Evaluating Anomaly Detection Algorithms: A Multi-Metric Analysis Across Variable Class Imbalances

1st Mohammad Sahadat Hossain
TU Dortmund University
Dortmund, Germany
mohammadsahadat.hossain@tu-dortmund.de

2nd Mohammad Sakhawat Hossain
TU Dortmund University
Dortmund, Germany
sakhawat.hossain@tu-dortmund.de

3rd Simon Klüttermann
TU Dortmund University
Dortmund, Germany
Simon.Kluettermann@cs.tu-dortmund.de

4th Emmanuel Müller
TU Dortmund University
Dortmund, Germany
emmanuel.mueller@cs.tu-dortmund.de

***Abstract*—Anomaly detection is critical for ensuring integrity and performance in numerous high-stakes domains, ranging from financial systems to network security. The effectiveness of anomaly detection algorithms is significantly affected by the class imbalances that characterize real-world data. This research conducts an in-depth analysis of five anomaly detection algorithms—Angle-based Outlier Detector (ABOD), K-Nearest Neighbors (KNN), Local Outlier Factor (LOF), Isolation Forest (IsoForest), and One-Class SVM (OCSVM). Our evaluation spans datasets with a spectrum of feature complexities and observation volumes, alongside a targeted resampling of anomaly percentages, shifting from a balanced 50/50 distribution to imbalances ranging from 10% to 40%. A multi-metric evaluation framework is deployed, encompassing F1, ROC AUC, PR AUC, MCC, and Kappa, to deliver a layered assessment of algorithmic performance. Our findings reveal distinct stability in the correlations of F1 with Kappa and MCC, and between MCC and Kappa, signifying their potential as consistent performance indicators across various datasets and models. In contrast, F1's correlation with ROC and PR AUC, and to a lesser degree PR AUC's correlation with ROC, displayed notable fluctuations, indicating a differential impact of class distribution on these metrics. The study underscores the imperative of utilizing a multi-metric approach for a comprehensive evaluation of anomaly detection algorithms, ensuring adaptability to the diverse and skewed distributions encountered in practice. The insights from this analysis provide a pathway for practitioners to make informed decisions in selecting and deploying anomaly detection models that can withstand the challenges posed by varying class imbalances.**

***Index Terms*—Anomaly Detection, Class Imbalance, Evaluation Metrics, Machine Learning Algorithms, and Performance Correlation Analysis**

I. Introduction

Anomaly detection, the process of identifying unusual patterns that do not conform to expected behavior, is vital across a wide array of sectors, including finance, cybersecurity, healthcare, and industrial operations. This field grapples with the inherent complexities posed by class imbalances, the sporadic nature of anomalies, and the varied manifestation of data. Traditional classifiers often underperform in these contexts due to the disproportionate representation of classes, prompting the need for specialized anomaly detection techniques. However, the efficacy of these specialized algorithms across different conditions remains a subject ripe for investigation.

This paper rigorously evaluates five advanced anomaly detection algorithms—Angle-based Outlier Detector (ABOD), K-Nearest Neighbors (KNN), Local Outlier Factor (LOF), Isolation Forest (IsoForest), and One-Class SVM (OCSVM)—across a multitude of dataset characteristics. These characteristics include the number of features, the volume of observations, and an array of anomaly proportions that have been methodically resampled from a balanced distribution to ratios that are indicative of more typical, imbalanced environments.

In recognition of the shortcomings inherent in single-metric evaluations, our study utilizes a multi-metric approach. We incorporate F1, ROC AUC, PR AUC, MCC, and Kappa scores to provide a holistic view of performance. Each metric sheds light from a distinct angle: the F1 score harmonizes precision with recall, ROC AUC offers a measure independent of classification thresholds, PR AUC accentuates the positive class performance, and both MCC and Kappa provide balanced assessments that consider true negatives and adjust for class imbalance.

