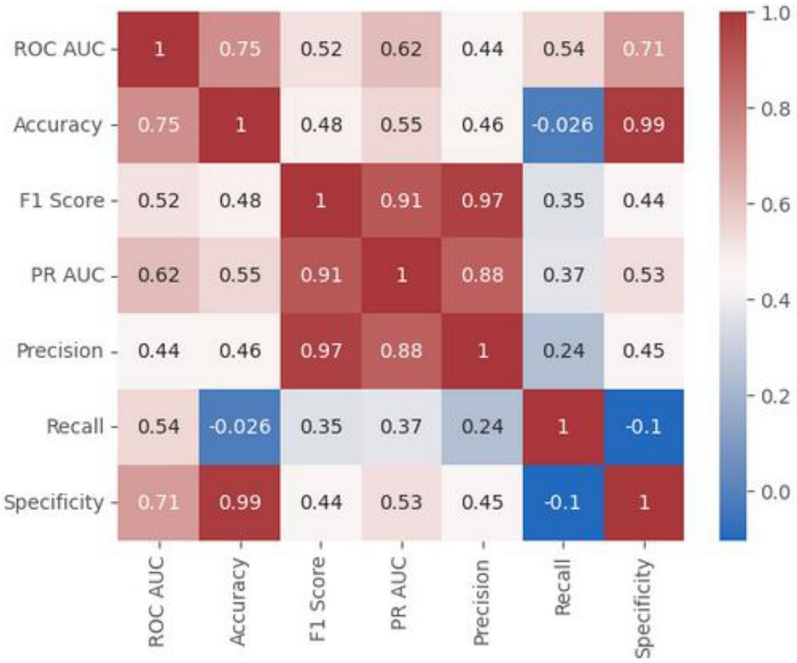

Figure 3. Correlation matrix of al metrics results for the analyzed algorithms and datasets

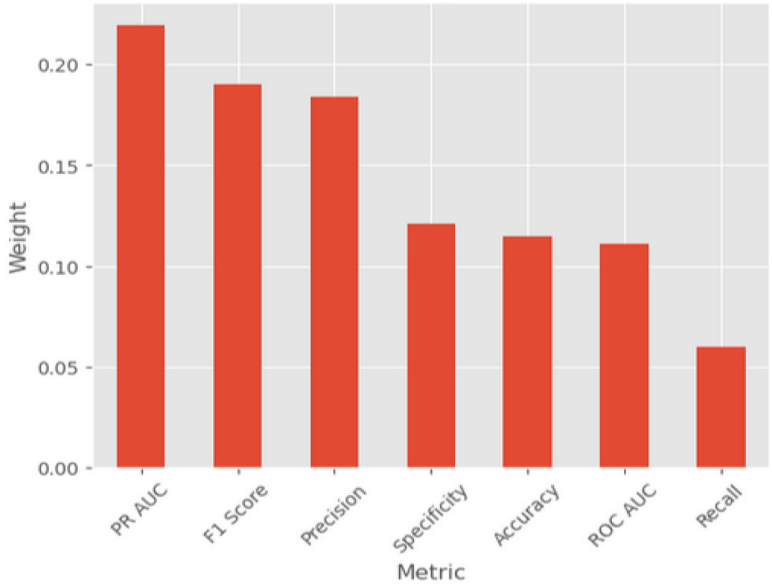

Figure 4. Weight selection for all metrics based on contribution to the first principal component

several sets of weights and see how the order of the models changes.

A comparison of the obtained aggregated metric's values was performed on the boxplot visible in Figure 6. Typically, the value of the metric fell within the range from 45 to 65%. For OWA aggregation, generally higher values of metrics are obtained with PCA weights than with expert weights. The opposite is true for weighted average and Choquet aggregation. Overall, the highest metrics' values were for OWA, OWA Simpson, OWA ONC3 (with PCA weights) with median value above 60, and generally the lowest for OWA trapezoidal and 3/8 with expert weights with median value above 50.

With the use of obtained aggregated metric, two sets of rankings are created. The ranking based on aggregation with PCA-derived weights is presented in Table 2, while the ranking obtained from aggregation with expert weights is in Table 3.

