

Impact of Synthetic Tabular Feature Dimensions on Self-Supervised Model Classification Accuracy

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of varying synthetic tabular feature dimensions on the final downstream classification accuracy of self-supervised foundation models. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. 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 impact of varying synthetic tabular feature dimensions on the final downstream classification accuracy of self-supervised foundation models?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

13 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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 consists of 72 datasets with sample sizes ranging from 500 to 92,000 and feature counts ranging from 5 to 3,	×	0.02
DLBench includes 11 datasets with sample sizes ranging from 7,000 to 1,000,000 and feature counts ranging from 10 to 2,0	×	0.02
TabularBench comprises 45 datasets with sample sizes ranging from 3,000 to 10,000 and feature counts ranging from 5 to 6	×	0.02
TabZilla contains 36 datasets with sample sizes ranging from 300 to 1,000,000 and feature counts ranging from 7 to 4,297	×	0.02
TP-BERTa includes 202 unlabeled datasets with sample sizes ranging from 10,000 to 100,000 and feature counts ranging from	×	0.02
TP-BERTa includes 145 labeled datasets with sample sizes ranging from 10 to 9,800 and feature counts ranging from 3 to 3	×	0.03
OpenTabs consists of 2,000 unlabeled datasets with an average of 23,000 samples and 24 features.	×	0.03
UniTabE consists of 283,000 unlabeled datasets with an average of 46,000 samples and 31 features.	×	0.03
Levin et al (2023) introduced a pseudo-feature approach for pre-training that predicts missing features in upstream data	×	0.06
Ye et al (2023) pre-trained a Transformer encoder with 2,000 high-quality cross-table datasets using masked table models	×	0.06
DoRA (Du et al, 2023) focuses on designing a pretext task based on domain knowledge in the financial domain for real est	×	0.04
DoRA uses an intra-sample pretext task by selecting the domain-specific feature of a sample as the self-supervised label	×	0.06
DoRA adopts inter-sample contrastive learning to separate dissimilar samples based on the domain-specific feature.	×	0.05
Tabular data represent ubiquitous practical utility in diverse domains, including medicine, finance, and many other area	×	0.06

References

- <http://arxiv.org/abs/2212.12411v1>
- <http://arxiv.org/abs/2505.15103v2>
- <http://arxiv.org/abs/2402.01204v4>