Our examination proceeds in a bifurcated manner: we initially consider algorithm performance across datasets with an equal anomaly-to-normal ratio (50/50). While such a balance is rare in practical scenarios, it establishes a baseline for comparative analysis. Subsequently, we investigate algorithm performance across datasets with resampled anomaly ratios of 10%, 20%, 30%, and 40%, thereby emulating more conventional operating conditions. Through the application of Spearman correlation coefficients, we discern the interrelations between various performance metrics and their evolution in response to shifting data properties.

A notable innovation of our research is the incorporation of variable anomaly proportions into our testing paradigm,

mirroring the complexity found in real-world datasets and evaluating the resilience of algorithms under these heterogeneous conditions. Our findings are intended to steer practitioners toward informed decision-making in the selection and deployment of anomaly detection algorithms, ensuring that the chosen models are robust not just in theory but under the practical exigencies presented by real-world data.

Through this endeavor, we contribute to the broader discourse on the practical application and assessment of anomaly detection algorithms, with the ultimate goal of enhancing their reliability and precision when operationalized in the real world.

II. Related Work

The landscape of anomaly detection has been shaped by the need to adapt traditional data analysis techniques to the challenges posed by outliers and rare events. Early methods were primarily statistical, relying on assumptions about data distribution that do not hold in many real-world scenarios. With the advent of machine learning, a plethora of algorithms have been introduced, each with their own strengths and limitations when dealing with high-dimensional, imbalanced data sets.

Angle-based Outlier Detection (ABOD) [1] exemplifies geometric approaches to anomaly detection, which consider the angular variance between points, proving effective in identifying outliers in high-dimensional spaces. K-Nearest Neighbors (KNN) is a distance-based method that classifies data points based on the distance to the nearest neighbors [2] and has been widely used due to its simplicity and effectiveness. Local Outlier Factor (LOF) [3] extends this idea by comparing local densities, allowing it to find anomalies that may not be outliers globally but are locally distinct.

Different approaches include the Isolation Forest (IsoForest) algorithm [4], which employs a forest of random trees to isolate anomalies, demonstrating high efficiency and scalability. Finally, one-Class SVM (OCSVM) [5] represents a kernel-based method tailored to anomaly detection, which constructs a decision boundary around the normal data instances, aiming to exclude outliers.

In terms of evaluation metrics, research has emphasized the inadequacy of conventional metrics like accuracy, especially in the context of imbalanced data sets [6]. The F1 score, ROC AUC, and PR AUC have been adopted as more appropriate measures, with the latter two providing more nuanced evaluations by considering model performance across different thresholds. However, these metrics have limitations; the F1 score can be overly optimistic in the presence of class imbalance, and ROC AUC can be misleading when the positive class is much smaller than the negative class.

Matthews Correlation Coefficient (MCC) and Cohen's Kappa [7] have been proposed as more balanced alternatives, accounting for true negatives and offering a correction for chance agreement. These metrics are particularly valuable in scenarios with varying class distributions, as they provide a more reliable indication of model performance.

The body of work on Re-sampling strategies to address class imbalance has also grown, with techniques such as SMOTE [8] and its variants being widely discussed. However, the impact of Re-sampling on the evaluation of anomaly detection algorithms has not been thoroughly explored, particularly in relation to the variation of anomaly percentages.

Our study builds upon this extensive body of work, critically analyzing the performance of a range of anomaly detection algorithms across different levels of class imbalance. We contribute to the discourse by not only employing a variety of performance metrics to evaluate the algorithms but also by examining how the correlation between these metrics evolves as we alter the anomaly percentages, offering insights that are closely aligned with the complexities of real-world data.