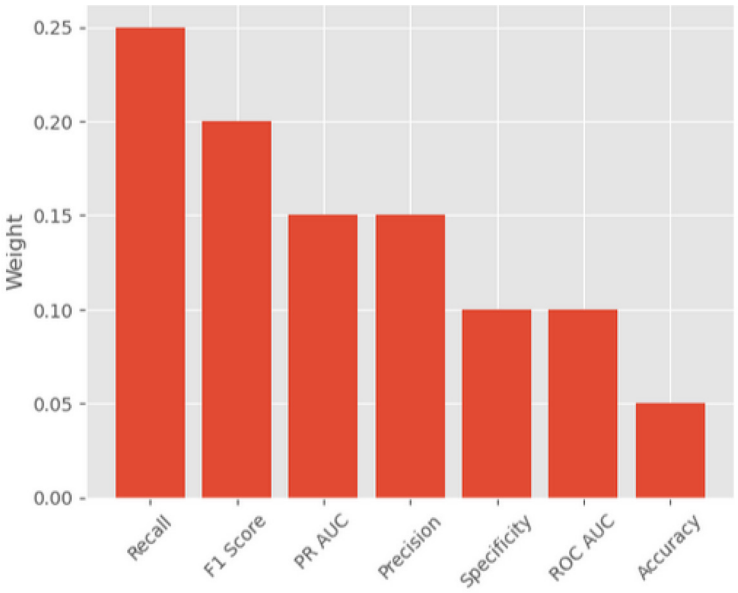

Figure 5. Weight selection for all metrics based on expert knowledge

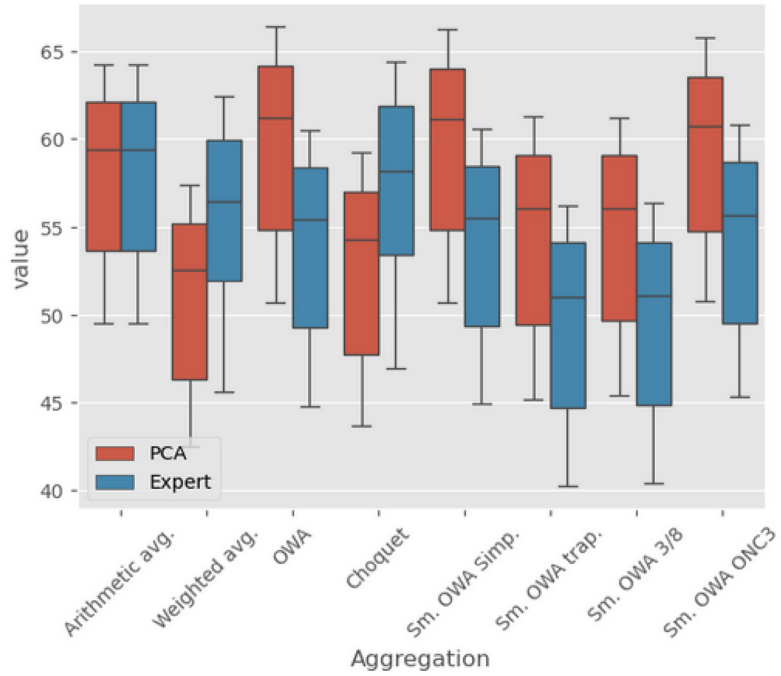

Figure 6. Comparison of the aggregated metric's value for different weights and aggregation methods; aggregation performed for all individual metrics

It can be seen that almost always the ratings do not differ between methods. Only the Algorithm G and Algorithm H swap the first two places when using the arithmetic mean. The same situation occurs when using PCA weights a for basic OWA aggregation.

To provide comparison for rankings based on aggregated metric, seven rankings originating from individual metrics' values are presented in Table 4. While some rankings seem similar to the aggregated (for example ROC AUC or PR AUC yields similar results), some differ for certain algorithms. Considering the accuracy or specificity metric, the Algorithm A is placed third, while other metrics rank it at usually fifth or sixth position. Algorithm D, generally placed near to bottom (seventh to ninth position), by accuracy or specificity is classified as second-best model. Algorithm I is also recognized as one of the worst, by all metrics except recall which placed in on fourth position. Accuracy and specificity produce similar rankings to each other, but differ from other

Table 2. Ranking obtained from aggregation of all metrics with PCA weights

Algorithm	Arithmetic average	Weighted average	OWA	Choquet	Smooth OWA (Simpson)	Smooth OWA (trapezoidal)	Smooth OWA (3/8)	Smooth OWA (ONC3)
Algorithm A	6	6	6	6	6	6	6	6
Algorithm B	9	9	9	9	9	9	9	9
Algorithm C	4	4	4	4	4	4	4	4
Algorithm D	7	7	7	7	7	7	7	7
Algorithm E	10	10	10	10	10	10	10	10
Algorithm F	3	3	3	3	3	3	3	3
Algorithm G	1	2	1	2	1	1	1	1
Algorithm H	2	1	2	1	2	2	2	2
Algorithm I	8	8	8	8	8	8	8	8
Algorithm J	5	5	5	5	5	5	5	5

