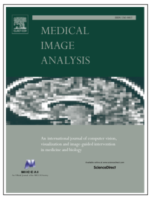
Medical Image Analysis 110 (2026) 103943

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media



Fréchet radiomic distance (FRD): A versatile metric for comparing medical imaging datasets

Nicholas Konz [a,1,*], Richard Osuala [b,c,d,1], Preeti Verma [b], Yuwen Chen [a], Hanxue Gu [a], Haoyu Dong [a], Yaqian Chen [a], Andrew Marshall [e,*], Lidia Garrucho [b], Kaisar Kushibar [b], Daniel M. Lang [c,d], Gene S. Kim [f], Lars J. Grimm [g], John M. Lewin [h], James S. Duncan [e,h,i], Julia A. Schnabel [c,d,j], Oliver Diaz [b,k], Karim Lekadir [b,l], Maciej A. Mazurowski [a,g,m,n]

[a] *Department of Electrical and Computer Engineering, Duke University, NC, USA*
[b] *Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain*
[c] *Institute of Machine Learning in Biomedical Imaging, Helmholtz Center Munich, Munich, Germany*
[d] *School of Computation, Information and Technology, Technical University of Munich, Munich, Germany*
[e] *Department of Biomedical Engineering, Yale University, CT, USA*
[f] *Department of Radiology, Weill Cornell Medical College, NY, USA*
[g] *Department of Radiology, Duke University, NC, USA*
[h] *Department of Radiology & Biomedical Imaging, Yale University, CT, USA*
[i] *Department of Electrical Engineering, Yale University, CT, USA*
[j] *School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK*
[k] *Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra, Spain*
[l] *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*
[m] *Department of Biostatistics & Bioinformatics, Duke University, NC, USA*
[n] *Department of Computer Science, Duke University, NC, USA*

ARTICLE INFO

Keywords:
Metrics
Evaluation
Image similarity
Generative models
OOD detection
Image-to-image translation

ABSTRACT

Determining whether two sets of images belong to the same or different distributions or domains is a crucial task in modern medical image analysis and deep learning; for example, to evaluate the output quality of image generative models. Currently, metrics used for this task either rely on the (potentially biased) choice of some downstream task, such as segmentation, or adopt task-independent perceptual metrics (*e.g.*, Fréchet Inception Distance/FID) from natural imaging, which we show insufficiently capture anatomical features. To this end, we introduce a new perceptual metric tailored for medical images, FRD (Fréchet Radiomic Distance), which utilizes standardized, clinically meaningful, and interpretable image features. We show that FRD is superior to other image distribution metrics for a range of medical imaging applications, including out-of-domain (OOD) detection, the evaluation of image-to-image translation (by correlating more with downstream task performance as well as anatomical consistency and realism), and the evaluation of unconditional image generation. Moreover, FRD offers additional benefits such as stability and computational efficiency at low sample sizes, sensitivity to image corruptions and adversarial attacks, feature interpretability, and correlation with radiologist-perceived image quality. Additionally, we address key gaps in the literature by presenting an extensive framework for the multifaceted evaluation of image similarity metrics in medical imaging—including the first large-scale comparative study of generative models for medical image translation—and release an accessible codebase to facilitate future research. Our results are supported by thorough experiments spanning a variety of datasets, modalities, and downstream tasks, highlighting the broad potential of FRD for medical image analysis.

---

* Corresponding authors.
E-mail addresses: nicholas.konz@duke.edu (N. Konz), richard.osuala@ub.edu (R. Osuala), yuwen.chen@duke.edu (Y. Chen), hanxue.gu@duke.edu (H. Gu), haoyu.dong151@duke.edu (H. Dong), yaqian.chen@duke.edu (Y. Chen), andrew.marshall@yale.edu (A. Marshall), lgarrucho@ub.edu (L. Garrucho), kaisar.kushibar@ub.edu (K. Kushibar), lang@helmholtz-munich.de (D.M. Lang), sgk4001@med.cornell.edu (G.S. Kim), lars.grimm@duke.edu (L.J. Grimm), john.lewin@yale.edu (J.M. Lewin), james.duncan@yale.edu (J.S. Duncan), julia.schnabel@tum.de (J.A. Schnabel), oliver.diaz@ub.edu (O. Diaz), karim.lekadir@ub.edu (K. Lekadir), maciej.mazurowski@duke.edu (M.A. Mazurowski).

[1] Equal Contribution.

https://doi.org/10.1016/j.media.2026.103943
Received 8 May 2025; Received in revised form 1 December 2025; Accepted 9 January 2026

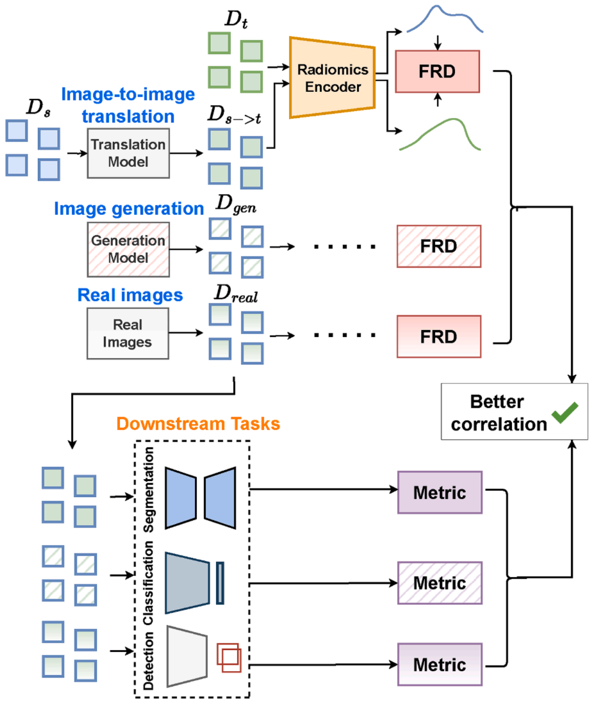


Fig. 1. Summary of our main contribution: **FRD**, a metric designed from the ground up for comparing unpaired distributions of real and/or generated medical images.

1. Introduction

Comparing image distributions is crucial in deep learning-driven medical image analysis. Example applications include out-of-domain (OOD) detection (Tschuchnig and Gadermayr, 2022), *e.g.*, for detecting if new medical images were acquired using different protocols; the evaluation of image-to-image translation models, *e.g.*, for converting MRI (magnetic resonance imaging) to CT (computed tomography) (Wolterink et al., 2017; Armanious et al., 2020); the quality assessment of images generated to supplement real training data (Chen et al., 2021; Pinaya et al., 2022); and others (Chan et al., 2020).

However, image distribution metrics from general computer vision (*e.g.*, Fréchet Inception Distance/FID (Heusel et al., 2017)) often miss key requirements for medical image analysis, and questioning of their unadapted use for this subfield has recently begun (Osuala et al., 2023a; Woodland et al., 2024; Konz et al., 2024b; Wu et al., 2025). For example, in medical image OOD detection and image-to-image translation, the focus extends beyond just general image quality to *image-level domain adaptation*: ensuring that source-domain images (*e.g.*, from one scanner, vendor or institution) are compatible with diagnostic models trained on target-domain images from another scanner/vendor/institution, addressing ubiquitous domain shift issues in medical imaging (Durrer et al., 2024; Beizaee et al., 2023; Wang et al., 2024b; Modanwal et al., 2020; Yang et al., 2019; Liu et al., 2021; Zhang et al., 2018; Guan and Liu, 2021; Yao et al., 2019; Mårtensson et al., 2020). Additionally, medical imaging requires metrics that specifically capture anatomical consistency and realism, as well as clinical interpretability (Maier-Hein et al., 2024; Salahuddin et al., 2022; Chen et al., 2022a; Singh et al., 2020). We argue that these specialized requirements are overlooked by the current metrics used for comparing sets of real and/or synthetic medical images.

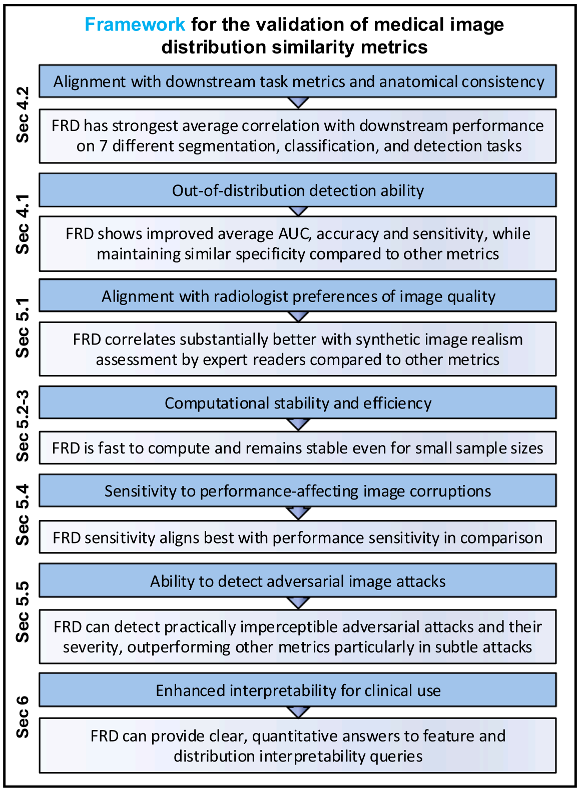
The common approach of comparing medical image distributions in terms of the performance of some downstream task such as segmentation (because direct qualitative image assessment by radiologists is expensive and non-standardized) is driven by the choice of task, and requires costly training and labeling efforts. A task-independent metric that captures general image quality *and* aligns with expected downstream task performance would therefore be preferable. In computer vision, perceptual metrics like FID are commonly used to evaluate image quality relative to real images (Heusel et al., 2017; Saxena and Teli, 2021; Bińkowski et al., 2018), yet these metrics are based on natural image features. Despite this, many applications of medical image translation (Li et al., 2023b; Wang et al., 2024b; Li et al., 2023a; Shi et al., 2023) and generation (Pinaya et al., 2022; Hashmi et al., 2024; Hansen et al., 2024; Sun et al., 2022) rely on FID (or the related Kernel Inception Distance/KID (Bińkowski et al., 2018)) for evaluation, even though recent findings suggest these metrics may poorly reflect medical image quality (Konz et al., 2024a,b; Chen et al., 2024b; Wu et al., 2025). Our experiments further support this issue. Moreover, to date, no studies have proposed *interpretable* metrics specifically tailored for comparing unpaired medical image distributions, despite the importance of explainability in medical image analysis.

In this paper, we showcase and address limitations in current metrics for comparing unpaired medical image distributions. We begin by evaluating "RadiologyFID" (RadFID) (Osuala et al., 2023b), a natural extension from FID which uses RadImageNet (Mei et al., 2026) features instead of ImageNet (Deng et al., 2009a) features, that has seen surprisingly little use. We find that RadFID improves upon prior metrics in some areas, yet lacks in interpretability, stability on small datasets, and other essential qualities.

To address these gaps, we introduce Fréchet Radiomic Distance (**FRD**), a metric leveraging pre-defined, interpretable *radiomic* features which are widely used in medical image analysis (see *e.g.*, Yip and Aerts (2016), Gillies et al. (2016), Van Griethuysen et al. (2017), Lambin et al. (2012)). In addition to the inherent interpretability of the radiomic features it uses (Cui et al., 2023; Orton et al., 2023; Bang et al., 2021; Ye et al., 2024; Rifi et al., 2023), FRD offers numerous advantages over learned feature metrics like FID and RadFID, which we demonstrate for various applications via an extensive evaluation framework for medical image distribution similarity metrics, summarized in Fig. 2. FRD is an improved version of our previous early version "FRD_{v0}" (Osuala et al., 2024), being both more robust and capturing a larger, more descriptive space of image features (Section 3.1); we further verify the resulting improvements in this work with far more extensive intrinsic and extrinsic evaluations than were present in our preliminary study.

We demonstrate these results in key application areas for unpaired medical image distribution comparison, such as out-of-domain (OOD) detection/analysis and the evaluation of image-to-image translation and image generation models. Our experiments span a wide range of medical image datasets and downstream tasks, image translation and generation models, and perceptual metrics. The datasets (see Table 1) cover broad medical imaging scenarios that possess different image domains, including breast MRI inter-scanner data from different vendors, inter-sequence brain MRI data involving varying sequences, inter-modality data such as lumbar spine and abdominal MRI and CT, and others, all of which present unique challenges for the explored tasks. **We summarize our contributions as follows:**

1. We highlight the shortcomings of common metrics for image distribution comparison (*e.g.*, FID) in meeting the unique requirements of medical imaging.
2. We introduce FRD, a task-independent perceptual metric based on radiomic features, which offers various improvements over prior metrics: **(1)** alignment with downstream tasks, **(2)** stability and computational efficiency for small datasets, **(3)** clinical interpretability, **(4)** sensitivity to image corruptions and adversarial attacks, and **(5)** alignment with radiologist perceptions of image quality.
3. We validate FRD across diverse medical imaging datasets and applications, such as practical out-of-domain detection (including proposing a novel, standardized *dataset-level* OOD metric), image-to-image


Framework for the validation of medical image distribution similarity metrics

Sec 4.2

Alignment with downstream task metrics and anatomical consistency

FRD has strongest average correlation with downstream performance on 7 different segmentation, classification, and detection tasks

Sec 4.1

Out-of-distribution detection ability

FRD shows improved average AUC, accuracy and sensitivity, while maintaining similar specificity compared to other metrics

Sec 5.1

Alignment with radiologist preferences of image quality

FRD correlates substantially better with synthetic image realism assessment by expert readers compared to other metrics

Sec 5.2-3

Computational stability and efficiency

FRD is fast to compute and remains stable even for small sample sizes

Sec 5.4

Sensitivity to performance-affecting image corruptions

FRD sensitivity aligns best with performance sensitivity in comparison

Sec 5.5

Ability to detect adversarial image attacks

FRD can detect practically imperceptible adversarial attacks and their severity, outperforming other metrics particularly in subtle attacks

Sec 6

Enhanced interpretability for clinical use

FRD can provide clear, quantitative answers to feature and distribution interpretability queries

Fig. 2. Our evaluation framework for FRD and other medical image distribution similarity metrics.

translation, image generation, and others, demonstrating its superiority in unpaired medical image distribution comparison.

4. We show that, thanks to various improvements, including more robust feature normalization and the inclusion of many additional frequency space-based features which form a more expressive space of features to describe medical images, FRD outperforms our initially proposed $\mathrm{FRD}_{\mathbf{v0}}$ (as well as other metrics) in essentially all tested scenarios.
5. Finally, we present a general-purpose evaluation framework for medical image similarity metrics, including the first large-scale comparative study of generative models for medical image-to-image translation, and release an accessible codebase to support future research.

We summarize FRD and visualize potential applications for it in Fig. 1. FRD can be readily used and integrated into experiment pipelines via our accessible codebase packaged as a Python library at https://github.com/RichardObi/frd-score. In addition, our evaluation framework for medical image similarity metrics can be utilized at github.com/mazurowski-lab/medical-image-similarity-metrics.

2. Related work

2.1. Metrics for comparing image distributions

The standard approach for comparing two unpaired sets/distributions of images $D_1, D_2 \subset \mathbb{R}^n$ involves defining a distance metric between them that satisfies basic properties (reflexivity, non-negativity, symmetry, and the triangle inequality) (Jayasumana et al., 2024). In image-to-image translation for example, D_2 represents images translated from a source domain to a target domain, and D_1 is a set of real target domain images. For unconditional generation, D_2 contains generated images, and D_1 serves as a real reference set.