III. Methodology

A. Data Description

Our study encompasses a collection of data sets, each varying in size, and feature complexity, initially balanced with a 50/50 split between anomalies and normal observations. The data sets originate from diverse domains, ensuring a breadth of features ranging from low-dimensional (e.g., 6 features) to high-dimensional (e.g., over 100 features), with sample sizes extending from hundreds to tens of thousands. For each data set, we define three levels: the training set (X_train), the test set (X_test), and the ground truth labels for the test set (y_test).

data set	No. of Observation	No. of Features
Aloi	49534	27
Campaign	41188	62
Cardio	1831	21
Letter	1600	32
Mammography	11183	6
Mnist	7603	100
Musk	3062	166
Optdigits	5216	64
Pendigits	6870	16
Satimage-2	5803	36
Shuttle	49097	9
Stamps	340	9
Thyroid	3772	6
Vowels	1456	12
Waveform	3443	21
Wbc	223	9

B. Re-sampling Technique

Re-sampling is carefully planned to keep the total number of test samples. This way, any changes seen in how well anomaly detection methods work can only be due to the changed anomaly rates and not to changes in the size of the train set. This careful management of factors is needed to compare algorithmic performance across a range of anomaly occurrence rates in a fair way. Ten percent, twenty percent, thirty percent, and forty percent of anomalies were picked to represent a range of situations, from very rare to fairly

common unusual events. The study can get a better idea of how well each method works as the number of anomalies goes down by looking at a range of anomaly rates.

Algorithm 1 Adjusting Anomaly Ratios in Test Sets

$X_{train} \leftarrow$ training data matrix
$X_{test} \leftarrow$ test data matrix
$y_{test} \leftarrow$ test data labels
$anomaly\_percentages \leftarrow [0.1, 0.2, 0.3, 0.4]$
function REMOVENAN(X)
    return X_{noNaN}
end function
function ADJUSTTEST-SET($X_{train}, X_{test}, y_{test}, anomaly\_percentage$)
    $normal\_data \leftarrow X_{train}$
    $anomaly\_data \leftarrow$ filter X_{test} by $y_{test} == 1$
    $num\_anomalies \leftarrow$ length($anomaly\_data$)
    $num\_normals \leftarrow$ calculate normal's from $num\_anomalies$ and $anomaly\_percentage$
    Combine and anomaly_data to X_test_adjusted
    Label $X_{test\_adjusted}$ to create $y_{test\_adjusted}$
    return $X_{train\_adjusted}, X_{test\_adjusted}, y_{test\_adjusted}$
end function

C. Algorithm Implementation

- KNN Anomaly Detection [9]:

  The k-nearest-neighbor (kNN) method is a commonly utilized distance-based technique in the field of anomaly detection. Under normal conditions, this method is uncomplicated and efficacious in identifying worldwide anomalies. The objective is to determine the k-nearest neighbors of each data point contained in a given data set. The anomaly score is calculated based on the adjacent data points. In order to determine the anomaly score, the maximum distance to the next k neighbors is computed.
- Local Outlier Factor (LOF) [10]:

  The LOF, or Local Outlier Factor, is a mechanism for detecting anomalies based on distance. It is utilized to identify local irregularities by analyzing the local densities. This approach involves identifying the k-nearest neighbors for every individual data point inside a provided streaming data set. The local density of each data point is calculated by estimating the local reachability density (LRD) using the k-nearest-neighbors method. The anomaly score is determined by comparing a data point's Local Reachability Density (LRD) with the LRDs of its k nearest neighbors.
- Isolation Forest [11]:

  This anomaly detection approach relies on the principle of 'isolation' instead of the commonly utilized distance and density metrics. In this methodology, the anomalies are segregated or separated from the usual cases. The data instances that are both rare and have attribute values that significantly differ from the majority of the data instances are more likely to be isolated. This approach uses an isolation tree, a binary tree structure, to separate instances of interest.
- One-class SVM (OCSVM) [12]:

  Various semi-supervised and unsupervised versions of One-class support vector machine (OCSVM)–based anomaly detection have been documented in the literature. The fundamental concept of this strategy, which is based on machine learning, is to acquire knowledge about a decision boundary that attains the utmost distinction between the points and the origin. In the absence of labels, OCSVM is typically susceptible to outliers. To fix this problem, Amer et al. (2013) [13] added two changes that made OCSVM better at finding anomalies without being supervised. These improvements reduce the impact of outliers on the decision boundary in comparison to normal cases. Hu et al. (2018) [14] introduced a technique for identifying aberrant sub-sequences inside the provided time-series data using an anomaly detection approach.
- Angle-based Outlier Detection (ABOD) [1]:

