Synthetic Data Rebalancing via Tabular Distribution Matching for Fair Code Generation

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does rebalancing synthetic training data via tabular distribution matching affect fairness metrics on imbalanced code generation benchmarks like HumanEval-X. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Diffusion and Flow Matching Models for Tabular Data: A Survey. Research question: How does rebalancing synthetic training data via tabular distribution matching affect fairness metrics on imbalanced code generation benchmarks like HumanEval-X?.

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 3.3/10.

3 Results

11 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
Future benchmarks should include missing-data settings, imbalanced classes, high-cardinality categorical features, relat	×	0.06
Formal differential privacy has begun to appear in tabular diffusion models [116], [117].	×	0.06
Recent memorization studies [118], [119] indicate that a small subset of records may dominate memorized generations.	×	0.02
Anomaly detection methods such as TCCM [71] show that feature-level residuals can support interpretability.	×	0.05
Healthcare models such as FlexGen-EHR [85] and PatientFlow [41] point toward multimodal and longitudinal tabular generat	×	0.04
Several recent methods combine diffusion or flow matching with autoencoders, transformers, tree models, or feature-token	×	0.13
Diffusion SOS [64] 2022 KDD Synthesis (single, generic) SDEs 6 Utility Generic.	×	0.02
STaSy [32] 2023 ICLR Synthesis (single, generic) (MM) (OH) SDEs 15 Fidelity, Utility, Diversity Generic.	×	0.02
TabDDPM [8] 2023 ICML Synthesis (single, generic) (QT) (OH) DDPM+MLD 16 Fidelity, Utility, Privacy Generic.	×	0.02
CoDi [34] 2023 ICML Synthesis (single, generic) (MM) (OH) DDPM+MLD 15 Utility, Diversity Generic.	×	0.02
AutoDiff [33] 2023 NeurIPSW Synthesis (single, generic) Any 15 Fidelity, Utility, Privacy Generic.	×	0.03
MissDiff [76] 2023 ICMLW Synthesis (single, generic) (MM) (OH) SDEs 3 Fidelity.	×	0.01
Data augmentation for tabular data is a long-standing research problem [75].	×	0.11
Data augmentation can be divided into two different tasks: 1) data synthesis and 2) over-sampling.	×	0.10
Over-sampling can be considered as a special case of single table synthesis where we only generate a part of the table.	×	0.03

References

- http://arxiv.org/abs/2104.11797v1
- http://arxiv.org/abs/2403.10075v2
- http://arxiv.org/abs/2502.17119v2