Typically, images from D_1 and D_2 are first encoded into a lower-dimensional feature space $F_1, F_2 \subset \mathbb{R}^m$ respectively via an encoder $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$. Then, a distance such as the Fréchet distance (Fréchet, 1957)—technically the 2-Wasserstein distance—is computed between these feature distributions. After assuming that F_1 and F_2 are Gaussian with respective estimated mean vectors μ_1, μ_2 and covariance matrices Σ_1, Σ_2, this distance becomes

$$d_F(F_1, F_2) = \left(||\mu_1 - \mu_2||_2^2 + \mathrm{tr}\left[\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{\frac{1}{2}}\right]\right)^{\frac{1}{2}}. \tag{1}$$

The popular **Fréchet Inception Distance (FID)** metric (Heusel et al., 2017) computes this distance utilizing an ImageNet-pretrained (Deng et al., 2009b) Inception v3 network (Szegedy et al., 2016) as the encoder. Other metrics include **KID (Kernel Inception Distance)** (Bińkowski et al., 2018), which uses Maximum Mean Discrepancy (MMD) and is suited for smaller datasets, and **CMMD (CLIP-MMD)** (Jayasumana et al., 2024), which employs CLIP-extracted (Radford et al., 2021) image features with MMD as an alternative to FID. However, note that all of these distances utilize features learned from natural images.

2.2. Radiology FID (RadFID)

Recent studies suggest that standard perceptual metrics like FID, which are pretrained on natural images, may be unsuitable for medical images (Konz et al., 2024a,b; Chen et al., 2024b; Wu et al., 2025). A straightforward solution to this is given by the usage of features from a model trained on a large "universal" medical image dataset, such as Radiology ImageNet (RadImageNet) (Mei et al., 2026); such an approach, termed RadFID, was introduced in Osuala et al. (2023b) and then further tested for unconditional generative model evaluation in Woodland et al. (2024). However, it has not seen widespread adoption, and this work is the first to explore its use for OOD detection as well as image translation.

2.3. Fréchet radiomic distance (FRD): version 0

We introduced a preliminary version of the Fréchet Radiomic Distance in (Osuala et al., 2024) in the context of evaluating multi-condition latent diffusion models for breast MRI generation, which we label here as "$\mathrm{FRD}_{\mathbf{v0}}$". $\mathrm{FRD}_{\mathbf{v0}}$ is computed by extracting 94 different radiomic features v_{ji} for each image x_i in the given dataset D (D_1 or D_2), and min-max normalizing each type of feature (to $[0, 1]$) given its distribution in the dataset, as

$$v_{ji} \leftarrow \frac{v_{ji} - \min_i v_{ji}}{\max_i v_{ji} - \min_i v_{ji}}. \tag{2}$$

Next, to make $\mathrm{FRD}_{\mathbf{v0}}$ values comparable to FID, these are re-scaled to the common range of ImageNet pre-trained InceptionV3 features used to compute FID, $[0, 7.456]$. Finally, $\mathrm{FRD}_{\mathbf{v0}}$ is computed as the Fréchet distance between these min-max normalized distributions of radiomic features for D_1 and D_2.

FRD makes several improvements to $\mathrm{FRD}_{\mathbf{v0}}$ (Section 3.1), including (1) the addition of many radiomic features that better capture the variation and nuances of different signals (*e.g.*, in frequency space) that are useful to compare imaging distributions; (2) using z-score normalization instead of min-max for better robustness to outliers; and (3) normalizing a given feature type for both D_1 and D_2 with respect to the same reference distribution. Our experiments in using these metrics for a wide range of applications (Section 4) show that these changes result in universal improvements.

2.4. Evaluating deep generative medical image models

We will study two types of generative models in medical imaging: image-to-image translation models and unconditional generative models. Image-to-image translation models, primarily used in radiology to

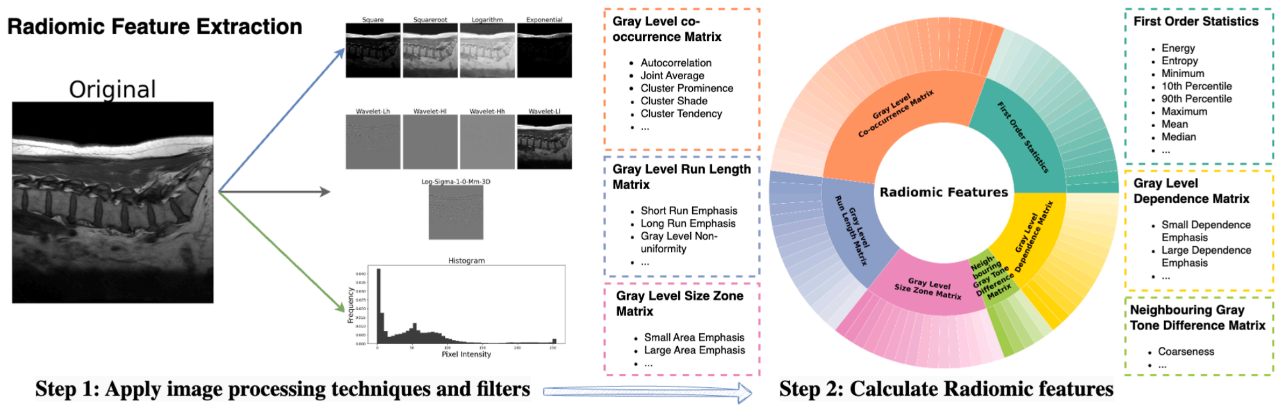


Fig. 3. Our extraction process and taxonomy for radiomic image features. From left to right, the steps of feature extraction include **(1)** the passing of the input image through possible filters, *e.g.*, wavelet/frequency-space conversion, and **(2)** the measurement of various first-order or higher-order features, *e.g.*, entropy and gray-level matrix features, respectively.

mitigate domain shift between datasets (Armanious et al., 2020; McNaughton et al., 2023), have applications such as inter-scanner translation (e.g., across different manufacturers) (Cao et al., 2023; Beizaee et al., 2023), inter-sequence translation (e.g., T1 to T2 MRI) (Durrer et al., 2024; Li et al., 2023b), and inter-modality translation (e.g., MRI to CT) (Yang et al., 2019; Phan et al., 2023; Zhang et al., 2018; Wang et al., 2024b). Our study includes datasets covering each of these scenarios.

Unconditional generative models, which learn to generate synthetic images given unlabeled real images for training, are commonly used to supplement medical image datasets, *e.g.*, for the training of diagnostic models (Kazeminia et al., 2020; Yi et al., 2019; Chen et al., 2022b), including the generation of rare cases (Chen et al., 2021). To our knowledge, no previous work has developed task-independent metrics specifically for medical image generation or translation models. The vast majority of works utilize FID (Heusel et al., 2017) (rather than *e.g.*, RadFID), despite its aforementioned limitations.

2.5. Radiomic features for medical image analysis

Radiomic features, which are typically hand-crafted, have long been used in diverse medical image diagnostic tasks (Van Griethuysen et al., 2017; Yip and Aerts, 2016; Gillies et al., 2016; Lambin et al., 2012), providing a meaningful, interpretable feature space for analyzing medical images. Applications include cancer screening (Jiang et al., 2021), outcome prediction (Aerts et al., 2014; Clark, 2008), treatment response assessment (Drukker et al., 2018; Cha et al., 2017; Li et al., 2016), and many others. A number of radiomic-based clinical tests have even received FDA clearance (Huang et al., 2023). While previous works mainly use learned network features over pre-defined radiomic features for diagnostics (Wagner et al., 2021), few studies have applied radiomics for out-of-domain detection or the evaluation of image translation/generation models, which we show has strong potential.

3. Methods

3.1. Towards a metric designed for medical images: FRD

RadFID is a seemingly suitable alternative to FID, which we will show improves on typically-used perceptual metrics for medical images—such as FID—in various aspects. However, it still lacks clear interpretability of the features being used to compare images (as well as various other limitations which we will demonstrate with upcoming experiments). For medical imaging, especially in image-to-image translation tasks, it is often critical to answer specific questions about how an image's features change or differ from some reference/original image—a need less relevant in natural image translation tasks like style transfer. However, the learned features used in RadFID/FID are difficult to interpret reliably (Section 6).

As a more interpretable alternative, we propose FRD, which utilizes a space of real-valued *radiomic features* of images. The taxonomy and extraction process of these features are illustrated in Fig. 3. They include image-level features such as basic first-order statistics, and textural statistics such as the gray level co-occurrence matrix (Haralick et al., 1973), gray level run length matrix (Galloway, 1975), and gray level size zone matrix (Thibault et al., 2009). Crucially, we improve on $\mathrm{FRD}_{\mathbf{v0}}$ (Osuala et al., 2024) by passing input images through various optional wavelet filters prior to radiomic computation: these filters first apply a spatial Fourier transform to an image, and then apply one of four different possible choices of low- or high-pass filter combinations along the two spatial directions (*e.g.*, low-low, low-high, high-low, and high-high).

This step results in a much more comprehensive space of features with which to compare distributions of images—in particular, the frequency-based features which can capture crucial subtleties in images—which we will show results in improved performance for a wide range of applications (Sections 4 and C.1.1, etc.). In total, this results in $m = 464$ features, for each combination of filter and radiomic feature. All computations are completed via the PyRadiomics library (Van Griethuysen et al., 2017), and radiomics are extracted from the entire, unmasked image. Note that we evaluate the importance of different types of features for FRD in Section C.1.1. Importantly, these features are in compliance with definitions provided by the Imaging Biomarker Standardization Initiative (Zwanenburg et al., 2016, 2020), mitigating past issues of poorly-standardized radiomics in the field (Yip and Aerts, 2016; Kocak et al., 2024).

Each image $x \in \mathbb{R}^n$ is mapped to its radiomic feature representation $f_{\mathrm{radio}}(x) \in \mathbb{R}^m$, and we compute FRD as the Fréchet distance between radiomic feature distributions D_1 and D_2, applying a logarithmic transformation for stability:

$$\mathrm{FRD}(D_1, D_2) := \log d_F(f_{\mathrm{radio}}(D_1), f_{\mathrm{radio}}(D_2)). \tag{3}$$

We also z-score normalize each feature with respect to its distribution in D_1. Note that we tested MMD distance as an alternative to Fréchet, but found it less effective (Section C.1.2).

3.2. Downstream task-based image metrics

FRD and the other perceptual metrics discussed so far compare image distributions in a downstream task-independent manner. However, in medical image analysis, task-dependent metrics are often more common, such as assessing how closely translated/generated images resemble real target images by evaluating them with downstream tasks such

Table 1
Main datasets evaluated in this paper. "Domains" are the source→target domain pairs used, *e.g.*, for image translation. Each dataset is labeled for how its domains are similar ("Intra-") and how they differ ("Inter-"), *e.g.*, the BraTS dataset has intra-modality, inter-sequence domains. "FGT" is fibroglandular tissue.

Abbrev. name	Full name/citation	Domains	Intra-	Inter-	Train/val/test sizes	Downstream tasks
Breast MRI	Duke Breast CancerMRI (DBC)(Saha et al., 2018a),(Lew et al., 2024)	Siemens→GE(T1 MRI)	Sequence	ScannerManuf.	12K/2.4K/2.6K	FGT segmentation,breast segmentation,cancer classification
Brain MRI	BraTS(Menze et al., 2015)	T1→T2	Modality	Sequence	28K/6K/6K	Tumor segmentation,tumor detection,cancer classification
Lumbar spine	TotalSegmentator(Wasserthal et al., 2023)and in-house MRIs	T1 MRI→CT	Body region	Modality	2K/0.6K/0.6K	Bone segmentation
CHAOS	CHAOS(Abdom. MRI & CT)(Kavur et al., 2021)	CT→T1 MRI(in-phase)	Body region	Modality	1.8K/1.1K/0.6K	Liver segmentation,liver classification

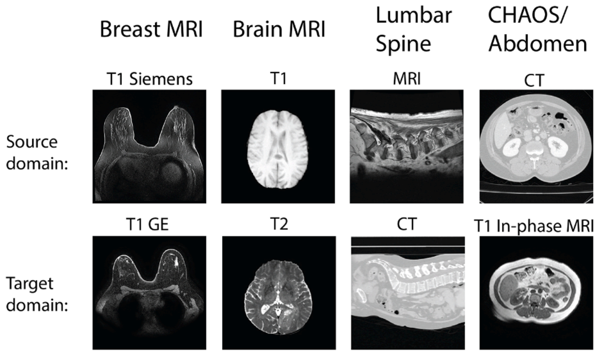


Fig. 4. Example images from each dataset, ordered left-to-right with respect to Table 1.

as semantic segmentation. For example, in image translation, this often involves training a model on real target domain images and testing it on translated images (Vorontsov et al., 2022; Kang et al., 2023), or vice versa (Yang et al., 2019; Chen et al., 2024b). If the task is segmentation, this approach also measures *anatomical consistency* between source and translated images. Such metrics will therefore serve as important targets for the task-independent metrics which we will evaluate.

While segmentation is the primary task of interest, we also assess other downstream tasks including object detection and classification. We denote downstream performance on a dataset D (*e.g.*, a translated image test set $D^{\text{test}}_{s\to t}$) as $\text{Perf}(D) \in \mathbb{R}$, with higher values indicating better performance. Specifically, we use the Dice coefficient for segmentation, mIoU (mean Intersection-over-Union) and mAP@$[0.5, 0.95]$ (mean Average Precision) for object detection, and AUC (area under the receiver operating characteristic curve) for classification—the latter computed on predicted logits to account for test set class imbalance (Mandrekar, 2010; Maier-Hein et al., 2024).

3.3. Datasets and downstream tasks

We evaluate a range of multi-domain medical (radiology) datasets of 2D image slices extracted from 3D volumes for out-of-domain detection, translation, and generation, covering inter-scanner, inter-sequence, and inter-modality cases (ordered from least to most severe visual differences between domains). The datasets include: **(1) breast MRI** (T1-weighted) from Siemens and GE scanners (DBC (Saha et al., 2018a; Lew et al., 2024)); **(2) brain MRI** (T1-weighted and T2-weighted sequences) from BraTS (Menze et al., 2015); **(3) lumbar spine** MRIs and CTs (from TotalSegmentator (Wasserthal et al., 2023) and in-house MRIs); and **(4) abdominal CT and in-phase T1 MRI** from CHAOS (Kavur et al., 2021). Each dataset is split by patient into training, validation, and test sets, which are each sub-split into domains (details in Table 1), resized to 256×256 and normalized to $[0, 1]$. Example images are in Fig. 4.

The lumbar spine and CHAOS datasets pose especially challenging scenarios due to their relatively small size and significant differences in visible features and anatomical structures between their respective domains.

Downstream Task Evaluation. In addition to prior task-independent/perceptual metrics (Section 2) and FRD (Section 3.1), we assess images using auxiliary models trained on downstream tasks (Table 1), as described in Section 3.2. These models, trained on target domain data, are tested on various domains, such as target (D^{test}_t), source (D^{test}_s), source-to-target translations ($D^{\text{test}}_{s\to t}$), or others, depending on the experiment. Full details on model training, architecture, dataset creation, and task labels are provided in Sections B.2 and A.

4. Evaluation and results

We will now demonstrate various applications of FRD, including out-of-domain detection (Section 4.1), image-to-image translation evaluation (Section 4.2), unconditional image generation evaluation (Section 4.3), and abnormality detection (Section 4.4).

4.1. FRD for out-of-domain detection

As discussed in the introduction, a common problem in deep learning for medical image analysis is *domain shift*: where when some diagnostic downstream task model is presented with images that were acquired from a site, sequence, or modality different from the one where its training data originated, there may be a performance drop due to the data being OOD from the training data (AlBadawy et al., 2018; Mårtensson et al., 2020; Guo et al., 2024). In this section, we will show how FRD is overall superior to prior perceptual metrics for detecting when medical images are OOD.