Table 3. Ranking obtained from aggregation of all metrics with PCA weights

Algorithm	Arithmetic average	Weighted average	OWA	Choquet	Smooth OWA (Simpson)	Smooth OWA (trapezoidal)	Smooth OWA (3/8)	Smooth OWA (ONC3)
Algorithm A	6	6	6	6	6	6	6	6
Algorithm B	9	9	9	9	9	9	9	9
Algorithm C	4	4	4	4	4	4	4	4
Algorithm D	7	7	7	7	7	7	7	7
Algorithm E	10	10	10	10	10	10	10	10
Algorithm F	3	3	3	3	3	3	3	3
Algorithm G	1	2	2	2	2	2	2	2
Algorithm H	2	1	1	1	1	1	1	1
Algorithm I	8	8	8	8	8	8	8	8
Algorithm J	5	5	5	5	5	5	5	5

Table 4. Ranking obtained on the basis of individual metrics' values

Algorithm	ROC AUC	Accuracy	F1 Score	PR AUC	Precision	Recall	Specificity
Algorithm A	5	3	6	6	6	8	3
Algorithm B	8	10	9	8	9	7	9
Algorithm C	4	5	4	4	3	5	4
Algorithm D	7	2	8	7	7	9	2
Algorithm E	10	8	10	10	10	10	8
Algorithm F	3	6	3	3	5	2	7
Algorithm G	1	1	2	2	2	3	1
Algorithm H	2	4	1	1	1	1	5
Algorithm I	9	9	7	9	8	4	10
Algorithm J	6	7	5	5	4	6	6

metrics in many cases. Recall also provides some diversification in ranking. Some general conclusions can be drawn for most of the algorithms. However we can observe that different metrics yield inconsistent results.

To compare the proposed ranking method with the classic approach of individual metrics, NDCG and MAP values are calculated. They should not be interpreted as a quality measure per se, but rather as a consistency measure between the proposed ranking and rankings by individual metrics. NDCG and MAP metrics are presented in Table 5.

Table 5. Consistency of individual metrics' rankings with proposed ranking (Choquet aggregation with fuzzy measures based on expert weights)

Metric used for ranking	NDCG value	MAP value
ROC AUC	0.986	0.967
Accuracy	0.924	0.696
F1 Score	0.999	1.000
PR AUC	1.000	1.000
Precision	0.995	1.000
Recall	0.982	0.927
Specificity	0.918	0.696

The lowest quality is achieved by ranking on the basis of specificity and very similarly on the basis of accuracy. The NDCG measure was 91.8% and 92.4% for these rankings, respectively, relative to our rankings. For the other metrics, the NDCG measure is around 98–100% which suggests very high similarity to our proposed ranking method.

For the computation of MAP measure, the ranking is split in two parts. The k best models are labeled 1, and the rest 0. Here $k=5$ is chosen (meaning the ranking is divided in half, into better and worse models). Noticeably the weakest ranking consistency appears if one are to use specificity alone or accuracy alone, it is around 69.6%. Much higher is the score for ranking by the recall metric with 92.7% and AUC with 96.7%, and the rankings by F1 and PR AUC are in accordance with ours by the MAP metric. The lower MAP scores as compared to NDCG may suggest that, although the rankings obtained by specificity and accuracy might seem similar to other metrics, they have much differences as to where they placed the algorithms considered as the best by the aggregation metric.

Aggregation of less correlated metrics

Metrics selected for aggregation after discarding the most strongly correlated ones include:

- Recall – strongly independent, low correlation with precision and specificity. Very important in anomaly detection.
- Precision – although correlated with F1 and PR AUC, carries its own score – added value when evaluating false positives.
- ROC AUC – fairly low correlation with precision/recall/F1 – looks at global discriminating ability.
- PR AUC – similar to F1 score (and highly correlated with it), but PR AUC is better with unbalanced data.
- Specificity – virtually identical to accuracy, but usually more informative in anomaly detection.

