Multimodal Anomaly Detection: CLIP-Based vs Diffusion Models on Cross-Domain OpenML Benchmarks

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How do multimodal anomaly detection models (e.g., CLIP-based) compare to diffusion models in terms of F1-score robustness when evaluated on cross-domain OpenML benchmarks with class-imbalanced test. 21 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Anomaly Detection: How to Artificially Increase your F1-Score with a Biased Evaluation Protocol. Research question: How do multimodal anomaly detection models (e.g., CLIP-based) compare to diffusion models in terms of F1-score robustness when evaluated on cross-domain OpenML benchmarks with class-imbalanced test sets?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.7/10.

3 Results

10 papers retrieved. 21 claims extracted; 0 independently verified. Quality review score: 2.7/10.

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.	×	0.05
Algorithm 1 presents the unbiased procedure to train and evaluate an anomaly detection model.	×	0.08
The anomalous samples from the train set are removed to get a clean set that is used to train a model.	×	0.04
The train set is also used to compute the contamination rate and fix the threshold.	×	0.06
The threshold is fixed such that the train set has as many anomalies as predicted anomalies, i.e., fp = fn.	×	0.03
This threshold is finally used on the predictions made on the new (unseen) samples composing the test set to measure the	×	0.07
The AUC and AVPR are computed using the predicted scores directly.	×	0.08
The anomalous samples in the train set are used only to compute the threshold for the F1-score and are then thrown away.	×	0.08
The more anomalous samples we can use to evaluate a model, the more precise the evaluation.	×	0.05
Algorithm 2 recycles the anomalous samples contained in the train set.	×	0.03
The threshold is then computed on the test set as there are no anomalies left in the train set to estimate it.	×	0.03
This leads to a situation where precision = recall = F1-score.	×	0.09
This recycling procedure makes sense in the context of anomaly detection as it obtains more precise results.	×	0.06
Algorithms 1 and 2 take as input any dataset and any trainable anomaly-score function.	×	0.05
The Arrhythmia and Thyroid datasets from the ODDS repository and the Kddcup dataset from the UCI repository are often us	×	0.08
The Arrhythmia dataset has 452 samples with a contamination rate of 14.6%.	×	0.03
The Thyroid dataset has 3772 samples.	×	0.03
The Kddcup dataset has 494020 samples.	×	0.03
The theoretical F1-score increases with the contamination rate of the test set.	×	0.11
The theoretical F1-score is shown to increase with the contamination rate in Figure 5.	×	0.13
When the threshold t for the F1-score is computed using the test set as done in Algorithm 2, recall = precision = F1-sco	×	0.10

References

- http://arxiv.org/abs/2410.01534v2
- http://arxiv.org/abs/2502.14293v2
- http://arxiv.org/abs/2106.16020v1