Perceptual metrics like FRD and FID can help detect whether a new image x_{test} is in-distribution (ID) or out-of-distribution (OOD) relative to a reference ID dataset D_{ID} (*e.g.*, some model's training set) without labels. The OOD score $s(x_{\text{test}}; D_{\text{ID}})$ can be defined as the distance (here, we use L_2) of x_{test}'s features from the mean features of D_{ID}:

$$s(x_{\text{test}}; D_{\text{ID}}) = ||f(x_{\text{test}}) - \mathbb{E}_{x_{\text{ID}}\sim D_{\text{ID}}} f(x_{\text{ID}})||_2, \tag{4}$$

where f is an image feature encoder (Reiss and Hoshen, 2023; Schlegl et al., 2019). OOD performance thus depends on the choice of feature space, so we compare radiomic (*i.e.*, FRD or $\text{FRD}_{\mathbf{v0}}$), RadImageNet (*i.e.*, RadFID) and ImageNet (*i.e.*, FID) features for OOD detection.

For each dataset, we use the target domain training set (Table 1) as D_{ID} and compute the OOD score on the ID and OOD images of the test set,[1] aggregating scores via AUC (Fawcett, 2006). OOD detection AUC results (top block of Table 2) and ID vs. OOD detection score distributions (Fig. 5) show that FRD radiomic features outperform learned feature spaces (FID, RadFID) and $\text{FRD}_{\mathbf{v0}}$ radiomic features on average, more clearly separating ID and OOD distributions. This is particularly

---

[1] Note that for datasets which have images from multiple domains of the same patient, *e.g.*, BraTS, we use random sampling to ensure that the domain subsets for a given train/val/test set do not overlap by patient.

Table 2
Using different feature spaces/distance metrics for OOD detection. Best result and runner-up for a given detection metric and dataset are shown in bold and underlined, respectively. 95th percentile thresholding (Eq. (5)) was used to obtain accuracy, sensitivity and specificity results.

Metric	Feature Space/Distance Metric	BreastMRI	Brain MRI	Lumbar	CHAOS	Avg.
AUC	ImageNet	0.43	**0.91**	0.89	0.94	0.79
	RadImageNet	0.35	0.64	0.99	0.99	0.74
	$\mathrm{FRD}_{\mathbf{v0}}$	0.60	0.90	0.78	**1.00**	0.82
	FRD	**1.00**	0.76	**1.00**	**1.00**	**0.94**
Accuracy	ImageNet	0.65	0.73	0.81	0.84	0.76
	RadImageNet	0.68	0.48	**0.98**	0.92	0.77
	$\mathrm{FRD}_{\mathbf{v0}}$	0.74	**0.84**	0.79	**1.0**	0.84
	FRD	**0.96**	0.57	0.92	0.95	**0.85**
Sensitivity	ImageNet	0.03	0.51	0.37	0.83	0.44
	RadImageNet	0.02	0.03	**1.00**	0.88	0.48
	$\mathrm{FRD}_{\mathbf{v0}}$	0.07	0.71	0.19	**1.00**	0.49
	FRD	**1.00**	**0.95**	0.71	**1.00**	**0.92**
Specificity	ImageNet	0.88	0.95	0.96	0.87	0.92
	RadImageNet	0.92	0.94	0.97	**1.00**	0.96
	$\mathrm{FRD}_{\mathbf{v0}}$	**0.99**	**0.97**	**1.00**	**1.00**	**0.99**
	FRD	0.95	0.92	**1.00**	**1.00**	0.85	0.93

true for the challenging case of breast MRI, where the domain shift is visually subtle (as shown in Fig. 4), likely due to the use of frequency-space features in FRD, which can capture subtle visual details. Overall, FRD is sensitive to even these subtle changes *because* it relies on rigid, hand-crafted radiomic features, making it highly sensitive to distributional deviations and particularly suited for OOD detection (as well as detecting image corruptions; see Sections 5.4, 5.5, and C.5).

However, for true practical use, a score threshold $\hat{s}$ would need to be set to binarily classify ID vs. OOD images *without* a validation set of OOD examples, as AUC simply integrates over all possible thresholds. Given the lack of OOD examples, this is doable if we heuristically set $\hat{s}$ as the 95th percentile of the scores for known ID points:

$$\hat{s} = \mathrm{Percentile}_{95}(S_{\mathrm{ID}}), \tag{5}$$

where S_{ID} is the reference distribution of ID scores defined by $S_{\mathrm{ID}} := \{s(x; D_{\mathrm{ID}} \setminus x) : x \in D_{\mathrm{ID}}\}$. We illustrate these computed thresholds for each dataset using different feature spaces in Fig. 5, and show quantitative detection results using them in Table 2. We see that using FRD or $\mathrm{FRD}_{\mathbf{v0}}$ features over ImageNet or RadImageNet results in noticeably improved average accuracy and sensitivity, and on-par specificity, especially for the challenging subtle domain shift case of breast MRI. FRD improves on $\mathrm{FRD}_{\mathbf{v0}}$ noticeably in AUC and sensitivity, and is roughly on-par for accuracy and specificity.

4.1.1. *OOD performance drop prediction*

Another closely related question is *"does FRD detect when performance will drop on new data?"* for some downstream task model. We evaluated this for each of the downstream tasks of Table 1, and we see that in almost all cases, there is a drop in average performance on test data that was detected as OOD using the binary threshold approach of Eq. (5), compared to ID performance (full table in Section C.2). Additionally, in Section C.3 we show that FRD outperforms other metrics in ranking which of different OOD datasets will result in worse downstream task performance.

4.1.2. *Towards practical dataset-level OOD detection*

We also propose a FRD-based metric for *dataset-level* OOD detection, $\mathrm{nFRD}_{\mathrm{group}}$, which is formulated to estimate the probability that some new test set D_{test} is OOD as a whole, relative to D_{ID}. This is particularly designed for the realistic scenario of receiving a new dataset from some outside hospital/site, and wanting an interpretable indication of if the dataset is suitable for some in-domain trained model. While preliminary, we found that this metric scores OOD datasets more consistently than other prior metrics, providing an estimated OOD probability of $\mathrm{nFRD}_{\mathrm{group}} \simeq 1$ for 3 out of 4 datasets which span a variety of modalities and body regions; see Section C.4 for the full details.

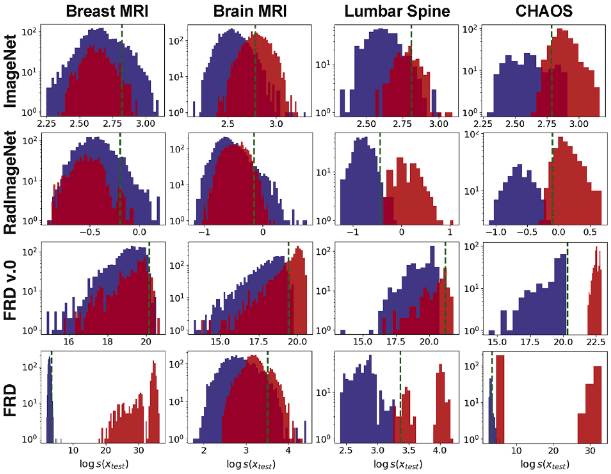


Fig. 5. OOD detection score distributions for in-domain (**blue**) and OOD (**red**) test images for each dataset (columns), using different feature spaces (rows). Computed detection thresholds (Eq. (5)) are shown as vertical dashed **green** lines.

4.1.3. *Summary: practical medical image OOD detection using FRD*

Finally, in the interest of practical usage, we provide a step-by-step guide for OOD detection of medical images using FRD in Algorithm 1.

Algorithm 1 Medical image OOD detection using FRD.

```
Require: Test image set $D_{test}$, reference ID image set $D_{ID}$, radiomic feature encoder $f := f_{radio}$.
1: $S_{ID} := \{s(x_{ID}; D_{ID} \setminus x_{ID}) : x_{ID} \in D_{ID}\}$
2: $\hat{s} = \mathrm{Percentile}_{95}(S_{ID})$
3: $\ell_{test} := \{\mathbf{1}[s(x_{test}; D_{ID}) \geq \hat{s}] : x_{test} \in D_{test}\}$
4: return Binary OOD labels $\ell_{test}$
5: return (Optional) dataset-level OOD score, $nFRD_{group}(D_{test}; D_{ID})$ (Section C.4)
```

Table 3

Perceptual/task-independent metrics $d(D^{\text{test}}_{s\to t}, D^{\text{test}}_{t})$ for image translation models. Best and runner-up models according to each metric in bold and underlined, respectively.

	Breast MRI						Brain MRI					
Method	FRD	FRD_{v0}	RadFID	FID	KID	CMMD	FRD	FRD_{v0}	RadFID	FID	KID	CMMD
CycleGAN	38.1	706	0.26	107	0.049	**0.308**	33.4	554	0.06	21.7	0.004	0.378
MUNIT	43.7	**626**	0.29	144	0.089	1.480	25.7	540	0.05	21.6	0.006	0.388
CUT	24.8	632	**0.17**	106	0.053	0.362	33.8	552	0.13	29.4	0.012	0.259
GcGAN	**24.6**	647	**0.17**	104	0.040	0.322	**12.1**	**523**	**0.04**	**19.0**	**0.003**	**0.239**
MaskGAN	48.7	1042	0.35	118	0.089	0.642	27.8	555	0.06	23.5	0.008	0.392
UNSB	**24.6**	645	0.19	**91**	**0.033**	0.388	**12.1**	525	0.08	26.0	0.010	0.563

	Lumbar						**CHAOS**					
Method	FRD	FRD_{v0}	RadFID	FID	KID	CMMD	FRD	FRD_{v0}	RadFID	FID	KID	CMMD
CycleGAN	6.71	350	0.25	210	0.161	2.950	42.8	470	0.11	**122**	**0.051**	0.379
MUNIT	9.31	367	0.30	**197**	**0.151**	**2.317**	**5.41**	**276**	**0.10**	136	0.073	0.904
CUT	**6.48**	417	**0.21**	245	0.206	3.373	6.84	514	**0.10**	145	0.083	0.444
GcGAN	6.52	**313**	0.25	226	0.161	3.300	6.38	434	0.12	141	0.064	0.507
MaskGAN	6.64	421	0.27	248	0.217	3.237	58.8	437	0.22	212	0.130	2.120
UNSB	6.59	375	0.23	208	0.172	2.579	51.8	542	0.11	135	0.078	**0.356**

4.2. FRD for evaluating image-to-image translation

4.2.1. Image-to-image translation models

Unpaired image-to-image translation for medical images, lacking paired data, is challenging and typically relies on adversarial learning. We evaluate a variety of state-of-the-art unpaired models: CycleGAN (Zhu et al., 2017), MUNIT (Huang et al., 2018), CUT (Park et al., 2020), GcGAN (Fu et al., 2019), MaskGAN (Phan et al., 2023), and UNSB (Kim et al., 2024), each representing diverse techniques such as contrastive learning and style/content disentanglement. All models are trained on source and target domain images from each dataset, with detailed training specifics in Section B.1.

4.2.2. Evaluation with perceptual metrics

We first evaluate each translation model using perceptual metrics to measure the distance between translated test set source domain images and real test set target domain images. We compare FRD to FRD_{v0}, RadFID, FID, KID, and CMMD for this task, with results in Table 3. We first qualitatively observe that FID often fails to capture visual quality and anatomical consistency, particularly when there is a high semantic shift between source and target domains, as shown in Fig. 6. For example, FID, KID and CMMD rate MUNIT as best for lumbar spine despite a clear loss of bone structure—shown by MUNIT being the *worst* by segmentation performance in Table 4, which FRD and RadFID capture successfully. This highlights certain limitations of using prior perceptual metrics for medical images.

4.2.3. Correlation with downstream task performance and anatomical consistency

Since a key goal of medical image translation is maintaining downstream task performance (*e.g.*, segmentation) and mitigating domain shift, we will now examine whether perceptual metrics can serve as proxies for task performance by correlating perceptual distances with downstream task metrics. We calculate the Pearson correlation r between each perceptual metric and downstream performance across all translation models (Tables 3 and 4).

As shown in Fig. 7, FRD has the strongest (most negative) average correlation with downstream task performance ($r = -0.43$), followed by RadFID ($r = -0.36$), while FRD_{v0}, FID, KID, and CMMD are less consistent ($r = -0.01$, $r = -0.17$, $r = -0.17$, $r = -0.08$, respectively), especially for datasets with larger domain shifts. This is particularly the case for segmentation tasks (which measure anatomical consistency), where—excluding CHAOS, which had low correlations likely due to it generally being a difficult dataset (see Section 3.3)—FRD and RadFID achieved mean correlations of $r = -0.58$ and $r = -0.70$, while FRD_{v0}, FID, KID and CMMD have $r = -0.04$, $r = -0.34$, $r = -0.42$, and $r = -0.08$, respectively.

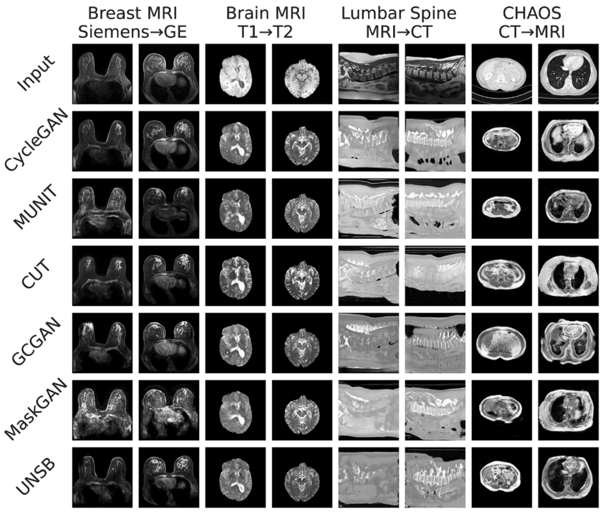


Fig. 6. Translations $x^{\text{test}}_{s\to t}$ from each translation model (non-top rows) given example inputs x^{test}_{s} (top row).

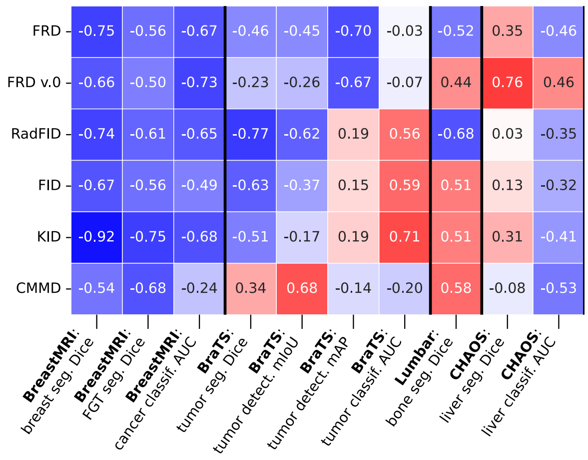


Fig. 7. Pearson correlation of perceptual metrics (vertical axis) (Table 3) with downstream task-based metrics (horizontal axis) (Table 4) for evaluating image translation, taken across all translation models **(lower r (colder color) is better)**.

Table 4
Downstream task performance metrics $\mathrm{Perf}(D^{\text{test}}_{s\to t})$ for image translation models. In-domain and out-of-domain performance shown at the bottom for reference for how susceptible each task is to domain shift, and as *expected* upper and lower performance bounds.

