Does the effectiveness of an overlap class in synthetic data generation for imbalanced tabular data scale cons

Assignee Research

June 10, 2026

Abstract

Due to their data-driven nature, Machine Learning (ML) models are susceptible to bias inherited from data, especially in classification problems where class and group imbalances are prevalent. Class imbalance (in the classification target) and group imbalance (in protected attributes like sex or race) can undermine both ML utility and fairness. Although class and group imbalances commonly coincide in real-world tabular datasets, limited methods address this scenario. While most methods use oversampling techniques, like interpolation, to mitigate imbalances, recent advancements in synthetic tab

1 Introduction

This paper examines: Synthetic Tabular Data Generation for Class Imbalance and Fairness: A Comparative Study. Research question: Does the effectiveness of an overlap class in synthetic data generation for imbalanced tabular data scale consistently across different deep generative architectures like VAEs and GANs?.

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 1.7/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 1.7/10.

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.

References

- http://arxiv.org/abs/2409.05215v1
- http://arxiv.org/abs/2502.17119v2
- http://arxiv.org/abs/2104.11797v1