  The multi-variate feature space angle that a set of three data points creates is what this anomaly detection method looks at. The variation in the angular enclosure's magnitude differs between outliers and normal spots. Typically, the variance is greater for the inlier points compared to the outliers. Therefore, this metric allows us to distinguish and group normal and outlier points separately.

D. Evaluation Metrics

Five evaluation metrics are used to measure the performance of the selected anomaly detection models, i.e. F1 score, ROC AUC, PR AUC, MCC, and Cohen's Kappa based on the Confusion matrix. These are defined as follows:

$$F1\ Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{1}$$

$$AUC = \int_0^1 \frac{TP}{TP + FN}\, d\frac{FP}{TN + FP} \tag{2}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3}$$

$$\kappa = \frac{2 \cdot (TP \cdot TN - FN \cdot FP)}{(TP + FP) \cdot (FP + TN) + (TP + FN) \cdot (FN + TN)} \tag{4}$$

Accuracy is defined as the proportion of correctly classified instances out of the total number of instances. The F1-score quantifies the correctness of a test. The Area Under the ROC Curve (AUC) quantifies the extent of the region beneath the ROC curve, which represents the balance between the True Positive Rate (TPR) and False Positive Rate (FPR) at varying probability thresholds. A machine learning model is considered superior if it exhibits greater levels of accuracy, precision, recall, F1-score, and/or AUC.

The MCC, or Matthews correlation coefficient, is a measure of the Pearson correlation coefficient applied to a confusion matrix. Cohen's Kappa is computed using the confusion matrix. The value for Kappa might be negative. However, Kappa considers the imbalance in class distribution rather than focusing just on total accuracy. High scores on MCC and Cohen Kappa are achieved only when the prediction yields favorable results for all four parameters of the confusion matrix (true positives, false negatives, true negatives, and false positives), relative to the proportions of positive and negative elements in the data set.

Conversely, True Positive (TP) is the proportion of samples that are accurately categorized as anomalies. False Negative (FN) is the proportion of anomaly samples that are mistakenly classed as normal data. False Positive (FP) is the proportion of the normal sample that is incorrectly categorized as an abnormality. True Negative (TN) is the proportion of anomaly samples that are correctly identified as normal data.

The Spearman rank correlation is a non-parametric statistical measure that evaluates the strength and direction of the association between two variables by examining their ranks and assessing how well they can be characterized using a monotonic function. Unlike Pearson's correlation, this method does not assume that the data follows a normal distribution. Spearman's technique is especially valuable when dealing with ordinal or non-normally distributed data, or where the focus is on evaluating the monotonicity of the connection rather than its linearity []. In our case, we choose our metrics score as variable.

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

d_i is the difference between the two metrics' ranks.

IV. Results

We first calculate the metric value for each model when the number of abnormal samples that are swapped to the normal samples is varied from 10% to 40%. Table I, II and III show the comparison of models based on different evaluation metrics on 17 data sets.

The line plot in Fig.1 illustrates the fluctuation of several performance measures based on the size of the data set before re-sampling. Each metric's performance seems to vary according to the number of samples, with ROC AUC and PR AUC demonstrating greater consistency across different data set sizes. The fluctuation of metrics such as Matthews Correlation Coefficient (MCC) and Cohen's Kappa indicates their susceptibility to changes in data set size. The graphic highlights the importance of taking into account the data set size when assessing model performance metrics. This is evident from the significant variations in metric values across different numbers of observations.