	Breast MRI			Brain MRI				Lumbar	CHAOS	
Method	Dice Breast	FGT	AUC Cancer	Dice Tumor	mIoU Tumor	mAP	AUC Cancer	Dice Bone	Dice Liver	AUC Liver
CycleGAN	0.871	**0.494**	0.530	0.348	0.164	0.126	0.805	0.232	0.284	0.591
MUNIT	0.832	0.201	0.511	0.337	0.168	0.125	0.844	0.101	0.182	0.323
CUT	0.843	0.373	0.544	0.303	0.159	0.133	**0.861**	**0.277**	**0.444**	**0.744**
GcGAN	0.876	0.389	0.492	0.360	0.165	**0.137**	0.835	0.126	0.167	0.702
MaskGAN	0.809	0.164	0.441	**0.375**	**0.170**	0.126	0.842	0.167	0.317	0.375
UNSB	**0.881**	0.308	**0.594**	0.353	0.169	0.135	0.839	0.138	0.381	0.405
In-domain	0.883	0.696	0.670	0.442	0.174	0.169	0.841	0.949	0.864	0.866
OOD	0.747	0.446	0.538	0.005	0.152	0.065	0.727	0.007	0.062	0.504

	Breast MRI		**Brain MRI**		**Lumbar CT**		**Abdom. CT**	
Model:	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
FID	67	54	157	26	210	103	232	118
KID	0.02	0.01	0.14	0.01	0.18	0.04	0.23	0.08
CMMD	0.30	0.33	0.82	0.28	3.57	0.87	4.72	0.48
RadFID	0.16	0.06	0.16	0.04	0.20	0.11	0.40	0.16
FRD_{v0}	585	568	709	633	232	167	897	710
FRD	47.4	43.4	58.5	56.3	8.3	6.3	71.3	68.6

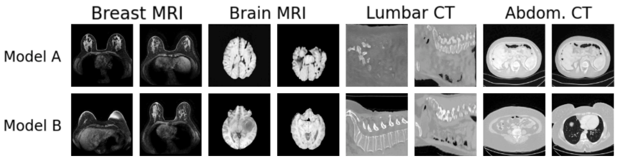


Fig. 8. Top: FRD and other perceptual metrics for evaluating unconditional generative models, comparing a poor model (**A**) to a better model (**B**). **Bottom:** example generated images.

These results align with recent work which showed that metrics like FID and KID, despite being popular in mainstream computer vision, do not reliably correlate with downstream task performance (Konz et al., 2024b; Wu et al., 2025). Our results also potentially indicate generally best translation models for medical images, which we will discuss in Section 7.

4.3. *FRD for evaluating unconditional image generation*

We have so far focused our experimental effort on image-to-image translation over image generation due to the direct relationship of it with the key problem of domain shift in medical imaging. We will now study FRD in evaluating unconditional generative medical image models, similar to FID's typical use. We trained StyleGAN2-ADA (Karras et al., 2020) with default settings on four single-domain image generation tasks: **(1)** GE T1 breast MRI, **(2)** T1 brain MRI (BraTS), **(3)** lumbar spine CT, and **(4)** an abdominal CT dataset (CT-Organ (Rister et al., 2020))[2].

We evaluate each perceptual metric (FRD, FRD_{v0}, RadFID, FID, CMMD, KID) by ranking samples from an early model iteration (**Model A**, 2×10^5 images seen in training) of visibly lower quality against a fully trained model (**Model B**, 10× more images seen in training)—similar to Jayasumana et al. (2024), with results and sample generated images shown in Fig. 8. FRD successfully identifies the lower-quality model in all cases, aligning with prior metrics, except for CMMD, which fails for breast MRI.

4.4. *FRD as a predictor of abnormality*

In this section, we will demonstrate a basic example of another use of FRD: as a predictor of abnormalities within medical images. We consider a dataset of axial breast DCE-MRI images cropped through the middle to only include one breast. We consider using FRD to predict whether a given test breast image is healthy or unhealthy in two scenarios: **(1)** where only a reference set of healthy breast images D_{healthy} is available (an anomaly detection/out-of-distribution detection scenario), and **(2)**, a somewhat easier case where a reference set of unhealthy breasts D_{cancer} can also be used. The test set and reference sets are sampled from the MAMA-MIA (Garrucho et al., 2025) test and train sets, respectively, with full dataset creation and pre-processing details provided in Section A.5.

In the first scenario, for each test image x, we measure the FRD of x from the reference healthy set via $s(x; D_{\text{healthy}})$ (Eq. (4)), and use this as the unhealthy prediction score for x. We then aggregate all results via the AUC, computed with the $s(x; D_{\text{healthy}})$ of each test image x and its true label. From this, we obtained an AUC of 0.950.

In the second scenario, a reference set of unhealthy examples D_{cancer} is available in addition to D_{healthy}, allowing us to simply classify some x according to which of the reference sets it is closest to. More precisely, we predict a binary unhealthiness label $\hat{y}(x)$ for x via

$$\hat{y}(x) = \begin{cases} 0 \text{ if } s(x; D_{\text{healthy}}) < s(x; D_{\text{cancer}}) \\ 1 \text{ else} \end{cases} . \tag{6}$$

We then aggregate the results over the entire test set via the AUC given the predicted labels and the true labels, resulting in an AUC of 0.989.

Both of these experiments indicate that the features captured by FRD are highly discriminative for this task, which points to the meaningfulness of the features for medical imaging domains and respective diagnostic applications. Understandably, the task was slightly easier in the second scenario, due to the availability of unhealthy cases for direct comparison.

5. Properties of FRD

In the following sections, we will demonstrate various intrinsic properties of FRD, including its relationship with user (radiologist) preference for image quality (Section 5.1), computational stability (Section 5.2), computational efficiency (Section 5.3), and sensitivity to image corruptions (Section 5.4) and adversarial attacks (Section 5.5).

5.1. *Relationship of FRD to user preference of image quality*

In this section, we investigate whether FRD (and other metrics) correlate with human expert-perceived quality of synthetically-generated images, in the context of breast cancer screening. We utilized the experimental design of Garrucho et al. (2023), where three experienced readers were asked to rate the realism of sets of high-quality synthetic mammography images generated by a CycleGAN (Zhu et al., 2017) trained

---

[2] CHAOS was not large enough to train on for high generation quality.

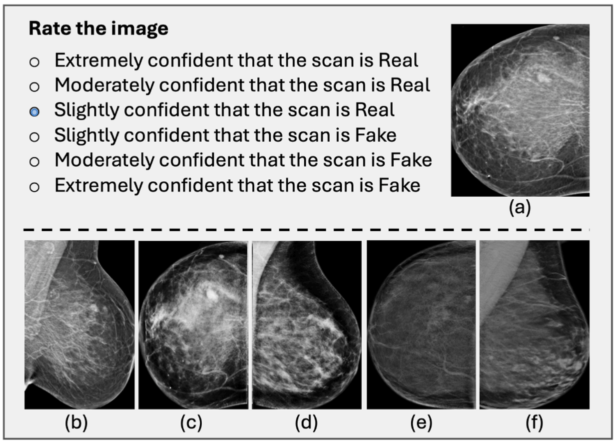


Fig. 9. Reader Study: Using an ImageJ plugin, each reader was shown one image at a time, which was randomly sampled from a set of 90 synthetic mammograms. The image set was equally distributed between craniocaudal (CC) and mediolateral oblique (MLO) views (45 each), as well as between the Optimam, CSAW, and BCDR datasets (30 each). Images (a) to (f) are synthetic samples that looked realistic to the readers: (a) BCDR CC, (b) BCDR MLO, (c) Optimam CC, (d) Optimam MLO, (e) CSAW CC, and (f) CSAW MLO. Figure based on Garrucho et al. (2023).

on images from a specific acquisition site and anatomical view. This was conducted for six (anatomical view, acquisition site) settings, where the anatomical view is CC or MLO, and the dataset is OPTIMAM (Halling-Brown et al., 2020), CSAW (Dembrower et al., 2020), or BCDR (Lopez et al., 2012). These three readers were two breast radiologists with over 9 and 11 years of experience, respectively, and a surgical oncologist with over 14 years of experience in image-guided breast biopsy.

As illustrated in Fig. 9, ratings were conducted on a Likert scale, where the user had to rate a given generated image on a scale of 1 (extremely confident in the scan appearing fake) to 6 (extremely confident in it appearing real), where 1, 2, 3, 4, 5, and 6 are respectively mapped to equally-distributed probabilities (final ratings) of 0.05, 0.23, 0.41, 0.59, 0.77, and 0.95 to compute the ROC curve of each reader as in Alyafi et al. (2020). In total, the study was completed on 90 synthetic images, containing 15 CC and 15 MLO images from each of the three mammography datasets.

Since this study essentially asked readers to visually compare the synthetic images to their known mental reference for how real mammography images generally should appear, we can assess whether FRD and the other perceptual metrics can exhibit the same behavior. To this end, we compute the distance between (a) the set of synthetic images for a given dataset and view and (b) a fixed reference set of 90 real mammography images sampled evenly from all views and datasets[3] Shown in Fig. 10, we analyzed this for each of FRD, $\text{FRD}_{\mathbf{v0}}$, FID, and RadFID, and measured the correlation (Pearson's linear correlation r, as well as Spearman's non-linear/rank r and Kendall's τ) between the metric and user rating, where each datapoint corresponds to each possible (view, dataset) combination. User ratings are averaged over all three readers and all 15 synthetic images for the given view and dataset. Additionally, we explore using user ratings *calibrated* by ratings of real data in Section C.7.

The desired behavior for a perceptual distance metric is for a metric to negatively correlate with user preference, as higher-quality/more realistic synthetic images—as measured by higher user rating—should correspond to the synthetic images having lower perceptual distance with respect to real images. Shown in Fig. 10, we see that out of all metrics, on average over all three correlation measures, FRD performs best in this regard, similarly to $\text{FRD}_{\mathbf{v0}}$ in linear correlation ($r \simeq -0.65$) and out-performing noticeably in non-linear correlation.

Absolute Rating

Metric	Pearson r	Spearman r	Kendall τ
FRD	-0.65	-0.71	-0.6
$\text{FRD}_{\mathbf{v0}}$	-0.66	-0.43	-0.20
FID	0.37	0.66	0.47
RadFID	0.91	0.94	0.87

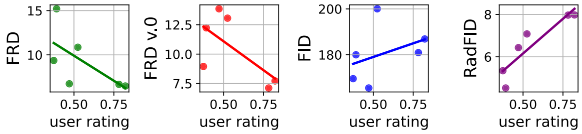


Fig. 10. Top: Correlation coefficients (linear Pearson r_P and non-linear/rank Spearman r_S and Kendall τ) of different distance metrics with average user (radiologist) preference, for the task of measuring synthetic image quality. **Bottom:** Associated plots and linear best fits for this data.

On the other hand, FID and RadFID actually *anti-correlate* with user perception of quality of synthetic images, particularly RadFID; FID is surprisingly less worse in this regard (aligning with recent similar findings in Woodland et al. (2024)), despite RadFID utilizing pretrained domain-specific medical image features, while FID uses natural image features. This surprising result for RadFID could be due to its underlying Inception encoder being trained to focus on very specific patterns to detect (often localized) disease in the pretraining images and tasks of RadImageNet (Mei et al., 2026). While such local patterns may have high influence on RadFID, they may have relatively small influence on the overall image appearance and therefore the readers' ratings of overall image quality, hence the seeming "mismatch" between RadFID and the readers' ratings. On the contrary, FRD captures many generic, global image features (that didn't rely on some training set) that will likely end up overlapping with reader perception in some aspect, which may be why FRD (negatively) correlates more predictably with reader preference. Overall, this provides further evidence that FID and RadFID should be used with caution for medical images, and that FRD provides a noticeable alternative that correlates substantially better with the perceptual preference of experienced readers.

5.2. *Sample efficiency and stability*

The stability of perceptual metrics at small sample sizes is key for medical image datasets due to them being typically smaller (*e.g.*, $N \approx 10^2$–10^4) than natural image datasets (*e.g.*, ImageNet with $N \approx 10^6$). FID generally requires $N \approx 10^4$–10^5 samples for stability (Heusel et al., 2017; Jayasumana et al., 2024), which can be prohibitive in this setting. We will now evaluate FRD, $\text{FRD}_{\mathbf{v0}}$, RadFID, and FID across varying sample sizes N to test for this stability.

We test this under the case of image translation evaluation for the main datasets (Table 1): CycleGAN for breast MRI, GcGAN for brain MRI, CUT for lumbar spine, and MUNIT for CHAOS, respectively. Shown in Fig. 11 left, FRD remains stable even for very small N (down to $N = 10$), while RadFID and FID—as well as $\text{FRD}_{\mathbf{v0}}$, to a lesser extent—diverge as N grows small across all datasets, indicating that these FID-based metrics are not suitable for comparing small medical datasets of different sizes.

The stability of FRD at small N, especially compared to that of FID and RadFID, is likely due to relatively few features dominating its computation (see *e.g.*, Fig. 15(a)), meaning the effective dimensionality $\tilde{m}$ is much smaller than the full $m \approx 500$. Thus, the Fréchet distance in FRD behaves as though it operates in a lower-dimensional space, enhancing stability even with limited samples.

---

[3] Note that it is crucial for the reference set to remain constant for these metrics to have fixed scales.

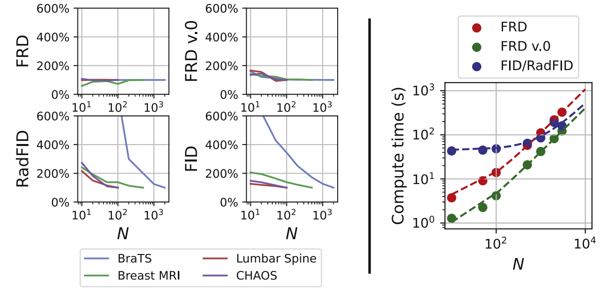


Fig. 11. Left: Sensitivity of FRD, $\mathrm{FRD}_{\mathbf{v0}}$, RadFID and FID to sample size N. Metric values (vert. axes) are relative to their highest-N result. **Right:** Computation time for the metrics w.r.t N, with linear best fits plotted in dashed lines.

More concretely, consider that computing the Fréchet distance (Eq. (1)) as is necessary for FRD and FID requires estimating the covariances matrices $\Sigma_1, \Sigma_2 \in \mathbb{R}^{m\times m}$ of the two data distributions in the m-dimensional feature space ($m \approx 500$ for FRD while $m \approx 2000$ for FID). For $N \ll m$, the estimation of these matrices can become very unstable, *assuming that* most of the m feature dimensions contribute similarly to the variability of the data distributions, *i.e.*, $\tilde{m} \simeq m$.

The computational stability results shown in Fig. 11 indicate $\tilde{m}$ to be typically higher for FID/RadFID features compared to FRD features, as the former are noticeably less stable for low N. As FRD features appear to have relatively low $\tilde{m}$, the true covariance matrices Σ in this feature space are approximately low rank, so that the estimation is stable for small N because they can be mostly represented by the first $\tilde{m}$ terms of their eigendecompositions, *e.g.*, $\mathrm{tr}(\Sigma) \simeq \sum_{i=1}^{\tilde{m}} \lambda_i$ given eigenvalues λ_i of Σ.

5.3. Computation time

We next compare the computation time of FRD to FID/RadFID (as well as $\mathrm{FRD}_{\mathbf{v0}}$) across sample sizes N, using data parallelism (`num_workers=8`) on UNSB-translated BraTS test images. As shown in Fig. 11 right, FRD is faster than FID/RadFID for small-to-moderate sample sizes ($N \lesssim 500$). For larger N, the computation time of FRD grows slightly faster, but both metrics remain efficient across different sample sizes (the computation time scales linearly with N asymptotically), with FRD and $\mathrm{FRD}_{\mathbf{v0}}$ particularly advantageous for small N. While FRD is slightly slower than $\mathrm{FRD}_{\mathbf{v0}}$—owing to the addition of various image filters prior to radiomic computation (Section 3.1)—the improved performance in essentially all tested applications (Section 4) make it well worth it.

