How does the robustness of OpenPangu-7B-MLA compare to other multilingual prosody-aware models on EchoMind und

Assignee Research

June 10, 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 does the robustness of OpenPangu-7B-MLA compare to other multilingual prosody-aware models on EchoMind under high-noise conditions, measured by F1-score and word error rate?.

2 Methodology

Systematic literature search across multiple databases yielded 11 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

11 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	×	0.06
In the unbiased evaluation procedure (Algorithm 1), anomalous samples are removed from the training set 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	×	0.01
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.	×	0.07
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 equal to recall, which is equal to the F1-score.	×	0.09
The study utilizes the Arrhythmia and Thyroid datasets from the ODDS repository.	×	0.03
The study utilizes the Kddcup dataset from the UCI repository.	×	0.04
The Arrhythmia dataset contains 452 samples with a contamination rate of 14.6%.	×	0.03
The Thyroid dataset contains 3772 samples.	×	0.03
The Kddcup dataset contains 494020 samples.	×	0.03
Recall (p+) does not depend on the contamination rate (α).	×	0.03
Precision increases as the ratio of anomalous samples to normal samples (N+t / N-t) increases, and therefore increases w	×	0.02
The AVPR increases with the contamination rate (α) because precision increases while other values in the equation remain	×	0.06
The F1-score with a fixed threshold increases with the contamination rate (α) because it is the harmonic mean of a const	×	0.07
Figure 5 illustrates the theoretical F1-score for varying contamination rates of the test set.	×	0.08

References

- http://arxiv.org/abs/2106.16020v1
- http://arxiv.org/abs/2510.22758v2
- http://arxiv.org/abs/2509.06951v2