Fig.2 and Fig. 3 reflect correlation pairs of each metric's performance before and after re-sampling. Before re-sampling, we see that F1 score vs MCC, F1 score vs Kappa, ROC AUC vs PR AUC, and MCC vs Kappa are more consistent while others fluctuate as the number of observations increases. After re-sampling, the same pair of metrics suggest the same consistency. This could reflect the sensitivity of different metrics to data volume and its underlying distribution.

TABLE I: F1 Scores across the data sets

Model	ABOD	IF	KNN	LOF	OCSVM
ALOI	0.596432	0.184035	0.403679	0.533096	0.220652
Stamps	0.935484	0.920635	0.920635	0.807018	0.847458
WBC	0.952381	0.909091	0.869565	0.818182	0.869565
WDBC	1.000000	0.952381	1.000000	1.000000	0.952381
Wform	0.447552	0.345865	0.549020	0.597403	0.330827
campaign	0.553960	0.482738	0.582289	0.426646	0.570310
cardio	0.898305	0.842730	0.802432	0.797508	0.885714
letter	0.735135	0.310078	0.732240	0.762431	0.350365
Mammo	0.000000	0.757112	0.739785	0.699332	0.733333
mnist	0.877941	0.610811	0.867061	0.853293	0.811018
musk	0.941748	0.923077	0.946341	0.937198	0.955665
optdigits	0.908475	0.694981	0.955414	0.937500	0.046784
pendigits	0.948328	0.934985	0.945455	0.942598	0.913183
satimg-2	0.959459	0.958333	0.959459	0.958333	0.952381
shuttle	0.952265	0.948623	0.955244	0.952123	0.949644
thyroid	0.000000	0.930000	0.938776	0.869565	0.912821
vowels	0.942308	0.493151	0.933333	0.910891	0.641975

TABLE II: ROC AUC's across the data sets

Model	ABOD	IF	KNN	LOF	OCSVM
ALOI	0.793958	0.554444	0.708755	0.790706	0.559692
Stamps	0.962539	0.916753	0.955255	0.941727	0.932362
WBC	1.000000	0.990000	0.990000	0.830000	0.990000
WDBC	1.000000	1.000000	1.000000	1.000000	1.000000
Wform	0.712600	0.719500	0.776600	0.772500	0.696300
campaign	0.777896	0.707176	0.785627	0.664042	0.778764
cardio	0.957419	0.939695	0.919970	0.899600	0.953028
letter	0.882800	0.627400	0.868200	0.896900	0.609000
momamo	0.570614	0.891413	0.868720	0.828780	0.893868
mnist	0.946367	0.854253	0.946208	0.931088	0.911657
musk	1.000000	0.969922	1.000000	1.000000	1.000000
optdigits	0.959156	0.851956	0.940978	0.981733	0.575556
pendigits	0.999343	0.967867	0.997247	0.992398	0.971359
satimg-2	0.997620	0.991668	0.998413	0.994049	0.996429
shuttle	0.999911	0.995958	0.999079	0.999766	0.996221
thyroid	0.944387	0.989594	0.989479	0.932362	0.984738
vowels	0.998000	0.768400	0.970000	0.966400	0.814800

Before re-sampling, F1 score vs MCC, F1 score vs Kappa, ROC AUC vs PR AUC, and MCC vs Kappa metric pairs, exhibit a stable correlation. This implies that regardless of the fluctuations in the data set's sample size, these couples consistently exhibit a discernible pattern of correlation. As the data set size rises, both the F1 score and MCC are likely to increase in a closely correlated way, suggesting a robust and persistent association. Unspecified metric pairs exhibit less predictable patterns when the data amount fluctuates. Their association may exhibit greater variability or lack a direct progression, suggesting that these measurements may not exhibit a consistent correlation as the data set size grows.

