CausalMixFT F1-Score Scaling with Dataset Size in TabularGLUE Benchmarks

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: To what extent does the F1-score improvement from CausalMixFT over SMOTE-based augmentation scale with increasing dataset size in the TabularGLUE benchmarks, and how does this scaling behavior differ. 17 claims were extracted from source literature; 1 was 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: To what extent does the F1-score improvement from CausalMixFT over SMOTE-based augmentation scale with increasing dataset size in the TabularGLUE benchmarks, and how does this scaling behavior differ across high-variance and low-variance tabular datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 12 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

12 papers retrieved. 17 claims extracted; 1 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
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 from the training set are removed to create a clea	×	0.05
In Algorithm 1, the threshold is computed using the training set such that the number of false positives equals the numb	×	0.02
In Algorithm 1, AUC and AVPR are computed using predicted scores directly on the test set.	×	0.08
Algorithm 2 recycles anomalous samples from the training set by moving them to the test set.	×	0.01
In Algorithm 2, the threshold is computed on the test set because no anomalies remain in the training set to estimate it	×	0.02
The recycling procedure in Algorithm 2 results in precision, recall, and F1-score being equal.	×	0.11
The study utilizes the Arrhythmia and Thyroid datasets from the ODDS repository.	×	0.03
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%.	×	0.03
The Thyroid dataset contains 3,772 samples.	×	0.04
The Kddcup dataset contains 494,020 samples.	×	0.04
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 increases, and therefore increases with	×	0.03
Theoretical analysis proves that AVPR increases with the contamination rate (α).	×	0.06
The F1-score with a fixed threshold increases with the contamination rate (α).	×	0.13
Figure 5 illustrates the theoretical F1-score for varying contamination rates, anomaly-detection capabilities (p+), and	✓	0.15

References

- http://arxiv.org/abs/2109.05633v1
- http://arxiv.org/abs/2506.16791v4
- http://arxiv.org/abs/2106.16020v1