

Synthetic Tabular Feature Dimensionality and Downstream Classification in Self-Supervised Models

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does varying the dimensionality of synthetic tabular features impact the downstream classification accuracy of self-supervised foundation models compared to standard normalization techniques. 19 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: A Survey on Self-Supervised Learning for Non-Sequential Tabular Data. Research question: How does varying the dimensionality of synthetic tabular features impact the downstream classification accuracy of self-supervised foundation models compared to standard normalization techniques?.

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. 19 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
The OpenML-CC18 benchmark contains 72 datasets.	×	0.02
The OpenML-CC18 benchmark datasets contain between 500 and 92,000 samples.	×	0.02
The OpenML-CC18 benchmark datasets contain between 5 and 3,073 features.	×	0.04
The DLBench benchmark includes both classification and regression tasks.	×	0.08
The DLBench benchmark contains 11 datasets.	×	0.02
The TabularBench benchmark contains 45 datasets.	×	0.02
The TabZilla benchmark contains 36 datasets focused on classification tasks.	×	0.03
The TP-BERTa benchmark includes 202 unlabeled datasets and 145 classification/regression datasets.	×	0.06
The OpenTabs dataset consists of 2,000 unlabeled datasets.	×	0.03
The average number of samples in the OpenTabs dataset is 23,000.	×	0.01
The UniTabE dataset consists of 283,000 unlabeled datasets.	×	0.03
Masking strategies completely remove targeted features, whereas perturbations leave partial information.	×	0.03
Levin et al. (2023) introduced a pseudo-feature approach to predict missing features in upstream data that are present i	×	0.04
Ye et al. (2023) pre-trained a Transformer encoder using 2,000 high-quality cross-table datasets.	×	0.04
Analyses indicate that pre-training provides more transferability over tree-based baselines.	×	0.05
The DoRA method (Du et al., 2023) designs a pretext task based on domain knowledge in the financial domain for real esta	×	0.04
DoRA uses an intra-sample pretext task where the self-supervised label is a domain-specific feature, such as predicting	×	0.06
DoRA adopts inter-sample contrastive learning to cluster samples with the same domain-specific feature, such as real est	×	0.05
Tabular data has practical utility in4diverse domains including medicine and finance.	×	0.09

References

- <http://arxiv.org/abs/2112.02962v4>
- <http://arxiv.org/abs/2008.10312v2>
- <http://arxiv.org/abs/2402.01204v4>