Following the process of re-sampling, which aims to establish a more balanced data set or address imbalances in

TABLE III: PR AUC across the data set

Model	ABOD	IF	KNN	LOF	OCSVM
ALOI	0.800463	0.539385	0.699337	0.773590	0.562348
Stamps	0.921850	0.825251	0.920345	0.911731	0.885658
WBC	1.000000	0.990909	0.990909	0.734614	0.990909
WDBC	1.000000	1.000000	1.000000	1.000000	1.000000
Wform	0.691296	0.691371	0.798010	0.805158	0.684456
campaign	0.767378	0.724542	0.776394	0.656668	0.775187
cardio	0.953925	0.923795	0.913020	0.877406	0.938055
letter	0.861802	0.614909	0.837153	0.889306	0.635441
mammo	0.491392	0.905397	0.879681	0.850055	0.900483
mnist	0.943939	0.820166	0.939398	0.925495	0.909297
musk	1.000000	0.947800	1.000000	1.000000	1.000000
optdigits	0.928909	0.815586	0.852483	0.942914	0.499832
pendigits	0.999328	0.954694	0.996620	0.969080	0.953961
satimg-2	0.997725	0.992268	0.998425	0.993956	0.996737
shuttle	0.999910	0.996503	0.996507	0.999772	0.995193
thyroid	0.855370	0.989855	0.989511	0.922402	0.985288
vowels	0.998182	0.755790	0.964054	0.961758	0.823288

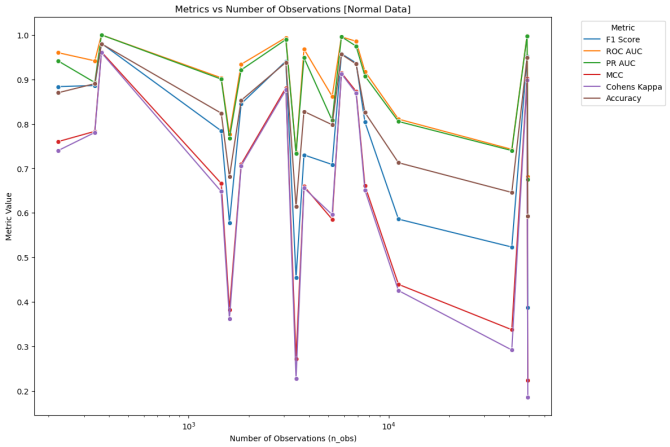

Fig. 1: Overall metrics performance over the number of observations

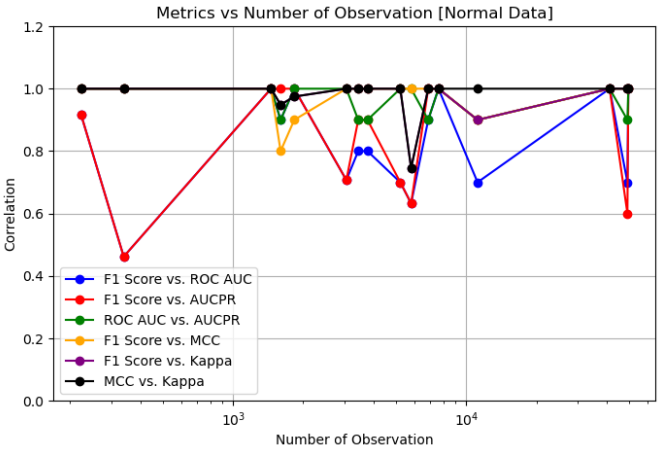

Fig. 2: Metric pairwise correlation trend over the number of observations before resampling applied