5.4. Sensitivity to image corruptions

In this section, we analyze the sensitivity of FRD to image corruptions that may affect downstream task performance, compared to prior common metrics. Given the importance of downstream task performance metrics for medical image translation models, as well as the typical increased sensitivity of medical image models to image corruptions compared to natural image models (Konz and Mazurowski, 2024a), we study if a corruption to a (translated) image that noticeably affects downstream task performance on that image is also captured by the perceptual metrics. For further results, see Section C.5 where we show that FRD and FRD are sensitive to various realistic image corruptions in MRI.

Consider some image transformation/corruption $T : \mathbb{R}^n \to \mathbb{R}^n$. We model a preferable perceptual metric d as approximately following the inverse proportionality

$$\frac{d(D_t, T(D_{s\to t}))}{d(D_t, D_{s\to t})} \underset{\sim}{\propto} \left(\frac{\mathrm{Perf}(T(D_{s\to t}))}{\mathrm{Perf}(D_{s\to t})}\right)^{-1}, \tag{7}$$

evaluated on the test set's target domain images D_t and translated source-to-target images $D_{s\to t}$. In other words, if the perceptual distance increases by some positive multiplicative factor K (implying the corruption made the translated images more distant from the target domain), we would expect the performance to go worse by $1/K$, up to a constant of proportionality—the sensitivity of the two metrics to image corruptions should match.

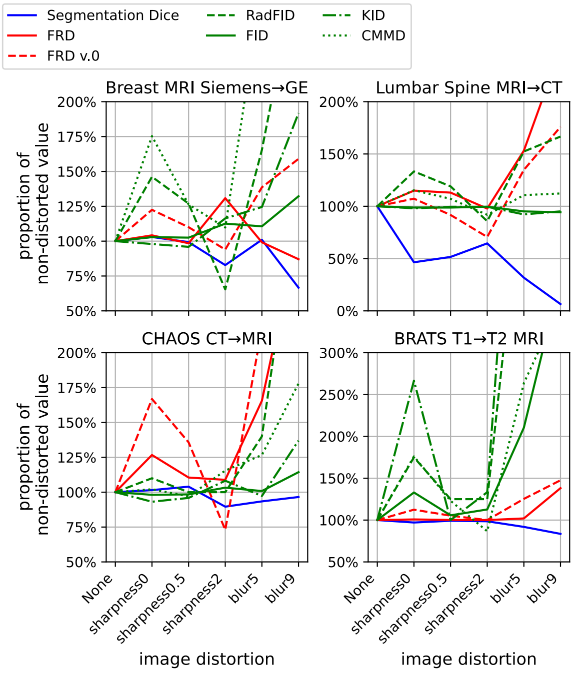


Fig. 12. Sensitivity of FRD, $\mathrm{FRD}_{\mathbf{v0}}$ (**red lines**), RadFID, FID, KID, and CMMD (**green lines**) to corruptions which affect downstream task performance (**blue line**) on translated images. Metric values (vertical axis) are relative to their undistorted result ("None").

We tested this on all datasets for downstream segmentation tasks that were sensitive to such corruptions (fibroglandular tissue (FGT) for breast MRI, tumor for brain MRI, bone for lumbar spine, and liver for CHAOS (Table 4)), for various translation models (CycleGAN, GcGAN, CUT, and MUNIT, respectively). We evaluate simple image corruptions T as Gaussian blurs with positive integer kernel k (labeled as `blur`k) or sharpness adjustments of non-negative factor γ (`sharpness`γ), from TorchVision (maintainers and contributors, 2016). We show the results in Fig. 12, plotting $d(D_t, T(D_{s\to t}))/d(D_t, D_{s\to t})$ and $\mathrm{Perf}(T(D_{s\to t}))/\mathrm{Perf}(D_{s\to t})$ for each corruption for all proposed and prior perceptual metrics d.

For breast MRI, we see that the corruption sensitivity of FID and FRD fairly well match the performance sensitivity, with the other perceptual metrics less so (for example, FRD and RadFID can be oversensitive to corruptions which barely affect performance). For lumbar spine, FRD, $\mathrm{FRD}_{\mathbf{v0}}$ and RadFID follow the performance sensitivity well, while other perceptual metrics are generally not as sensitive. For BraTS, performance is typically not sensitive to corruptions, which FRD and $\mathrm{FRD}_{\mathbf{v0}}$ follow; other perceptual metrics are generally oversensitive, but all still increase when performance decreases due to blurring. CHAOS is an interesting case, where all perceptual metrics except for FID are overly sensitive to corruptions despite only small changes to performance, likely due in part to challenges of training the downstream task model on such a small and challenging dataset (Section 3.3). Overall, the sensitivity of FRD aligns best with performance sensitivity. Other perceptual metrics are not as consistent over all datasets, which aligns with the results of Section 4.2.3.

5.5. Sensitivity to targeted adversarial attacks

Another type of image corruption that is important to consider is adversarial examples, a type of adversarial attack on some trained downstream task model f where an input image x is modified in a targeted manner to drastically change the model's prediction $f(x)$ to be incorrect (Goodfellow et al., 2015), while constraining the modifications to x to be as subtle as possible. This scenario is important to consider due to the safety-critical nature of medical image diagnosis applications. Here, we will consider the case of attacking a binary classification neural network to judge whether FRD (and other distance metrics) can detect when images have been attacked.

We use the FGSM method (Goodfellow et al., 2015) to attack some input image x; FGSM uses the gradient of the network's prediction loss to modify the input image x with true binary domain label y to some $\tilde{x}$ in order for f to misclassify x, as

$$\tilde{x} = \mathrm{FGSM}(x, y) := x + \epsilon \,\mathrm{sign}\left(\frac{\partial L(f(x), y)}{\partial x}\right), \tag{8}$$

where L is the (binary cross-entropy) loss between the model's prediction $f(x)$ and the true label y. The use of the real-valued parameter $\epsilon > 0$ and sign function constrains the attacked image to be imperceptible with a tolerance of ϵ, *i.e.*, $||\tilde{x} - x||_\infty < \epsilon$.

Given some test dataset D of images x and corresponding labels y, we will attack each image in D to obtain the attacked set $\tilde{D}_\epsilon := \{\tilde{x} = \mathrm{FGSM}(x, y) : (x, y) \sim D\}$, using various ϵ (note that $\epsilon = 0$ is the baseline case of non-attacked images, *i.e.*, $\tilde{D}_0 = D$). We will then determine if various distance metrics d, including FRD, can detect the attacked images compared to a separate reference set of clean images (the training set for f) D_{ref}, via $d(D_{\mathrm{ref}}, \tilde{D}_\epsilon)$. We will measure two desiderata: **(a)** if the metrics can differentiate between clean and attacked images—*i.e.*, if $d(D_{\mathrm{ref}}, D) < d(D_{\mathrm{ref}}, \tilde{D}_\epsilon)$ for various ϵ—and **(b)** if the distances $d(D_{\mathrm{ref}}, \tilde{D}_\epsilon)$ increase as the attack becomes more severe (ϵ grows larger).

We use a subset of the experimental setup of Konz and Mazurowski (2024a) (see that paper for more details), performing attacks on binary classification models for seven medical image datasets. These are (1) brain MRI glioma detection (**BraTS**, Menze et al. (2015)); (2) breast MRI cancer detection (**DBC**, Saha et al. (2018b)); (3) prostate MRI cancer risk scoring (**Prostate MRI**, Sonn et al. (2013)); (4) brain CT hemorrhage detection (**RSNA-IH-CT**, Flanders et al. (2020)); (5) chest X-ray pleural effusion detection (**CheXpert**, Irvin et al. (2019)); (6) musculoskeletal X-ray abnormality detection (**MURA**, Rajpurkar et al. (2018)); and (7) knee X-ray osteoarthritis detection (**OAI**, Tiulpin et al. (2018)). We evaluate f as being either a ResNet-18 (He et al., 2016) or a VGG-13 (Simonyan and Zisserman, 2015), and train them with training sets (D_{ref}) of size 1750, before evaluating and attacking them on class-balanced test sets of size 750, via the dataset creation/sampling procedures described in Konz and Mazurowski (2024a). As shown in full in Section C.6, these attacks are typically quite successful, even for very small values of ϵ, despite them being practically undetectable visually.

We show the results of these experiments in Fig. 13, where we show how $d(D_{\mathrm{ref}}, \tilde{D}_\epsilon)$ changes with respect to attack strength ϵ (including $\epsilon = 0$ for the unattacked case of $\tilde{D}_\epsilon = D$), for various distance metrics d, on all datasets and models. We see that FRD clearly differentiates attacked images ($\epsilon > 0$) from non-attacked images ($\epsilon = 0$) in almost all cases (desiderata (a)), and that it typically increases with higher attack strength/ϵ (desiderata (b)). Moreoever, FRD is typically more sensitive to the most subtle attacks ($\epsilon = 1/255$) compared to FRD and almost all other metrics, possibly due to the inclusion of frequency/wavelet features. These results show that FRD can detect adversarial attacks and their severity, despite the fact that these attacks are practically imperceptible.

6. FRD for interpretability

In this section, we will demonstrate how FRD aids in interpreting differences between large sets of medical images, *i.e.*, understanding the main features that differ between the two sets. The example we will study is interpreting the effects of image-to-image translation models, but this formalism could be applied to any two distributions of images.

At the single-image level, an input image x_s and output translated image $x_{s\to t}$ can be converted to feature representations (radiomic or learned) $h_s := f(x_s)$ and $h_{s\to t} := f(x_{s\to t})$, and we can interpret the feature change vector $\Delta h := h_{s\to t} - h_s$ and it's *absolute change* counterpart $|\Delta h|$ defined element-wise by $|\Delta h|^i := |h^i_{s\to t} - h^i_s|$. At the image distribution level, we can define $\Delta h := \mu_{s\to t} - \mu_s$ (and similarly $|\Delta h|$ via $|\Delta h|^i := |\mu^i_{s\to t} - \mu^i_s|$), where $\mu_s, \mu_{s\to t} \in \mathbb{R}^m$ are the mean vectors of the input and output feature distributions, respectively. In this case, we also define the distributions of values for *individual features* as $F^i_s := \{h^i_s : h_s \in F_s\}$ and $F^i_{s\to t} := \{h^i_{s\to t} : h_{s\to t} \in F_{s\to t}\}$.

In either case, Δh is simply the linear direction vector in feature space between the input and output distributions, analogous with other interpretability works that utilize the linear representation hypothesis (Park et al., 2024; Kim et al., 2018; Alain and Bengio, 2017; Konz et al., 2023). We will next discuss the options and challenges for interpreting Δh, for either learned features or fixed (radiomic) features.

Attempting Interpretability with Learned vs. Radiomic Features. A common method for interpreting directions v in a deep encoder's feature space, such as Δh, is *feature inversion* (Olah et al., 2017; Mahendran and Vedaldi, 2015), which uses gradient-based optimization to find an input image x_v that aligns with v in feature space, *i.e.*,

$$x_v = \mathrm{argmax}_x \mathrm{cossim}(v, f(x)). \tag{9}$$

However, we found that doing so using either ImageNet or RadImageNet features resulted in abstract visualizations that lack clear, quantitative insights useful for clinical interpretation (Fig. 14; see Section C.8 for details).

Alternatively, the *individual* features of Δh could be examined statistically with questions like "*which features changed the most?*" or "*did only a few features account for most of the cumulative change?*". However, concretely interpreting individual *learned* features remains challenging due to the qualitative nature of feature inversion, so we face the same problem.

Thankfully, the clear definitions of radiomic features (Section 3.1) allow for clear, quantitative answers to feature interpretability questions, beyond what is possible for learned feature techniques like feature inversion. Here we will exemplify this by interpreting a CUT model trained for lumbar translation, with the following questions.

1. **Which features changed the most?** Sorting features by their values in the $|\Delta h|$ between input and output image distributions (Fig. 15(a)) identifies those with the highest change, primarily textural/gray-level matrix features, reflecting appearance shifts from MRI to CT (Fig. 15(b)).
2. **Did only a few features change significantly?** Yes—50% of cumulative feature changes (measured by $|\Delta h|$) are covered by only 37 out of 500 features, indicating a light-tailed distribution (Fig. 15(a)).
3. **Which images changed the most or least?** Sorting input/output image pairs $(x_s, x_{s\to t})$ by their absolute feature change $||h_{s\to t} - h_s||_2 = ||\Delta h||_2$ (Fig. 15 (c)) shows that the most-changed images have distinct anatomical differences, while the least-changed images mainly differ in texture and intensity (Fig. 15(d)).

This interpretability methodology could also help compare translation models on the same dataset and assess model effects on the images, or analyze the domain shift between two datasets.

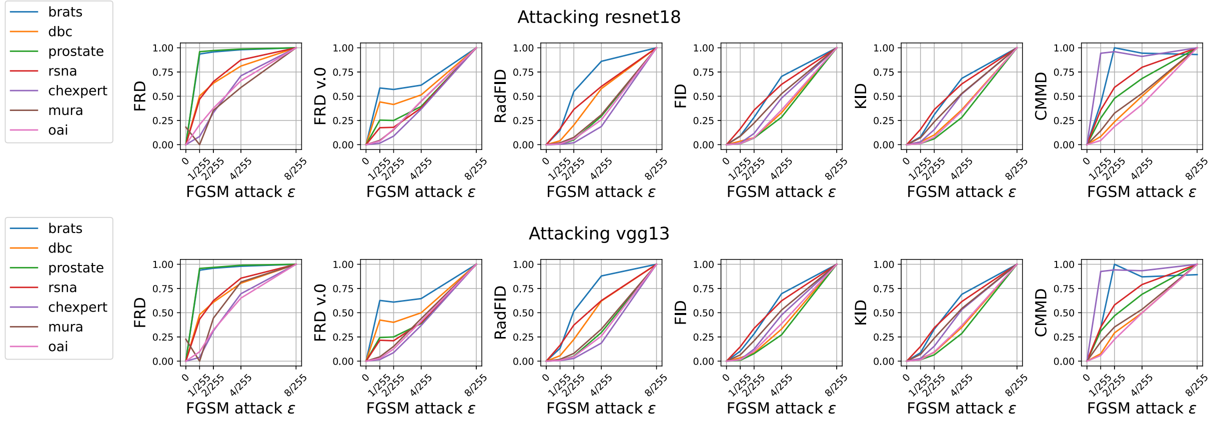


Fig. 13. Using different distance metrics (vertical axes) to detect FGSM adversarial attacks of various strength (horizontal axes), for ResNet-18 (top) and VGG-13 (bottom) classification models. Note that distance metrics are scaled to $[0, 1]$ over their plotted range of values for the sake of visualization.

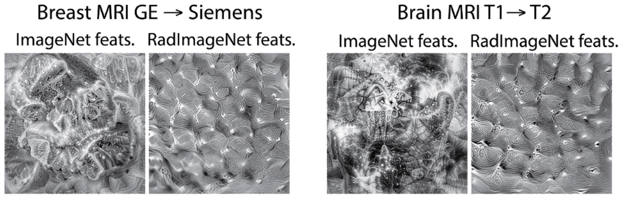


Fig. 14. Attempts at medical image translation interpretability via learned feature inversion.

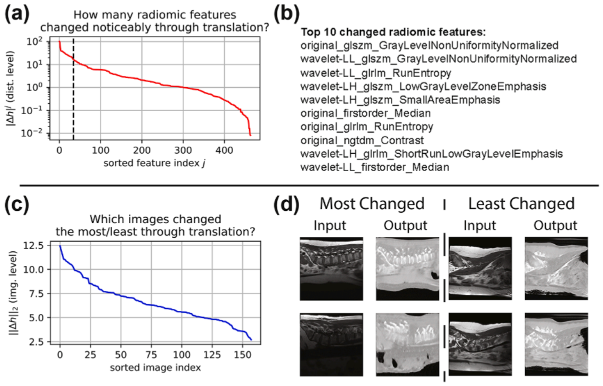


Fig. 15. Translation interpretability using radiomic features.

