

Multimodal vs. Tabular-Only Generative Models on Imbalanced Data Performance

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: How do multimodal generative models (e.g., combining tabular and textual data) perform compared to tabular-only models like CausalMixFT on imbalanced datasets, using Frchet Distance and downstream. 18 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Frchet Radiomic Distance (FRD): A Versatile Metric for Comparing Medical Imaging Datasets. Research question: How do multimodal generative models (e.g., combining tabular and textual data) perform compared to tabular-only models like CausalMixFT on imbalanced datasets, using Frchet Distance and downstream classification accuracy as evaluation metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

7 papers retrieved. 18 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
FRDv0 features over ImageNet or RadImageNet result in noticeably improved average accuracy and sensitivity for breast MR	×	0.02
FRDv0 achieves on-par specificity compared to ImageNet and RadImageNet features for breast MRI domain shift cases.	×	0.02
FRD improves on FRDv0 noticeably in AUC and sensitivity.	×	0.03
FRD is roughly on-par with FRDv0 for accuracy and specificity.	×	0.02
For datasets with images from multiple domains of the same patient (e.g., BraTS), random sampling is used to ensure train	×	0.03
The Duke Breast Cancer MRI (DBC) dataset contains 12K training, 2.4K validation, and 2.6K test images.	×	0.03
The BraTS dataset involves a domain shift from T1 to T2 modalities.	×	0.02
The BraTS dataset contains 28K training, 6K validation, and 6K test images.	×	0.04
The TotalSegmentator and in-house MRIs dataset for lumbar spine involves a domain shift from T1 MRI to CT.	×	0.01
The CHAOS dataset involves a domain shift from CT to T1 MRI (in-phase).	×	0.03
For OOD detection on Breast MRI using AUC, FRD achieved a score of 1.00 while ImageNet achieved 0.43.	×	0.04
For OOD detection on Breast MRI using Sensitivity, FRD achieved a score of 1.00 while ImageNet achieved 0.03.	×	0.04
FRD achieved an average AUC of 0.94 across all evaluated datasets, outperforming ImageNet (0.79) and RadImageNet (0.74).	×	0.02
In almost all evaluated downstream tasks, there is a drop in average performance on test data detected as OOD compared to	×	0.04
FRD outperforms other metrics in ranking which OOD datasets will result in worse downstream task performance.	×	0.13
Common perceptual metrics like FID are based on natural image features rather than medical image features.	✓	0.17
Recent findings suggest that FID and KID metrics may poorly reflect medical image quality.	×	0.11
Direct qualitative image assessment by radiologists is expensive and non-standardized.	×	0.04

References

- <http://arxiv.org/abs/2504.20900v1>
- <http://arxiv.org/abs/2412.01496v2>
- <http://arxiv.org/abs/2201.07932v1>