

# Impact of Data Augmentation Strategies on Calibration in Tabular Foundation Models for Medical Imaging

Assignee Research

June 9, 2026

## Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of different data augmentation strategies (e.g., CausalMixFT, MixUp, SMOTE) on the calibration performance (e.g., expected calibration error) of tabular foundation models. 8 claims were extracted from source literature; 0 were 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: What is the impact of different data augmentation strategies (e.g., CausalMixFT, MixUp, SMOTE) on the calibration performance (e.g., expected calibration error) of tabular foundation models fine-tuned on medical imaging datasets under domain shifts?.

## 2 Methodology

Systematic literature search across multiple databases yielded 12 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

12 papers retrieved. 8 claims extracted; 0 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 results in noticeably improved average accuracy and sensitivity, and on-par	×	0.04
FRD improves on FRDv0 noticeably in AUC and sensitivity, and is roughly on-par for accuracy and specificity.	×	0.03
FRD detects when task model performance drops on OOD data compared to ID performance.	×	0.05
FRD outperforms other metrics in ranking which of different OOD datasets will result in worse downstream task performance	×	0.12
In almost all cases, there is a drop in average performance on test data that was detected as OOD using the binary thres	×	0.04
FRD is designed for comparing unpaired distributions of real and/or generated medical images.	×	0.07
FID and KID may poorly reflect medical image quality.	×	0.08
No studies have proposed interpretable metrics specifically tailored for comparing unpaired medical image distributions.	×	0.09

## References

- <http://arxiv.org/abs/2408.08058v1>
- <http://arxiv.org/abs/2412.01496v2>
- <http://arxiv.org/abs/1808.05205v1>