7. Discussion and conclusions

Overall, our results with FRD show the value of using interpretable, medical image-specialized feature spaces like radiomic features for comparing unpaired medical image distributions. We showed that FRD consistently and substantially improves over FRD, FID, RadFID, as well as other prior common metrics for unpaired medical image comparison on all evaluated tasks. These include stronger alignment with downstream task metrics and anatomical consistency (Section 4.2.3) for the task of evaluating image-to-image translation, OOD detection ability (Section 4.1), alignment with radiologist preferences of image quality (Section 5.1), computational stability and efficiency, especially for small sample sizes (Sections 5.2 and 5.3), sensitivity to performance-affecting image corruptions (Section 5.4) and adversarial image attacks (Section 5.5), as well as, notably, enhanced interpretability for clinical use (Section 6).

Medical Image Similarity from the Perspective of Radiologists. It is also important to consider the task of measuring the similarity of medical images from the perspective of radiologists. To explore this, we performed a semi-structured expert interview of two experienced radiologists (authors L.G. and J.L.). According to them, distinguishing between images from clearly different acquisition types, such as T1- vs. T2-weighted brain MRI or between modalities like abdominal/lumbar spine MRI and CT scans (Fig. 4), is a straightforward task. In contrast, differentiating between more subtly distinct domains, such as breast MRI scans acquired using the same sequence type but different scanner manufacturers or settings (Fig. 4), is considerably more challenging. Nonetheless, identifying such differences is crucial, as even subtle domain shifts can significantly impact the performance of downstream tasks (Table 4, bottom two rows), despite the almost imperceptible visual differences between domains.

Notably, FRD is able to detect these subtle domain differences with high consistency (see the OOD detection results of Table 2, first column). If a radiologist was to use such a downstream task model in clinical practice, FRD could help inform whether the model is well-calibrated for some new image at hand by detecting if the new image is OOD of the model's training set. More generally, this could be used to detect any gradual distribution shift of newly acquired data over time compared to the model's original training set, indicating if the model may require fine-tuning on the new data.

In general, we have found that FRD is sensitive to variations in both scanner protocol and sequence type in a way that follows intuition. For example, FRD detects the subtle differences between GE-acquired and Siemens-acquired T1-weighted images in the breast MRI dataset (Table 1), as FRD between the GE training set and the GE test set is 8.561, while FRD between the former and the Siemens test set is 60.54, noticeably higher. We also found that FRD captures how intra-modality (sequence type) variation should be less than inter-modality variation on the brain MRI dataset (Table 1), where the FRD with respect to the T1 sequence training set is noticeably higher for the T2 and FLAIR sequences' test sets (7.762 and 7.147, respectively), compared to the FRD computed for the same sequence type (but different patients) of the T1 test set (2.923).

Recommendations for Best Medical Image Translation Models. A further contribution of our study comprises a comprehensive set of experiments to compare source-to-target domain image translation methods. To this end, our empirical findings suggest notable general recommendations for medical image translation models. For severe domain shifts (*e.g.*, lumbar spine and CHAOS), CUT performed best in both downstream tasks and perceptual metrics of FRD and RadFID, likely due to its contrastive learning approach that preserves image structure. For more subtle shifts (breast MRI and brain MRI), GcGAN performs well, with UNSB and CycleGAN also effective for breast MRI. As such, our findings recommend incorporating contrastive learning approaches such as CUT for high domain shifts, while adversarial learning may suffice for moderate shifts, as demonstrated by GcGAN.

Limitations. Our work covered a range of diagnostic tasks and imaging domains, but is still limited to radiology. Other modalities could be explored, though radiomic features may need adjustment or (partial) replacement by other types of modality-specific imaging biomarkers. In this regard, our work demonstrates the vast potential of using imaging biomarker variability as a dataset comparison technique across medical domains.

For instance, in histopathology (Gurcan et al., 2009; Gupta et al., 2019; Hölscher et al., 2023) and cytopathology (Rodenacker and Bengtsson, 2003), quantitative pathomics features can be extracted, characterizing tissue architecture, cellular morphology, cell density, texture, intensity, and spatial relationships. Therefore, measuring the variability between such pathomics feature distributions across sets of cytology or histology images or image patches can provide insights into out-of-distribution detection, tissue characterization, disease localization, and generative model evaluation in this domain. This motivates future work to define and empirically assess variants of a respective “Fréchet Pathomic Distance”. On another note, in further medical imaging domains, such as dermatology, radiomic features have already shown utility in disease characterization (Attallah and Sharkas, 2021; Wang et al., 2024a), suggesting such features and, thus, FRD can be directly applicable for comparing distributions in such domains. Therefore, there is much potential for future work in quantifying the descriptive power of radiomics across different medical fields to identify where additional biomarkers are needed and where radiomics features suffice, followed by experiments evaluating FRD in these contexts.

Finally, our interpretability contributions are nascent, and further work is needed to extract more qualitative, yet concrete, insights. Additionally, the absolute values of FRD are only meaningful in relative comparisons between models on the same dataset (*i.e.*, measured using the same reference dataset D_1), similar to FID and other metrics.

Future Work and Expansions. There are many potential further applications of FRD with one example being the evaluation of multimodal models, such as vision-language models, *e.g.*, CLIP-type (Radford et al., 2021) models or text-to-image generative models, *e.g.*, Stable Diffusion-type (Rombach et al., 2022) models. FRD may also hold potential for evaluating video generative models (e.g., in ultrasound imaging (Chen et al., 2024a; Reynaud et al., 2025)), prompting further research into optimal strategies for feature handling–such as whether to (a) extract aggregated features from video sequences or (b) aggregate features extracted from individual frames. Additionally, applying FRD to sets of video frames could support identification of frames of interest and detection of transitions where the informational content of the video changes.

We evaluated the relationship between FRD and downstream task performance on single-task, single-domain models by correlating the FRD (computed between evaluation and reference sets) with model performance (Section 3.3). Given the field’s shift towards generalist foundation models (*e.g.*, those trained across multiple domains, modalities, or tasks) (Paschali et al., 2025), it’s crucial to assess FRD’s utility in this context. Following the initial pre-print release of this work, we applied FRD to the recently-released MRI-CORE foundation model for MRI analysis (Dong et al., 2025). We found that the FRD computed between MRI-CORE’s pre-training set and its zero-shot and few-shot evaluation sets was strongly correlated with the change in performance (3D Dice coefficient) gained from pre-training from initial SAM weights (Spearman $r = -0.721$ and -0.842, respectively), across ten segmentation tasks (Fig. 4 of Dong et al. (2025)). This result supports the broader utility of FRD for characterizing the performance of a range of models for medical image analysis.

Future work can consider certain modifications and expansions to FRD. For instance, radiomic features can be computed exclusively within a mask or region of interest (ROI), and, as another alternative, subsequently even further extended to a weighted combination of local (within-mask) and global (whole-image) features. This would enable additional applications, such as evaluating the quality or variability of image annotations by computing FRD between radiomic features extracted from a “gold standard” mask and those from candidate annotations. Such an approach would not only quantify the impact of inter- and intra-observer variability on radiomic biomarkers but also offer a complementary way to assess segmentation and object detection model performances. This assessment can be complementary alongside traditional annotation similarity metrics such as the Dice Score, Intersection over Union and Hausdorff Distance, which can be overly influenced by annotation size, spatial extent, and outliers (Reinke et al., 2024). On another note, the effect of weighting certain types of features (*e.g.*, image-space, frequency-space/wavelet, etc.) over others within FRD could be modified, which may be advantageous for certain applications.

CRediT authorship contribution statement

Nicholas Konz: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Richard Osuala:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Preeti Verma:** Software, Investigation, Formal analysis; **Yuwen Chen:** Validation, Formal analysis, Data curation; **Hanxue Gu:** Visualization, Formal analysis, Data curation; **Haoyu Dong:** Visualization, Formal analysis; **Yaqian Chen:** Writing – review & editing, Conceptualization; **Andrew Marshall:** Writing – review & editing, Formal analysis, Conceptualization; **Lidia Garrucho:** Project administration, Methodology; **Kaisar Kushibar:** Methodology, Conceptualization; **Daniel M. Lang:** Supervision, Project administration, Funding acquisition; **Gene S. Kim:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization; **Lars J. Grimm:** Writing – review & editing, Supervision, Conceptualization; **John M. Lewin:** Writing – review & editing, Investigation, Conceptualization; **James S. Duncan:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization; **Julia A. Schnabel:** Writing – review & editing, Supervision, Project administration, Funding acquisition; **Oliver Diaz:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization; **Karim Lekadir:** Writing – review & editing, Supervision, Project administration, Funding acquisition; **Maciej A. Mazurowski:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Maciej A. Mazurowski reports financial support was provided by National Institute of Biomedical Imaging and Bioengineering. Karim Lekadir reports financial support was provided by Horizon Europe. Karim Lekadir reports financial support was provided by Horizon 2020. Oliver Diaz reports financial support was provided by Ministry of Science, Innovation and Universities of Spain. Richard Osuala reports financial support was provided by Helmholtz Association Helmholtz Information & Data Science Academy. Daniel M. Lang reports financial support was provided by Helmholtz Information & Data Science Incubator. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Research reported in this publication was supported by the National Institute Of Biomedical Imaging And Bioengineering of the National Institutes of Health under Award Number R01EB031575. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research received funding from the European Union's Horizon Europe and Horizon 2020 research and innovation programme under grant agreement no 101057699 (RadioVal) and no 952103 (EuCanImage), respectively. It was further partially supported by the project FUTURE-ES (PID2021-126724OB-I00) and AIMED (PID2023-146786OB-I00) from the Ministry of Science, Innovation and Universities of Spain. Richard Osuala acknowledges a research stay grant from the Helmholtz Information and Data Science Academy (HIDA). Daniel M. Lang and Julia A. Schnabel received funding from HELMHOLTZ IMAGING, a platform of the Helmholtz Information & Data Science Incubator.

Appendix A. Dataset and task labeling details

A.1. Breast MRI

For breast MRI we use the 2D slices of the pre-contrast scan volumes from the Duke Breast Cancer dataset (Saha et al., 2018a), using the same train/validation/test splits (by patient) and preprocessing of (Konz et al., 2024a) from the 100 patient volumes with FGT and breast segmentation annotations (see the following paragraph). This results in train/validation/test splits with source, target domain sub-splits of size $\{4096, 7900\}/\{432, 1978\}/\{688, 1890\}$ images.

FGT and breast segmentation. FGT (fibroglandular/dense tissue) and breast segmentation masks for this dataset are provided from (Lew et al., 2024).

Cancer classification/slice-level detection. For the cancer classification task, we follow the same convention of (Konz and Mazurowski, 2024b), and label slice images as cancer-positive if they contain any tumor bounding box annotation, and negative if they are at least 5 slices away from any positive slices (ignoring the intermediate ambiguous slices). We then train a basic ResNet-18 (He et al., 2016) (modified for 1-channel input images) as our binary cancer classification model, on the positive and negative slices from the training set's GE scans. The model's evaluation datasets are otherwise unchanged from the other downstream tasks (besides the labels used for the images).

A.2. Brain MRI

For brain MRI, we utilized the multi-modal brain tumor dataset from the BraTS 2018 challenge (Menze et al., 2015). Since the BraTS's own validation set doesn't have masks available, we began by extracting the original shared training set and dividing the patients into training, validation, and test sets with a ratio of 0.7:0.15:0.15 for this paper. Next, we focused on the T1 and T2 sequence volumes along with their corresponding masks. Each slice of the image volume was normalized and saved as 2D PNG files to construct our 2D dataset. Note that because by default, each patient has both T1 and T2 scans, we used randomly sampling to construct the T1 and T2 subsets of the train, validation, and test sets such that there is no overlap between patients for the T1 and T2 sets (for example, the T1 and T2 test set images).

Tumor segmentation and detection. The original mask contains multiple classes of segmentation, including: Background (Label 0), Enhancing Tumor (Label 4), Tumor Core (Label 1), Whole Tumor (Label 2), Peritumoral Edema (Label 3). We conducted a binary tumor/not-tumor segmentation by combing all pixels with label larger than 0. For tumor detection, the tumor bounding box is generated by the smallest box that covers the entire tumor region. For those cases without tumor shown in that slice, we excluded them during model training/validation/testing.

Cancer classification/slice-level detection. We also further modify the task into a binary tumor classification task: whether this slice contains tumor or not. For those slices that are near the boundary of the tumor, specifically the 5 slices before and after the tumor presence (switching between positive and negative in each volume), we excluded them from the classification as they are considered ambiguous slices.

A.3. Lumbar spine

The CT lumbar spine dataset is obtained from TotalSegmentator (Wasserthal et al., 2023), and the T1 MRI data is private (to be revealed upon paper acceptance). We split the 2D source and target data in train/val/test as $\{495, 1466\}/\{175, 409\}/\{158, 458\}$.

Bone segmentation. We perform binary classification on each pixel to determine whether it includes bone or not. The ground truth masks for MRI are reviewed by experts, while the CT masks are sourced from Wasserthal et al. (2023).

A.4. CHAOS

We extract 2D CT and T1 in-phase MRI slices from the CHAOS dataset (Kavur et al., 2021). For each domain, we randomly split the data by patient in the ratio of 10:5:5, resulting in the 2D slices for the source and target domains being divided into train/val/test as $\{1488, 322\}/\{926, 182\}/\{460, 182\}$.

Liver segmentation. Liver masks for both modalities are provided by Kavur et al. (2021).

Liver classification. We assign positive labels to slices which contain the liver and negative labels to those that do not.

A.5. Single breast cancer prediction experiments

For the experiments of Section 4.4, we used the MAMA-MIA (Garrucho et al., 2025) breast DCE-MRI dataset. For the train and test sets of MAMA-MIA, each 2D slice image from each axial MRI volume was split in half down the middle to result in two images of single breasts. A given single breast image was then labeled as either healthy or cancerous if its accompanying lesion segmentation (from Garrucho et al. (2025)) was non-zero within the image. Applying this procedure to MAMA-MIA's test set resulted in our experiment's test set of both healthy and cancerous single breast images, and applying it to MAMA-MIA's train set resulted in our reference sets of healthy and cancerous images, D_{healthy} and D_{cancer}, respectively.

Appendix B. Model training/architectural details

In this section we describe the training details of all networks in the paper. All experiments were completed on four 48GB NVIDIA A6000 GPUs.

B.1. Translation models

All six translation models (CycleGAN (Zhu et al., 2017), MUNIT (Huang et al., 2018), CUT (Park et al., 2020), GcGAN (Fu et al., 2019), MaskGAN (Phan et al., 2023), and UNSB (Kim et al., 2024)) were trained with their default settings (besides being modified to input and output 1-channel images), except for a few exceptions to be described shortly; these settings are shown in Table B.5.

The exceptions are that for MUNIT and CUT, training for too long resulted in drastic changes in image content for breast MRI and lumbar so we chose earlier model iterations of $10,000$ and $20,000$, respectively for MUNIT, and 20 epochs for both for CUT.

Table B.5
Translation model training details.

Model	Training time	Batch size
CycleGAN (Zhu et al., 2017)	200 epochs	4
MUNIT (Huang et al., 2018)	1M iters.	1
CUT (Park et al., 2020)	200 epochs	1
GcGAN (Fu et al., 2019)	200 epochs	32
MaskGAN (Phan et al., 2023)	200 epochs	4
UNSB (Kim et al., 2024)	200 epochs	1

B.2. Downstream task models

In this section we describe the architectural and training details of all models trained for the downstream tasks of each dataset (Table 1) on its respective target domain data from the training set. All models are trained with Adam (Kingma and Ba, 2015) and a weight decay strength of 10^{-4} for 100 epochs.