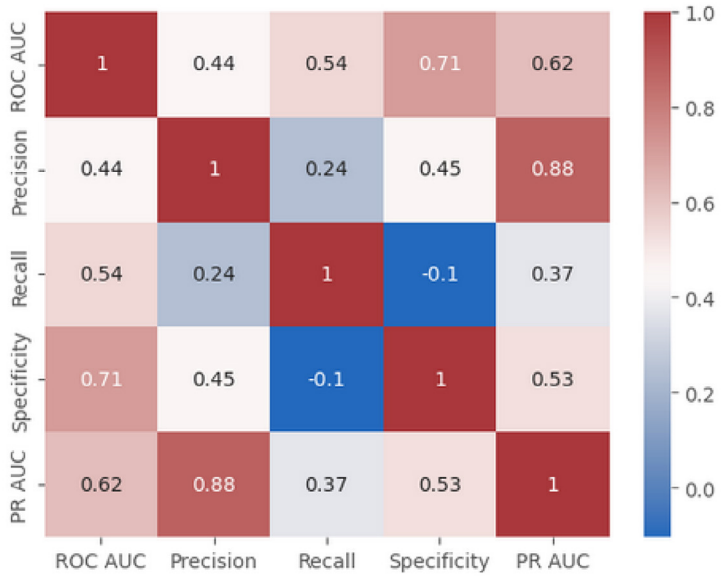

Figure 7. Correlation matrix after removing the most redundant metrics

The correlation after removal of other metrics is presented in Figure 7. All correlations higher than 0.9 are removed. The highest correlation now is 0.88 between precision and PR AUC.

The weights for aggregation of less correlated metrics are derived in similar manner as before for the full set – basing on PCA contribution and by expert knowledge. The PCA weights are presented in Figure 8. The highest weight is assigned to PR AUC metric. The next is precision, and then specificity and ROC AUC with very similar contributions. Recall holds the smallest value, similarly as is the case for PCA for all metrics.

The weights obtained by expert are presented in . Likewise as with analysis of all metrics, noticeably highest value is assigned to recall metric. Following it are PR AUC and precision, both ranked very similarly (and as we know from correlation matrix, they carry somewhat similar information). ROC AUC and specificity are graded least significant on similar level to each other.

The boxplot on Figure 10 presents the comparison of aggregated metric value when considering only less correlated metrics for aggregation. The range of values is broader this time than for

