Train-Test Split Contamination and F1-Score Inflation in Code Generation Models

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of train-test split contamination on F1-score inflation for code generation models on CodeXGLUE security subsets. 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. 18 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. 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: What is the impact of train-test split contamination on F1-score inflation for code generation models on CodeXGLUE security subsets.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

14 papers retrieved. 18 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
Machine learning theory dictates that algorithm evaluation should be performed on a test set completely separated from t	×	0.05
In the unbiased evaluation procedure (Algorithm 1), anomalous samples are removed from the training set to create a clea	×	0.04
In Algorithm 1, the threshold is computed using the training set such that the number of predicted anomalies equals the	×	0.01
In Algorithm 1, the threshold computed on the training set is applied to predictions on the unseen test set to measure t	×	0.07
In Algorithm 1, AUC and AVPR are computed using predicted scores directly.	×	0.08
In the biased evaluation procedure (Algorithm 2), anomalous samples from the training set are recycled into the test set	×	0.09
In Algorithm 2, the threshold is computed on the test set because no anomalies remain in the training set to estimate it	×	0.03
The recycling procedure described in Algorithm 2 results in precision, recall, and F1-score being equal.	×	0.09
The study utilizes the Arrhythmia and Thyroid datasets from the ODDS repository.	×	0.02
The study utilizes the Kddcup dataset from the UCI repository.	×	0.03
The Arrhythmia dataset contains 452 samples with a contamination rate of 14.6%.	×	0.04
The Thyroid dataset contains 3,772 samples.	×	0.03
The Kddcup dataset contains 494,020 samples.	×	0.02
Recall (p+) does not depend on the contamination rate (α).	×	0.03
Precision increases as the ratio of anomalous to normal samples in the test set (N+t/N-t) increases, and therefore incre	×	0.03
The AVPR increases with the contamination rate (α) because it is driven by increasing precision.	×	0.07
The F1-score with a fixed threshold increases with the contamination rate (α) as it is the harmonic mean of a constant r	×	0.08
Figure 5 illustrates the theoretical F1-score for varying contamination rates of the test set.	×	0.09

References

- http://arxiv.org/abs/2504.20900v1
- http://arxiv.org/abs/2106.16020v1
- http://arxiv.org/abs/2303.12869v1