Segmentation. For all segmentation downstream tasks we train a standard UNet (Ronneberger et al., 2015) with five encoding blocks, at a batch size of 8 with a learning rate of 0.01. The model is trained with equally-weighted cross-entropy and Dice losses, the latter implemented with MONAI (Cardoso et al., 2022).

Object Detection. For detection downstream tasks, we trained a Faster-RCNN (Girshick, 2015) with a batch size of 4 and a learning rate of 0.005. The model is implemented using Torchvision (maintainers and contributors, 2016), with the number of predicted classes modified to 2. The loss function is the default loss from this built-in model.

Classification/Slice-level Detection. For classification tasks we train a standard ResNet-18 (He et al., 2016), modified to take in one-channel inputs and output one logit (as all tasks are binary classification). We use a batch size of 64 and a learning rate of 0.001, with a cross-entropy loss.

Appendix C. Additional experiments

C.1. Ablation studies

C.1.1. Radiomic feature importance

To better interpret FRD, we assess the importance of different radiomic feature groups (textural/gray-level matrix, wavelet, first-order) by ablation: examining how removing each group affects the translation model downstream task performance results (Fig. 7). For the wavelet features, which correspond to those where the image is passed through a frequency-based filter (-LL, -LH, -HL or -HH) before the computation of the feature, we evaluate both completely excluding all features ("no wavelet") as well as removing those for only one filter type (-LL, -LH, -HL or -HH). Results are shown in Fig. C.16.

Overall, wavelet and first-order features are most crucial for FRD, as excluding them significantly worsens correlation results. In evaluating sub-types of wavelet features, we see that excluding wavelet-HH improves performance for three tasks (breast MRI segmentation (FGT) and brain MRI/BraTS segmentation and detection mIoU), yet worsens performance for others (breast MRI segmentation (breast) and classification, BraTS detection mAP, and CHAOS classification). We see similar tradeoffs for excluding wavelet-HL, -LH, or -LL features. In general, there is no consistent advantage to excluding one type of wavelet feature while including the others.

Textural features are somewhat important for breast MRI but have limited impact on other datasets. Breast MRI is generally the most sensitive to feature exclusion, suggesting that subtle domain shifts require a broader range of features for accurate analysis. Overall, these findings indicate that including all types of radiomic features for computing FRD results in a better general-purpose metric.

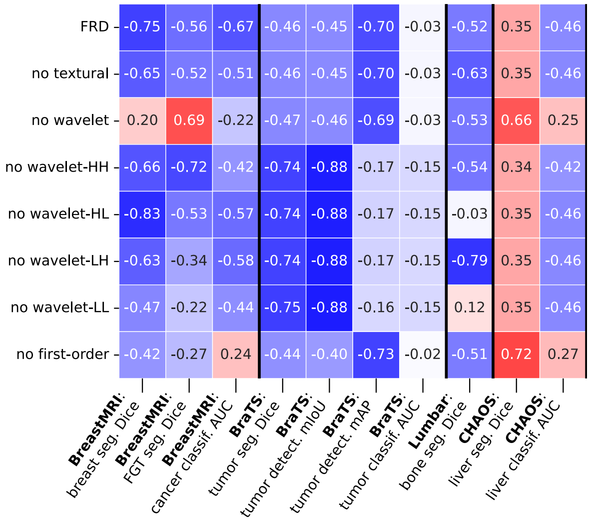


Fig. C.16. Importance of different radiomic features for FRD. Pearson correlation r between FRD and downstream task performance metrics across all translation models (as in Fig. 7), comparing using standard FRD with all features (top row) to removing certain groups of features (lower rows).

Table C.6
Correlation r of FRD computed with MMD distance (Eq. (C.1)) with standard Fréchet distance FRD (Eq. (3)), across all translation models.

Corr.type	BreastMRI	BrainMRI	Lumbar	CHAOS
Pearson	-0.75	-0.79	0.81	-0.84
Spearman	-0.83	-0.83	0.49	-0.54

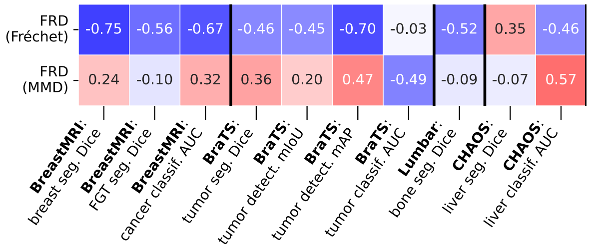


Fig. C.17. Pearson correlation r between FRD and downstream task performance metrics across all translation models (as in Fig. 7), comparing using Fréchet or MMD distance for FRD.

C.1.2. Using MMD instead of Fréchet distance

Perceptual metrics such as CMMD and KID use the MMD (Maximum Mean Discrepancy) distance metric d_{MMD} (Gretton et al., 2012) over the more common Fréchet distance, due to advantages such as lacking the Gaussianity assumption and being suitable for smaller datasets (Jayasumana et al., 2024; Bińkowski et al., 2018). Here we will evaluate calculating our proposed FRD distance using MMD (with a standard Gaussian RBF kernel) as

$$\mathrm{FRD}_{\mathrm{MMD}}(D_1, D_2) := d_{\mathrm{MMD}}(f_{radio}(D_1), f_{radio}(D_2)), \tag{C.1}$$

instead of via Fréchet distance as d_{radio} (Eq. (3)). We compare the two metrics (FRD and "FRD-MMD") in terms of (1) how much they correlate with downstream task performance metrics (as in Fig. 7), in Fig. C.17, and (2) whether they rank translation models similarly (linearly or non-linearly), in Table C.6.

We first see that FRD-MMD is noticeably inferior to FRD in terms of its negative correlation to downstream task performance (Fig. C.17); for all but one task, the correlation r is either close to zero, or in the wrong direction (positive r, as higher perceptual distance should correlate with worse performance, not better). Moreover, FRD-MMD is not consistent

Table C.7
Downstream task performance on test points detected as ID vs. OOD using our thresholding method.

	Breast MRI			Brain MRI				Lumbar	CHAOS	
Detected as:	Dice Breast	FGT	AUC Cancer	Dice Tumor	mIoU Tumor	mAP	AUC Cancer	Dice Bone	Dice Liver	AUC Liver
In-Domain	0.904	0.698	0.670	0.434	0.179	0.175	0.619	0.856	0.848	0.789
Out-of-Domain	0.731	0.473	0.535	0.498	0.215	0.223	0.618	0.001	0.129	0.463

in terms of its relationship to standard FRD (Table C.6). We hypothesize that these issues could potentially be due to the dependence of MMD on the choice of kernel, which could require further tuning, or the fact that MMD does *not* have an assumption of Gaussianity unlike the Fréchet distance, which may result in a metric that is too unconstrained.

C.2. OOD performance drop prediction

In Table C.7 we evaluate if images detected as out-of-domain with our OOD score thresholding approach (Section 4.1) also result in lowered performance compared to on detected ID cases.

C.3. OOD performance drop severity ranking

Here, we assess how well FRD and other perceptual metrics predict performance drops on out-of-domain (OOD) data. Given a model trained on target domain data D_t and two new OOD datasets $D_{\mathrm{OOD},1}$ and $D_{\mathrm{OOD},2}$, we examine if a metric d can correctly indicate which OOD dataset will suffer a greater performance drop. Specifically, we test if $\mathrm{Perf}(D^{\mathrm{test}}_{\mathrm{OOD},2}) < \mathrm{Perf}(D^{\mathrm{test}}_{\mathrm{OOD},1})$ aligns with $d(D^{\mathrm{test}}_{\mathrm{OOD},2}, D_t) > d(D^{\mathrm{test}}_{\mathrm{OOD},1}, D_t)$, and vice versa.

We evaluate this scenario with the datasets in Table 1 which possess additional data domains beyond the target domain and default source domain D_s, namely, BraTS using its T2-FLAIR data (Menze et al., 2015), and CHAOS using its T1 Dual Out-Phase and T2 SPIR MRI data (Kavur et al., 2021). We show these task performance vs. perceptual distance agreement results in Table C.8 for each type of downstream task, and for each possible pair of $D_{\mathrm{OOD},1}$ and $D_{\mathrm{OOD},2}$ for each dataset (T1 MRI and T2 FLAIR MRI for BraTS, respectively, and all 2-combinations of {T1 Dual Out-Phase MRI, T2 SPIR MRI, and CT} for CHAOS). Shown in Tables C.9 and C.10 are the specific results that generated Table C.8.

C.4. Towards dataset-level OOD detection

In Section 4.1, we showed how FRD/radiomic features can be used for single image-level binary OOD detection. However, a more realistic scenario may be that some new dataset is acquired from an outside hospital/site, and we wish to know if the dataset is generally OOD relative to our own reference ID dataset D_{ID} that we used to train some downstream task model, to get some idea of how our model will perform on the new dataset D_{test}. For example, our ID dataset could be breast MRI collected from GE scanners, and the new dataset could potentially have OOD (*e.g.*, Siemens) images. Our goal is therefore to have a metric that returns an (approximately) standardized value if D_{test} is OOD.

A naive prior approach to this could be to measure the FID or RadFID between D_{ID} and D_{test}, but as we will show, such distances are not clearly interpretable due to the distance value being noticeably affected by the specific dataset used, as well as the sample size (Section 5.2). To this end, we propose a FRD-based metric for dataset-level OOD detection which is designed to return 1 (or a value close to it) when the test set is completely OOD.

We do so by considering an ID reference point $x_{\mathrm{ID}} \sim D_{\mathrm{ID}}$ and test set point $x \sim D_{\mathrm{test}}$, both randomly sampled. Now, we wish to have a metric that estimates the probability that the test set is OOD. The key insight here is that the higher this probability, the higher the chance that x is OOD, such that it's expected score/distance from D_{ID}, $s(x)$ (Eq. (4)) will in turn be more likely to be larger than that of a typical ID point x_{ID}. Assuming that OOD points will not be typically *closer* to D than ID points, which is true by the definition of OOD, then the minimum value of this probability is $\Pr[s(x) > s(x_{\mathrm{ID}})] = 0.5$ if D_{test} is 100% ID (no clear difference between the test set and reference set score distributions), and $\Pr[s(x) > s(x_{\mathrm{ID}})] = 1$ if D_{test} is 100% OOD.

We then convert this to a metric, $\mathrm{nFRD}_{\mathrm{group}}$ (FRD for group-level OOD detection normalized to a fixed range) that ranges from 0 to 1 with

$$\mathrm{nFRD}_{\mathrm{group}} := 2(\Pr[s(x) > s(x_{\mathrm{ID}})] - 0.5). \tag{C.2}$$

The final question is then how $\Pr[s(x) > s(x_{\mathrm{ID}})]$ can be computed in practice; thankfully, the area under the ROC curve (AUC) *by definition* is this quantity (Fawcett, 2006), which can be easily computed, giving

$$\mathrm{nFRD}_{\mathrm{group}} := 2(\mathrm{AUC}[S_{\mathrm{test}}, S_{\mathrm{ID}}] - 0.5), \tag{C.3}$$

where $S_{\mathrm{test}} := \{s(x) : x \in D_{\mathrm{test}}\}$ and S_{ID} is the reference distribution of ID scores, $S_{\mathrm{ID}} := \{s(x_{\mathrm{ID}}; D_{\mathrm{ID}} \setminus x_{\mathrm{ID}}) : x_{\mathrm{ID}} \in D_{\mathrm{ID}}\}$, as in Section 4.1.

We evaluate $\mathrm{nFRD}_{\mathrm{group}}$ for OOD-scoring OOD test sets in Table C.11, averaged over 10 randomly sampled test sets of size 100 for each trial, compared to using the FID or FRDFID between D_{test} and D_{ID}. While all metrics assign a higher score for the OOD test set than the ID test set, we note that the scale of FID and RadFID OOD test set distance values changes noticeably depending on the dataset, a factor which would be even more pronounced if considering datasets of different sizes, as those metrics can be highly unstable for different sample sizes (Section 5.2). On the other hand, $\mathrm{nFRD}_{\mathrm{group}}$ is ≈ 1 for the OOD test set in 3/4 datasets (besides BraTS, due to it generally proving difficult for disentangle the ID and the OOD distributions (Section 4.1)), making it a more standardized, interpretable and practical metric. This enables us to posit that a $\mathrm{nFRD}_{\mathrm{group}}$ score of ≈ 1 for some new dataset means that the dataset is likely OOD.

We similarly see that for completely ID test sets (Table C.12), $\mathrm{nFRD}_{\mathrm{group}}$ is ≈ 0 in 3/4 cases. While RadFID does so for 4/4 cases, this doesn't account for the fact that RadFID is still highly sensitive to sample size, hurting its interpretable, standardized, realistic use in this case. Finally, we also show ablation studies in these two tables of using ImageNet or RadImageNet features to compute $\mathrm{nFRD}_{\mathrm{group}}$ instead of radiomic features, where we see that using these learned features results in less stable OOD test set distance values.

C.5. FRD vs. $\mathrm{FRD}_{\mathbf{v0}}$: A study on realistic MRI corruptions

Now we will compare the sensitivity of FRD to image corruptions compared to its predecessor, $\mathrm{FRD}_{\mathbf{v0}}$ Osuala et al. (2024), in the context of realistic MRI corruptions. Given an image dataset D, we will measure the effect that randomized transformations T have on the distance $d(D, T(D))$ (measured via FRD or $\mathrm{FRD}_{\mathbf{v0}}$), as T becomes more severe. The dataset that we use is the CHAOS T1-Dual In-Phase MRI training set (Table. 1). These transformations (besides random swapping) were chosen to simulate realistic artefacts to which MRI may be susceptible, due to issues such as motion, noise, etc.

Given some image $x \sim D$ with pre-established maximum intensity $I_{\max}$, the transformations T that we consider are listed as follows, each controlled by a percentage/severity parameter $p \in [0, 100]$.

1. **Gaussian noise.** Random Gaussian noise is added to x via $x \leftarrow \mathrm{clip}(x + \frac{pI_{\max}}{100}\epsilon, [0, I_{\max}])$ where $\epsilon \sim \mathcal{N}(0, 1)$.
2. **Gaussian blur.** Apply Gaussian blur with kernel size $k = \frac{p}{100} \times$ `max_size`, rounded to the nearest odd integer, where `max_size` is the pixel length of the longest side of the image.

Table C.8
Can FRD predict OOD performance drop severity? For each downstream task type (sub-tables) trained on a given dataset's target domain data D_t (first row) and for two OOD test sets $D^{\text{test}}_{\text{OOD},1}$ and $D^{\text{test}}_{\text{OOD},2}$ (second row), whether $\text{Perf}(D^{\text{test}}_{\text{OOD},2}) < \text{Perf}(D^{\text{test}}_{\text{OOD},1})$ does (✓) or does not (✗) correspond to $d(D^{\text{test}}_{\text{OOD},2}, D_t) > d(D^{\text{test}}_{\text{OOD},1}, D_t)$ and vice-versa. "-" denotes that the given perceptual metric was only negligibly affected by the change of the OOD dataset. "DOP" is "Dual Out-Phase".