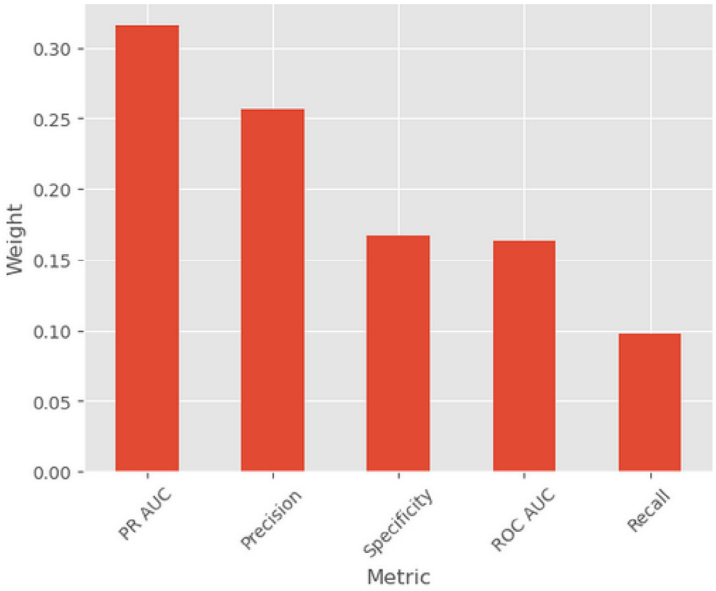

Figure 8. Weights for less correlated metrics based on PCA contribution

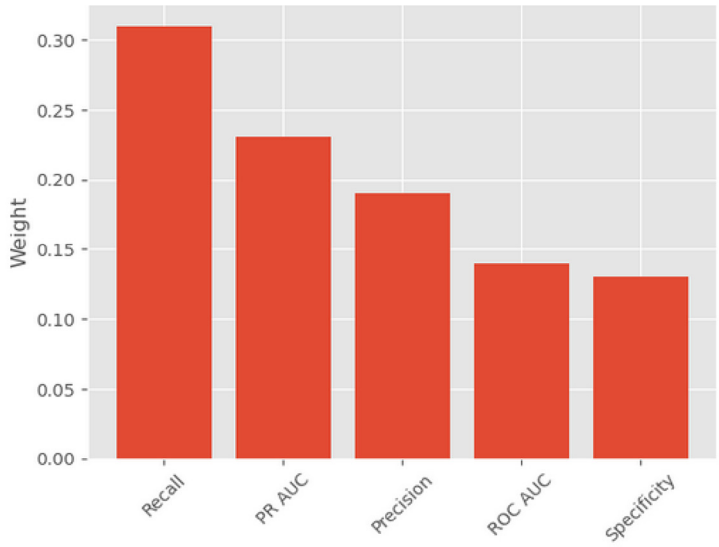

Figure 9. Weights for less correlated metrics based on expert knowledge

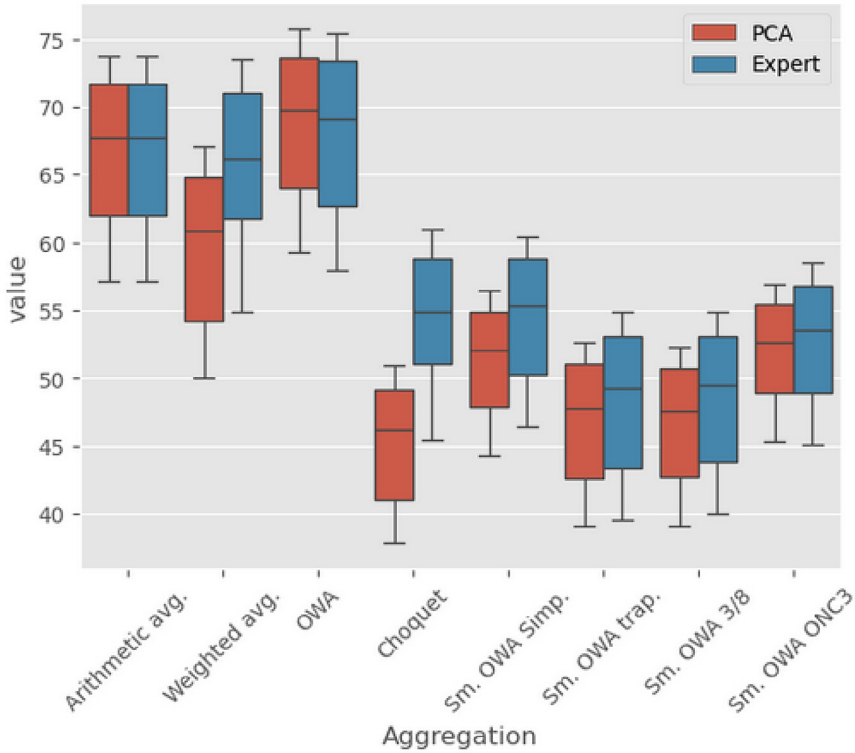

Figure 10. Comparison of the aggregated metric's value for different weights and aggregation methods; aggregation performed for all individual metrics

aggregation of all metrics. Smallest median value is achieved for Choquet method with fuzzy measures based on PCA weights, around 45. OWA with 3/8 and trapezoidal smoothing methods fell between 45 and 50 with the median value (regardless of weights used). Median values around 50–55 are achieved for OWA with 3/8 and Simpson smoothing (for both weight sets) and for Choquet with fuzzy measures based on expert weights. Significantly higher values are obtained from aggregation by arithmetic mean, weighted average (especially with expert weights) and basic OWA operator (with both sets of weights). The highest median is achieved for OWA with PCA weights, reaching almost 70.

Large differences in the absolute values of the metrics do not reflect in significant changes in ranking. Only the 8 to 9 and 1 to 2 positions differ sometimes for PCA weights (as seen in Table 6). For example, ranking for Choquet with fuzzy measures based on PCA weights differs in these positions from the rankings presented for aggregation of all metrics. Same situation occurs for weighted average with PCA weights. For expert weights (presented in Table 7) the ranking is less

Table 6. Ranking obtained from aggregation of less correlated metrics with PCA weights

Algorithm	Arithemtic average	Weighted average	OWA	Choquet	Smooth OWA (Simpson)	Smooth OWA (trapezoidal)	Smooth OWA (3/8)	Smooth OWA (ONC3)
Algorithm A	6	6	6	6	6	6	6	6
Algorithm B	9	8	9	8	9	9	9	9
Algorithm C	4	4	4	4	4	4	4	4
Algorithm D	7	7	7	7	7	7	7	7
Algorithm E	10	10	10	10	10	10	10	10
Algorithm F	3	3	3	3	3	3	3	3
Algorithm G	2	2	1	2	1	2	2	2
Algorithm H	1	1	2	1	2	1	1	1
Algorithm I	8	9	8	9	8	8	8	8
Algorithm J	5	5	5	5	5	5	5	5

