Impact of Evaluation Protocols on F1-Score and AVPR in Anomaly Detection Benchmarks

Assignee Research

June 13, 2026

Abstract

Anomaly detection is a widely explored domain in machine learning. Many models are proposed in the literature, and compared through different metrics measured on various datasets. The most popular metrics used to compare performances are F1-score, AUC and AVPR. In this paper, we show that F1-score and AVPR are highly sensitive to the contamination rate. One consequence is that it is possible to artificially increase their values by modifying the train-test split procedure. This leads to misleading comparisons between algorithms in the literature, especially when the evaluation protocol is not

1 Introduction

This paper examines: Anomaly Detection: How to Artificially Increase your F1-Score with a Biased Evaluation Protocol. Research question: How do different evaluation protocols (e.g., stratified vs. random splits) affect the F1-score and AVPR metrics in anomaly detection benchmarks across diverse domains?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

13 papers retrieved. 22 claims extracted; 16 independently verified. Quality review score: 7.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The evaluation of an algorithm should be done on a test set completely separated from the train set.	✓	0.26
Algorithm 1 presents the unbiased procedure to train and evaluate an anomaly detection model.	✓	0.23
The anomalous samples from the train set are removed to get a clean set that is used to train a model.	✓	0.33
The train set is also used to compute the contamination rate and fix the threshold.	✓	0.20
The threshold is fixed such that the train set has as many anomalies as predicted anomalies, i.e., fp = fn.	✓	0.22
The threshold is finally used on the predictions made on the new (unseen) samples composing the test set to measure the	✓	0.28
The AUC and AVPR are computed using the predicted scores directly.	✓	0.21
The anomalous samples in the train set are used only to compute the threshold for the F1-score and are then thrown away.	✓	0.37
The more anomalous samples we can use to evaluate a model, the more precise the evaluation.	✓	0.29
Algorithm 2 recycles the anomalous samples contained in the train set.	✓	0.23
The threshold is computed on the test set as there are no anomalies left in the train set to estimate it.	✓	0.30
This recycling procedure leads to a situation where precision = recall = F1-score.	✓	0.21
The recycling procedure makes sense in the context of anomaly detection as it obtains more precise results.	✓	0.26
The recycling procedure can be found in the literature [24,31].	×	0.15
Algorithms 1 and 2 take as input any dataset and any trainable anomaly-score function.	✓	0.30
The Arrhythmia and Thyroid datasets from the ODDS repository [20] and the Kddcup dataset from the UCI repository [4] are	✓	0.32
The Arrhythmia dataset has 452 samples with a contamination rate of 14.6%.	×	0.13
The Thyroid dataset has 3772 samples.	×	0.08
The Kddcup dataset has 494020 samples.	×	0.14
The theoretical F1-score increases with the contamination rate of the test set.	×	0.12
The theoretical F1-score is shown to be increasing with the contamination rate in Figure 5.	×	0.12
When the threshold t for the F1-score is computed using the *test set* as done in Algorithm 2	✓	0.19

References

- http://arxiv.org/abs/2509.09030v2
- http://arxiv.org/abs/2106.16020v1
- http://arxiv.org/abs/2509.00042v1