Novel Tabular Generative Metrics and Downstream Classification Accuracy in OpenML-CC18 Hybrids

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How do novel tabular generative metrics correlate with downstream classification accuracy when evaluating synthetic data used to augment multimodal models on OpenML-CC18 hybrids. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Tabular Data Augmentation for Machine Learning: Progress and Prospects of Embracing Generative AI. Research question: How do novel tabular generative metrics correlate with downstream classification accuracy when evaluating synthetic data used to augment multimodal models on OpenML-CC18 hybrids?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

10 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 4.7/10.

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
Tabular data is heterogeneous, typically containing both dense numerical and sparse categorical attributes.	×	0.09
Tabular data has complicated structure, such as row and column permutation invariance and hierarchical organization, whe	×	0.05
Many TDA tasks involve large-scale table pools, sometimes encompassing millions of tables.	×	0.05
These tables often have inconsistent attribute naming and value formatting, and table pools themselves are dynamic, chan	×	0.03
TDA methods can be broadly categorized into retrieval-based approaches, which involve retrieving data from table pools,	×	0.11
The initial ML model yields sub-optimal results due to insufficient data and numerous missing or incorrect values.	×	0.04
Augmentation can be achieved through retrieval-based methods or generation-based methods that synthesize new data.	×	0.14
After augmentation, evaluation steps evaluate the effectiveness of the TDA process.	×	0.07
The result-TDA table enables the scientist to train a more accurate price prediction model.	×	0.04
Generative AI methods for TDA include VAE, GAN, Diffusion, Pretrained LM, and Large LM.	×	0.08
External data sources can contain various table sources, including databases and Web tables.	×	0.06
Generative methods for TDA include statistical approaches such as MICE and deep generative models like diffusion models.	×	0.10
The augmented analytics market is growing, as indicated by market research reports.	×	0.02

References

- http://arxiv.org/abs/2407.21523v1
- http://arxiv.org/abs/2504.20900v1
- http://arxiv.org/abs/2512.03307v1