Table 7. Ranking obtained from aggregation of less correlated metrics with expert weights

Algorithm	Arithemtic average	Weighted average	OWA	Choquet	Smooth OWA (Simpson)	Smooth OWA (Trapezoidal)	Smooth OWA (3/8)	Smooth OWA (ONC3)
Algorithm A	6	6	6	6	6	6	6	6
Algorithm B	9	9	9	9	9	9	9	9
Algorithm C	4	4	4	4	4	4	4	4
Algorithm D	7	7	7	7	7	7	7	7
Algorithm E	10	10	10	10	10	10	10	10
Algorithm F	3	3	3	3	3	3	3	3
Algorithm G	2	2	1	2	1	2	2	2
Algorithm H	1	1	2	1	2	2	2	2
Algorithm I	8	8	8	8	8	8	8	8
Algorithm J	5	5	5	5	5	5	5	5

dependent on aggregation used. As opposed to PCA weights, the positions 8 and 9 never changed. Only on the first two positions changes occurred between Algorithms G and H. OWA and smooth OWA with Simpson smoothing favored Algorithm G over H, while all other methods gave opposite result. No other changes are observed.

DISCUSSION

In addition to standard performance metrics, other characteristics of algorithms – such as computational complexity or execution time – can also be included in comparative evaluations. This approach was presented in [33], where metrics such as F-Score, Training Time, Testing Time, and Consistency were considered. Such an extension could be easily integrated into our framework by recording the mentioned training and testing parameters for the evaluated algorithms and including them among the aggregated features, with defining their respective importance (and thus their weight) in the evaluation process. Furthermore, the cited work applies an interesting multi-metric decision-making method based on AHP. While this is an advanced and well-established approach, it relies heavily on expert input for algorithm evaluation. In contrast, our research aims to increase the level of automation in the evaluation process.

A related study [34] also considers training time alongside accuracy and uses a k-NN-based meta-model to generate rankings. The ranking system proposed in that work offers a complementary perspective to ours: while our ranking is derived from model performance across multiple datasets, the cited method focuses on generating a ranking for a specific test dataset by identifying the most similar datasets among previously evaluated, usingthe k-NNframework and building the ranking accordingly. In contrast, our proposed method is primarily designed for use in scientific evaluation settings, where algorithms are assessed based on experiments conducted across a diverse set of datasets.

To compare rankings obtained in our study, we employed metrics such as NDCG, which is commonly used in research for evaluating ranking quality [35]. Other well-known approaches include statistical tests, such as the Friedman test and post-hoc tests to determine whether rankings differ significantly[36].For instance, the Nemenyi test has been applied to sports rankings [37], but could easily be adapted to other ranking domains. In addition, Spearman's rank correlation is also used for ranking comparisons [34]. However, a recurring challenge in such comparisons is the lack of an ideal or reference ranking, which often necessitates the manual creation of a "ground truth" ranking [38]. Our study also encounters this issue, which is why for now we have limited our comparisons to rankings derived from individual metrics. Designing a more reliable evaluation method for rankings remains an open challenge for future work.

CONCLUSIONS

This paper presents an interesting and novel methodology for aggregating multiple metrics used to assess anomaly detection algorithms. The proposed aggregated metric is designed to simplify interpretation of the results and ranking

of the best models. Several variants of the method are presented, including a more automatic approach involving aggregation of all metrics, as well as a correlation-aware approach that enables the exclusion of redundant information. Two weighting strategies for aggregation are discussed: one based on the contribution of metrics to the PCA components, and another derived from expert knowledge regarding the metrics' significance.

The algorithm rankings obtained via different metric aggregation strategies showed minimal variation, despite the fact that the aggregated metric values differed significantly (with median values ranging from approximately 45% to 70%). For the ten ranked algorithms, differences typically occurred in the ordering of the top two algorithms, with occasional changes in the placement of algorithms ranked 8th and 9th. When comparing the rankings derived from the proposed aggregated metric to those based on individual metrics, the aggregated results most closely resembled the rankings obtained from F1-score, PR AUC, and Precision. Slightly larger, but still moderate, differences are observed when using ROC AUC or Recall alone. In contrast, substantial discrepancies, especially at the top of the rankings, are found when using Specificity or Accuracy as sole metrics. This aligns with well-known concerns in anomaly detection that Accuracy may be a misleading metric, often inflated due to class imbalance.

