Reconstruction-Based vs. Permutation-Based Self-Supervised Learning for Tabular Model Robustness

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the comparative effect of reconstruction-based versus permutation-based self-supervised learning on the corruption robustness scores of tabular models in the RBT framework. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Survey on Self-Supervised Learning for Non-Sequential Tabular Data. Research question: What is the comparative effect of reconstruction-based versus permutation-based self-supervised learning on the corruption robustness scores of tabular models in the RBT framework?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

15 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
OpenML-CC18 contains 72 datasets with sample sizes ranging from 500 to 92,000 and feature counts from 5 to 3,073.	×	0.00
DLBench includes 11 datasets for classification and regression, with sample sizes ranging from 7,000 to 1,000,000 and fe	×	0.02
TabularBench consists of 45 datasets for classification and regression, with sample sizes ranging from 3,000 to 10,000 a	×	0.03
TabZilla contains 36 classification datasets with sample sizes ranging from 300 to 1,000,000 and feature counts from 7 t	×	0.01
TP-BERTa includes 202 unlabeled datasets with sample sizes ranging from 10,000 to 100,000 and feature counts from 1 to 3	×	0.02
TP-BERTa includes 145 labeled datasets for classification and regression, with sample sizes ranging from 10 to 9,800 and	×	0.03
OpenTabs contains 2,000 unlabeled datasets with an average of 23,000 samples and 24 features.	×	0.02
UniTabE contains 283,000 unlabeled datasets with an average of 46,000 samples and 31 features.	×	0.02
Levin et al (2023) introduced a pseudo-feature approach for pre-training deep tabular models to predict missing features	×	0.07
Ye et al (2023) pre-trained a Transformer encoder with 2,000 high-quality cross-table datasets using masked table modeli	×	0.09
DoRA (Du et al, 2023) focuses on designing a pretext task based on domain knowledge in the financial domain for real est	×	0.04
SSL4NS-TD is highly application-oriented and represents ubiquitous practical utility in diverse domains, including medic	×	0.10

References

- `http://arxiv.org/abs/2405.08815v1`

- http://arxiv.org/abs/2210.10599v1
- http://arxiv.org/abs/2402.01204v4