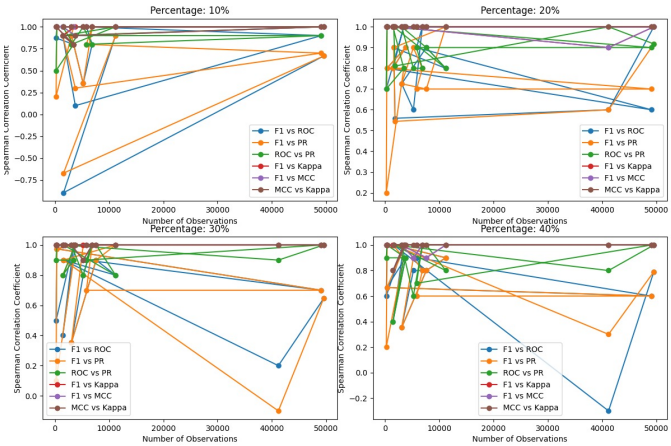

Fig. 3: Metric Pairwise Correlation trend over the number of observations

the existing data set, the previously consistent metric pairs maintain their degree of consistency. This indicates that the relationship between these indicators remains strong even after the data distribution is altered due to re-sampling. The consistent patterns found in certain pairings of measures, both before and after re-sampling, may suggest that these metrics are sensitive to variations in data volume and distribution. Metrics that exhibit a consistent correlation may be less influenced by the data set's size and distribution, therefore making them more dependable for estimating model performance, irrespective of changes in the data set.

For the model, the pattern follows the same as for the number of observations. Fig. 4 suggests that F1 scores are doing better with MCC and Cohen's Kappa rather than ROC AUC and PR AUC.

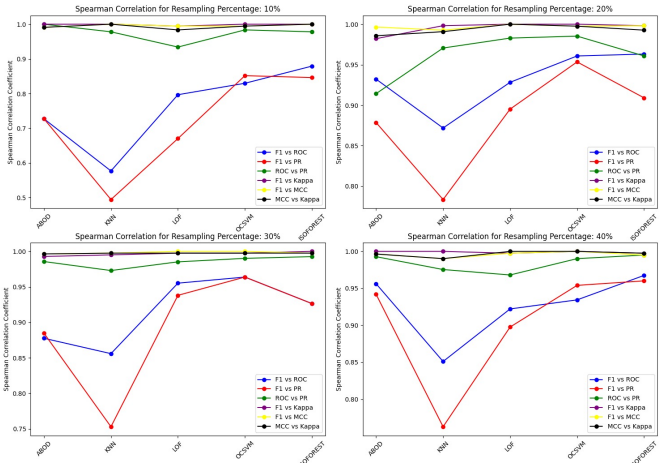

Fig. 4: Pairwise metric correlation to check model performance

V. Conclusion

The exploration of anomaly detection algorithms within varied and complex data landscapes underscores the criticality of metric selection in performance evaluation. Our extensive study, encompassing the analysis of Angle-based Outlier Detector (ABOD), K-Nearest Neighbors (KNN), Local Outlier Factor (LOF), Isolation Forest (IsoForest), and One-Class SVM (OCSVM), has provided a granular view of how these algorithms fare across datasets with resampled anomaly proportions. We found that while metrics like F1 score, ROC AUC, and PR AUC are widely used, they exhibit a degree of fluctuation in their correlation, particularly under extreme class imbalances and across different models. This was contrasted by the relative stability of MCC and Kappa, which proved to be less sensitive to the proportion of anomalies and offered a balanced measure of algorithm performance.

Our investigation reveals that the number of features in a dataset does not significantly influence the correlation of F1 vs. ROC, suggesting feature count may be a less critical factor than previously assumed. However, other metric pairs did not maintain this consistency, which may indicate a complex relationship between feature dimensions and metric reliability.

When metrics diverge, it presents a multifaceted challenge. Such disagreements necessitate a deeper dive into the distribution characteristics of the data and the particular goals of the anomaly detection task at hand. Discrepancies between metrics could indicate a need for recalibrating the decision threshold or could be a prompt to consider ensemble methods, potentially yielding a more cohesive performance narrative [17].

Interestingly, the PR AUC vs. ROC AUC correlation did not show as much variation [16] with resampled percentages as expected. This may suggest that in the context of the algorithms and datasets evaluated, both metrics are capturing aspects of model performance that are similarly impacted by changes in the anomaly percentage.

Our contribution to the field lies in advocating for a diversified metric approach in the evaluation of anomaly detection algorithms. Such an approach is instrumental for practitioners to select and fine-tune models that not only stand up to theoretical scrutiny but also exhibit practical resilience against the challenges presented by class imbalances.