The proposed aggregated metric framework addresses a longstanding challenge in model evaluation: the simultaneous interpretation of multiple, often conflicting performance indicators. By introducing a principled aggregation and ranking mechanism, our approach enables the comparison of models using a single metric that should reflecta balanced, task-aware synthesis of traditional measures. The novelty lies not only in the integration of diverse metrics into one scale, but in the flexible design space that allows tailoring the metric to domain-specific priorities, such as recall-dominant sensitivity or robustness to class imbalance. The flexibility is achieved due to choice possibilities of aggregation methods and weighting strategies. The system contributes to a more interpretable, standardized, and scalable evaluation process, offering practical utility in benchmarking, model selection, and automated reporting.

Future work may involve the development of methods to evaluate the quality of rankings obtained through metric aggregation. The proposed methodology could potentially be extended beyond anomaly detection to tasks such as classification or regression, where interpreting multiple performance metrics is also a common challenge. An intriguing direction for future research would be to use the aggregated metric as an objective function in neural networks or as a guiding criterion in cross-validation-based fine-tuning.

REFERENCES

1. Yager RR. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Trans Syst Man Cybern. 1988;18(1):183–190. 10.1109/21.87068
2. Murofushi T, Sugeno M. An interpretation of fuzzy measures and the Choquet integral as an integral with respect to a fuzzy measure. Fuzzy Sets Syst. 1989; 29(2):201–227. 10.1016/0165-0114(89)90194-2
3. Alves TL, Correia JP, Visser J. Benchmark-Based Aggregation of Metrics to Ratings. In: Joint Conference of the 21st International Workshop on Software Measurement and the 6th International Conference on Software Process and Product Measurement. IEEE; 2011. 20–29. 10.1109/IWSM-MENSURA.2011.15
4. Mordal K, Anquetil N, Laval J, Serebrenik A, Vasilescu B, Ducasse S. Software quality metrics aggregation in industry. J Softw Evol Process. 2013 Oct 13; 25(10):1117–1135. 10.1002/smr.1558
5. BenHassine-Guetari S, Darmont J, Chauchat JH. Aggregation of data quality metrics using the Choquet integral. In: 8th International Workshop on Quality in Databases (VLDB/QDB 10). Singapore; 2010.
6. Redondo AR, Navarro J, Fernández RR, de Diego IM, Moguerza JM, Fernández-Muñoz JJ. Unified performance measure for binary classification problems. Lecture Notes in Computer Science, 2020. 12490, 104–112. 10.1007/978-3-030-62365-4_10
7. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015 Mar 4;10(3):e0118432. 10.1371/journal.pone.0118432
8. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning – ICML '06. New York, New York, USA: ACM Press; 2006. 233–240. 10.1145/1143844.1143874
9. Lipton ZC, Elkan C, Narayanaswamy B. Thresholding Classifiers to Maximize F1 Score. 2014 Feb 8;
10. Chicco D, Jurman G. The advantages of the

Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics; 2020 Dec 2. 21(1):6. 10.1186/s12864-019-6413-7

11. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn; 2009 Oct 16; 77(1):103–23. 10.1007/s10994-009-5119-5
12. Yang Y, Liu X. A re-examination of text categorization methods. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA: ACM; 1999, 42–9. 10.1145/312624.312647
13. Mccallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. In: Proceedings of the AAAI-98 Workshop on Learning for Text Categorization; 1998.
14. Naidu G, Zuva T, Sibanda EM. A review of evaluation metrics in machine learning algorithms. Lecture Notes in Networks and Systems, 2023, 724, 15–25. 10.1007/978-3-031-35314-7_2
15. Sørbø S, Ruocco M. Navigating the Metric Maze: A Taxonomy of Evaluation Metrics for Anomaly Detection in Time Series; 2023 Mar 2;
16. Huet A, Navarro JM, Rossi D. Local evaluation of time series anomaly detection algorithms. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2022, 635–645. 10.1145/3534678.3539339
17. Kim GY, Lim SM, Euom IC. A study on performance metrics for anomaly detection based on industrial control system operation data. Electronics; 2022 Apr 12;11(8):1213. 10.3390/electronics11081213
18. Tatbul N, Lee TJ, Zdonik S, Alam M, Gottschlich J. Precision and Recall for Time Series; 2018 Mar 8;
19. Najgebauer P, Scherer R, Grycuk R, Walczak J, Wojciechowski A, Łada-Tondyra E. Fast visual imperfection detection when real negative examples are unavailable. Lecture Notes in Computer Science, 2023, 14126, 58–68. 10.1007/978-3-031-42508-0_6
20. Saad A, Ullah Sheikh U, Moslim MS. Developing convolutional neural network for recognition of bone fractures in X-ray images. Adv Sci Technol Res J; 2024 Aug 1, 18(4): 228–237. 10.12913/22998624/188656
21. Obrzut B, Kusy M, Semczuk A, Obrzut M, Kluska J. Prediction of 5–year overall survival in cervical cancer patients treated with radical hysterectomy using computational intelligence methods. BMC Cancer; 2017 Dec 12. 17(1):840. 10.1186/s12885-017-3806-3
22. Otoom M, Abdul Sattar K, Al Sadig M. Ensemble model for network intrusion detection system based on bagging using J48. Adv Sci Technol Res J; 2023 Apr 1. 17(2):322–9. 10.12913/22998624/161820
23. Bereta M. Negative selection algorithm for unsupervised anomaly detection. Appl Sci; 2024 Nov 27. 14(23):11040. 10.3390/app142311040
24. Sezgin A, Boyacı A. Enhancing intrusion detection in industrial internet of things through automated preprocessing. Adv Sci Technol Res J; 2023 Apr 1. 17(2):120–35. 10.12913/22998624/162004
25. Almutairi Y, Alhazmi B, Munshi A. Network intrusion detection using machine learning techniques. Adv Sci Technol Res J.; 2022 Jul 1. 16(3):193–206. 10.12913/22998624/149934
26. Rachwał A, Karczmarek P, Rachwał A. Smooth ordered weighted averaging operators. Inf Sci (Ny); 2025 Jan. 686:121343. 10.1016/j.ins.2024.121343
27. Karczmarek P, Gregosiewicz A, Łagodowski ZA, Dolecki M, Gałka Ł, Powroźnik P, et al. Smooth and enhanced smooth quadrature-inspired generalized choquet integral. Available at SSRN 4543000, 2023. 10.2139/ssrn.4543000
28. Beliakov G, Pradera A, Calvo T. Aggregation Functions: A Guide for Practitioners. Vol. 221. Berlin: Springer; 2007.
29. Grabisch M. Fuzzy integral in multicriteria decision making. Fuzzy Sets Syst; 1995 Feb. 69(3):279–98. 10.1016/0165-0114(94)00174-6
30. Pedrycz W, Gomide F. An introduction to fuzzy sets: analysis and design. MIT Press; 1998.
31. Wang Y, Wang L, Li Y, He D, Liu TY, Chen W. A theoretical analysis of NDCG type ranking measures. In: Proceedings of the 26th Annual Conference on Learning Theory, 2013 Apr 24, 25-54.
32. Maćkiewicz A, Ratajczak W. Principal components analysis (PCA). Comput Geosci. 1993 Mar. 19(3):303–342. 10.1016/0098-3004(93)90090-R
33. Ali R, Lee S, Chung TC. Accurate multi-criteria decision making methodology for recommending machine learning algorithm. Expert Syst Appl; 2017 Apr. 71:257–278. 10.1016/j.eswa.2016.11.034
34. Brazdil PB, Soares C, da Costa JP. Ranking Learning Algorithms: Using IBL and Meta-Learning on Accuracy and Time Results. Mach Learn; 2003 Mar. 50(3):251–277. 10.1023/A:1021713901879
35. Sun Q, Pfahringer B. Pairwise meta-rules for better meta-learning-based algorithm ranking. Mach Learn; 2013; 93(1): 141–161. 10.1007/s10994-013-5387-y
36. Brazdil PB, Soares C. A comparison of ranking methods for classification algorithm selection. Lecture Notes in Computer Science, 2003, 1810, 63–75. 10.1007/3-540-45164-1_8
37. Barrow D, Drayer I, Elliott P, Gaut G, Osting B. Ranking rankings: an empirical comparison of the predictive power of sports ranking methods. J Quant Anal Sport; 2013 Jan 1. 9(2), 10.1515/jqas-2013-0013
38. Héberger K. Sum of ranking differences compares methods or models fairly. TrAC Trends Anal Chem; 2010 Jan. 29(1):101–109. 10.1016/j.trac.2009.09.00