In closing, our research elucidates the strengths and potential pitfalls of various anomaly detection algorithms when confronted with different levels of class imbalance, emphasizing the value of a comprehensive and multi-dimensional evaluation strategy. By adopting a suite of evaluation metrics, we enable practitioners to more confidently navigate the complex terrain of anomaly detection, assuring that the deployed models accurately reflect their operational efficacy and are not disproportionately influenced by the peculiarities of any single metric.

While our research has offered substantive insights into the performance of anomaly detection algorithms across class imbalances, it also opens avenues for future work. Subsequent research could explore the integration of additional machine learning models, especially those employing recent advancements in deep learning, which may provide different perspectives on anomaly detection in imbalanced datasets. Moreover, an examination of the impact of feature selection and engineering on the performance of these algorithms could yield further valuable findings.

Another promising area of future work involves the development of new metrics or the refinement of existing ones to better handle the nuances of highly imbalanced data, potentially leading to improved anomaly detection strategies. The exploration of hybrid models or ensemble techniques that combine the strengths of various algorithms could also be a path worth investigating, as they may offer a solution to the disagreements between metrics observed in our study.

Finally, considering the dynamic nature of data, longitudinal studies on the drift of anomaly characteristics over time and their impact on detection algorithms would significantly contribute to the robustness of anomaly detection systems in real-world applications.

By addressing these potential research areas, the academic and professional communities can continue to enhance the efficacy and applicability of anomaly detection methodologies, ensuring that they remain effective as data landscapes evolve.

VI. Acknowledgement

This work was supported by the Lamarr-Institute for ML and AI and the Research Center Trustworthy Data Science and Security.

References

[1] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-Based Outlier Detection in High-dimensional Data," in Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2008, pp. 444–452.

[2] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets," in Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data, 2000, pp. 427–438.

[3] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-based Local Outliers," in Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data, 2000, pp. 93–104.

[4] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in 2008 Eighth IEEE Int. Conf. on Data Mining, 2008, pp. 413–422.

[5] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," Neural Computation, vol. 13, no. 7, 2001, pp. 1443–1471.

[6] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," in Proc. of the 23rd Int. Conf. on Machine Learning, 2006, pp. 233–240.

[7] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric," PLOS ONE, vol. 12, no. 6, 2017, e0177678.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, 2002, pp. 321–357.

[9] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces", European Conference on Principles of Data Mining and Knowledge Discovery, pp. 15-27, 2002.

[10] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, "LOF: identifying density-based local outliers" in ACM sigmod record., ACM, 2000.

[11] F. T. Liu, K. M. Ting and Z.-H. Zhou, "Isolation-based anomaly detection", ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 6, no. 1, pp. 3, 2012.

[12] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola and R. C. Williamson, "Estimating the support of a high-dimensional distribution", Neural computation, vol. 13, no. 7, pp. 1443-1471, 2001.

[13] M. Amer, M. Goldstein and S. Abdennadher, "Enhancing one-class support vector machines for unsupervised anomaly detection", ACM SIGKDD Workshop on Outlier Detection and Description, 2013.

[14] M. Hu, Z. Ji, K. Yan, Y. Guo, X. Feng, J. Gong, et al., "Detecting anomalies in time series data via a meta-feature based approach", IEEE Access, vol. 6, pp. 27 760-27 776, 2018.

[15] E. Szmidt and J. Kacprzyk, "The Spearman rank correlation coefficient between intuitionistic fuzzy sets," 2010 5th IEEE International Conference Intelligent Systems, London, UK, 2010, pp. 276-280, doi: 10.1109/IS.2010.5548399.

[16] T. Fawcett, "An Introduction to ROC Analysis," in Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006.

[17] S. Klüttermann, E. Müller, "Evaluating and Comparing Heterogeneous Ensemble Methods for Unsupervised Anomaly Detection," in *2023 International Joint Conference on Neural Networks*, doi: 10.1109/IJCNN54540.2023.10191405 .