Segmentation (*Dice*)				
	Brain MRI:trained on T2	CHAOS: trained on T1 Dual In-Phase		
	T1vs. T2 FLAIR	T1 DOPvs. T2 SPIR	T1 DOPvs. CT	T2 SPIRvs. CT
FRD	✓	✗	✓	✓
RadFID	✓	✓	✓	✓
FID	✗	✓	✓	✓
KID	-	✓	✓	✓
CMMD	✗	✓	✓	✓
Detection				
Brain MRI:trained on T2				
T1 vs. T2 FLAIR				
mIoU	*mAP*			
✗	✓			
✗	✓			
✓	✗			
-	-			
✓	✗			
Classification (*AUC*)				
Brain MRI:trained on T2	CHAOS: trained on T1 Dual In-Phase			
T1vs. T2 FLAIR	T1 DOPvs. T2 SPIR	T1 DOPvs. CT	T2 SPIRvs. CT	
✓	✓	✓	✓	
✓	✗	✓	✓	
✗	✗	✓	✓	
-	✗	✓	✓	
✗	✗	✓	✓	

Table C.9
Downstream task performance $\text{Perf}(D^{\text{test}}_{\text{OOD}})$ (left block) and perceptual distances $d(D^{\text{test}}_{\text{OOD}}, D_t)$ (right block) on out-of-domain data $D^{\text{test}}_{\text{OOD}}$ (each row) for downstream task models trained on (in-domain) BraTS T2 MRI data D_t, to supplement Table C.8.

	Downstream task performance				Perceptual distance metrics				
$D^{\text{test}}_{\text{OOD}}$	Dice	mIoU	mAP	AUC	FRD	RadFID	FID	KID	CMMD
T1 MRI	0.005	0.152	0.065	0.727	6.18	0.25	108	0.089	0.179
T2 FLAIR MRI	0.286	0.144	0.108	0.885	5.09	0.19	117	0.088	0.394

Table C.10
Downstream task performance $\text{Perf}(D^{\text{test}}_{\text{OOD}})$ (left block) and perceptual distances $d(D^{\text{test}}_{\text{OOD}}, D_t)$ (right block) on out-of-domain data $D^{\text{test}}_{\text{OOD}}$ (each row) for downstream task models trained on (in-domain) CHAOS T1 Dual In-Phase MRI data D_t, to supplement Table C.8.

	Downstream task performance		Perceptual distance metrics				
$D^{\text{test}}_{\text{OOD}}$	Dice	AUC	FRD	RadFID	FID	KID	CMMD
T1 Dual Out-Phase MRI	0.779	0.853	7.87	0.09	143	0.096	0.205
T2 SPIR MRI	0.262	0.867	7.55	0.20	189	0.126	0.507
CT	0.062	0.504	60.6	0.65	277	0.268	1.666

Table C.11
Dataset-level OOD detection scores for OOD test sets.

Metric	BreastMRI	Brain MRI	Lumbar	CHAOS
FID	178	223	277	338
RadFID	0.22	0.35	1.23	1.64
$\text{nFRD}_{\text{group}}$	1.00	0.62	0.95	1.00
+ ImageNet	0.01	0.84	0.75	0.94
+ RadImageNet	0.14	0.32	0.99	0.94

Table C.12
Dataset-level OOD detection scores for ID test sets.

Metric	BreastMRI	Brain MRI	Lumbar	CHAOS
FID	92	73	77	48
RadFID	0.09	0.06	0.04	0.04
$\text{nFRD}_{\text{group}}$	0.00	0.04	0.07	0.44
+ ImageNet	0.22	-0.04	0.05	-0.05
+ RadImageNet	0.1	-0.05	0.07	-0.13

3. **Random swap.** Randomly chosen small square patches of size $k = 15$ are swapped in the image, repeated $\text{round}(p)$ times.
4. **Motion artifacts.** MRI motion artifacts are simulated with realistic movement transformations (rotation and/or translation) via the `RandomMotion` function of TorchIO (Pérez-García et al., 2021). We used `RandomMotion` with parameters `degrees` $= \frac{p}{100} \times 10$, `translation` $= \frac{p}{100} \times 10$, and `num_transforms` $= \max(1, \text{round}(\frac{p}{100} \times 2))$.
5. **Random bias field.** MRI bias field artifacts are simulated via spatially-varying low-frequency intensity variations, implemented via TorchIO's `RandomBiasField` function with `coefficients` $= \frac{p}{100} \times 0.5$.

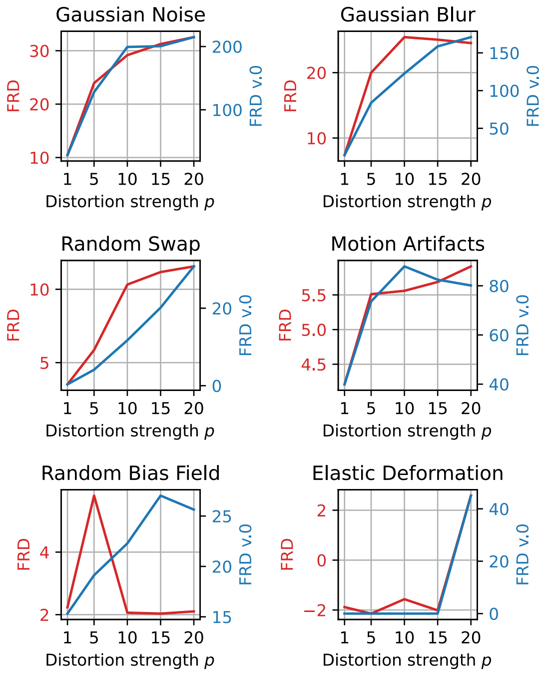


Fig. C.18. Comparison of FRD and FRD_{v0} (Osuala et al., 2024) in terms of sensitivity to abdominal MR image corruptions.

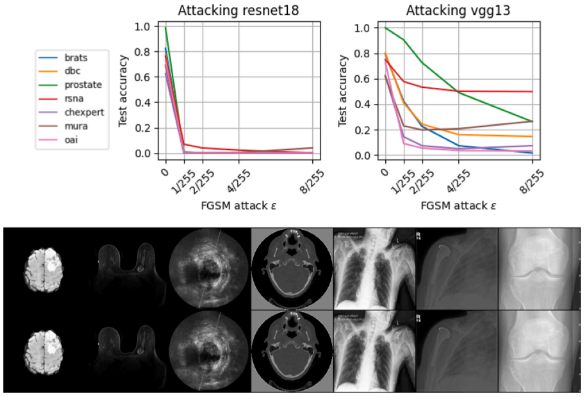


Fig. C.19. Top: Adversarial attack success (accuracy drop) on ResNet-18 and VGG-13 classification models. **Bottom:** Example non-attacked (first row) and attacked (second row) versions of the same images, from $\epsilon = 1/255$ FGSM attacks on ResNet-18 models for different datasets (columns of image; ordered left-to-right by legend labels in the top plots).

6. **Deformation.** MRI soft tissue deformations are simulated via random non-linear spatial corruptions, using TorchIO's `RandomElasticDeformation` function. We use parameters of `max_displacement` = $\frac{p}{100} \times 7.5$ and `num_control_points` = $\max(5, \text{round}(\frac{p}{100} \times 7))$.

We plot the results of this study in Fig. C.18. We see that FRD and FRD_{v0} have similar sensitivity behavior to image corruptions, besides the random bias field results for $p = 5$. In general, the two distance metrics increase as the corruption severity increases, which is the desired behavior.

C.6. Adversarial attack success results

We show the success of our adversarial attacks in terms of decreased model accuracy in Fig. C.19, alongside example clean and attacked images.

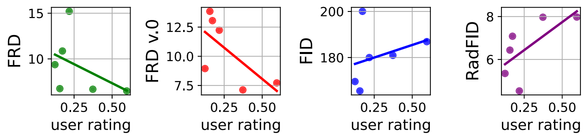


Fig. C.20. Top: Correlation coefficients (linear Pearson r_P and non-linear/rank Spearman r_S and Kendall τ) of different distance metrics with **calibrated** average user (radiologist) preference, for the task of measuring synthetic image quality. **Bottom:** Associated plots and linear best fits for this data.

C.7. Calibrating user preference ratings

Continuing from Section 5.1, a related way to measure a reader's rating of the quality of synthetic images is not just by their absolute average Likert score for the images, but by calibrating this measure by subtracting the average score for a set of relevant real images from it (in particular, 15 random real images for a given dataset and anatomical view). This aims to mitigate confounding effects on user preference, such as certain readers having a general bias in rating both synthetic and real images, higher or lower. We show the correlation results of this in Fig. C.20.

Similar to Fig. 10, we see that FRD correlates (negatively) fairly well with calibrated user preference, while FID and RadFID actually, undesirably, correlate positively. While FRD is substantially more correlated with user preference than FID and RadFID, FRD_{v0} does show a somewhat stronger correlation compared to FRD. Intuitively, this is plausible due to FRD including numerous additional frequency/wavelet radiomic features that are not present in FRD_{v0}, many of which (those not at low frequencies) are typically barely perceptible by the human eye (Dieleman, 2020; Salimans et al., 2017); thus, FRD will focus more on such features when comparing images. As demonstrated, the inclusion of these features resulted in noticeable improvements to FRD over FRD_{v0} in many medical image analysis applications (see Sections 4 and C.1.1), *i.e.*, where accounting for these subtle details is helpful for the underlying task.

C.8. Attempting to interpret differences between medical image distributions with learned features

We have applied feature inversion (Olah et al., 2017; Mahendran and Vedaldi, 2015) to visualize Δh using ImageNet and RadImageNet features (via Lucent (Kiat, 2021)), for breast MRI and brain MRI UNSB translation models, shown in Fig. 14. While the results hint at general textural and shape changes from translation, they lack clear, quantitative insights useful for clinical interpretation.

Appendix D. Additional discussion

D.1. Downstream task metrics as image distribution distance metrics

Segmentation performance metrics are themselves distance functions between two distributions of image features. If the predictions of the downstream task model on its test set and the corresponding ground truth labels/segmentations are taken as "features" of the images, then performance metrics such as Dice segmentation coefficient, IoU, etc. are image distribution metrics that clearly follow the topological requirements of distance metrics: reflexivity, non-negativity, symmetry, and the triangle inequality (Section 2).

References

Aerts, H. J.W.L., Velazquez, E.R., Leijenaar, R. T.H., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M.M., Leemans, C.R., Dekker, A., Quackenbush, J., Gillies, R.J., Lambin, P., 2014. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. 5 , 4006. https://doi.org/10.1038/ncomms5006

Alain, G., Bengio, Y., 2017. Understanding intermediate layers using linear classifier probes. In: International Conference on Learning Representations. https://openreview.net/forum?id=ryF7rTqgl.

AlBadawy, E.A., Saha, A., Mazurowski, M.A., 2018. Deep learning for segmentation of brain tumors: impact of cross-institutional training and testing. Med. Phys. 45 (3), 1150–1158.

Alyafi, B., Diaz, O., Elangovan, P., Vilanova, J.C., del Riego, J., Marti, R., 2020. Quality analysis of DCGAN-generated mammography lesions. In: 15th International Workshop on Breast Imaging (IWBI2020). Vol. 11513. SPIE, pp. 80–85.

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B., 2020. MedGAN: medical image translation using GANs. Comput. Med. Imaging Graph. 79, 101684.

Attallah, O., Sharkas, M., 2021. Intelligent dermatologist tool for classifying multiple skin cancer subtypes by incorporating manifold radiomics features categories. Contrast Med. Molecul. Imaging 2021, 7192016.

Bang, M., Eom, J., An, C., Kim, S., Park, Y.W., Ahn, S.S., Kim, J., Lee, S.-K., Lee, S.-H., 2021. An interpretable multiparametric radiomics model for the diagnosis of schizophrenia using magnetic resonance imaging of the corpus callosum. Transl. Psychiatry 11 , 462.

Beizaee, F., Desrosiers, C., Lodygensky, G.A., Dolz, J., 2023. Harmonizing flows: unsupervised MR harmonization based on normalizing flows. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 347–359.

Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A., 2018. Demystifying MMD GANs. In: International Conference on Learning Representations.

Cao, S., Konz, N., Duncan, J., Mazurowski, M.A., 2023. Deep learning for breast mri style transfer with limited training data. J. Digit. Imaging 36 , 666–678.

Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murray, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Zalbagi Darestani, M., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L. A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A., 2022. MONAI: An Open-Source Framework for Deep Learning in Healthcare. arXiv:2211.02701.

Cha, K.H., Hadjiiski, L., Chan, H.-P., Weizer, A.Z., Alva, A., Cohan, R.H., Caoili, E.M., Paramagul, C., Samala, R.K., 2017. Bladder cancer treatment response assessment in CT using radiomics with deep-learning. Sci. Rep. 7, 8738.

Chan, H.-P., Samala, R.K., Hadjiiski, L.M., Zhou, C., 2020. Deep learning in medical image analysis. Deep Learning in Medical Image Analysis: Challenges and Applications , 3–21.

Chen, H., Gomez, C., Huang, C.-M., Unberath, M., 2022a. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit. Med. 5 , 156.

Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D. F.K., Mahmood, F., 2021. Synthetic data in machine learning for medicine and healthcare. Nat. Biomed. Eng. 5 , 493–497.

Chen, T., Shi, Y., Zheng, Z., Yan, B., Hu, J., Zhu, X.X., Mou, L., 2024a. Ultrasound image-to-video synthesis via latent dynamic diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 764–774.

Chen, Y., Konz, N., Gu, H., Dong, H., Chen, Y., Li, L., Lee, J., Mazurowski, M.A., 2024b. Contourdiff: Unpaired image-to-image translation with structural consistency for medical imaging. arXiv:2403.10786.

Chen, Y., Yang, X.-H., Wei, Z., Heidari, A.A., Zheng, N., Li, Z., Chen, H., Hu, H., Zhou, Q., Guan, Q., 2022b. Generative adversarial networks in medical image augmentation: a review. Comput. Biol. Med. 144, 105382.

Clark, G.M., 2008. Prognostic factors versus predictive factors: examples from a clinical trial of erlotinib. Mol. Oncol. 1, 406–412.

Cui, S., Traverso, A., Niraula, D., Zou, J., Luo, Y., Owen, D., El Naqa, I., Wei, L., 2023. Interpretable artificial intelligence in radiology and radiation oncology. Br. J. Radiol. 96, 20230142.

Dembrower, K., Lindholm, P., Strand, F., 2020. A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks–the cohort of screen-aged women (CSAW). J. Digit. Imaging 33, 408–413.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009a. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009b. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Dieleman, S., 2020. Musings on typicality. https://benanne.github.io/2020/09/01/typicality.html.

Dong, H., Chen, Y., Gu, H., Konz, N., Chen, Y., Li, Q., Mazurowski, M.A., 2025. Mri-core: A foundation model for magnetic resonance imaging. arXiv:2506.12186.

Drukker, K., Li, H., Antropova, N., Edwards, A., Papaioannou, J., Giger, M.L., 2018. Most-enhancing tumor volume by MRI radiomics predicts recurrence-free survival “early on” in neoadjuvant treatment of breast cancer. Cancer Imaging 18, 1–9.

Durrer, A., Wolleb, J., Bieder, F., Sinnecker, T., Weigel, M., Sandkuehler, R., Granziera, C., Yaldizli, Ö., Cattin, P.C., 2024. Diffusion models for contrast harmonization of magnetic resonance images. In: Medical Imaging with Deep Learning. PMLR, pp. 526–551.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874.

Flanders, A.E., Prevedello, L.M., Shih, G., Halabi, S.S., Kalpathy-Cramer, J., Ball, R., Mongan, J.T., Stein, A., Kitamura, F.C., Lungren, M.P., Choudhary, G., Cala, L., Coelho, L., Mogensen, M., Morón, F., Miller, E., Ikuta, I., Zohrabian, V., McDonnell, O., Lincoln, C., Shah, L., Joyner, D., Agarwal, A., Lee, R.K., Nath, J., et al., 2020. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiol.: Artifi. Intell. 2, e190211. Publisher: Radiological Society of North America. https://doi.org/10.1148/ryai.2020190211

Fréchet, M., 1957. Sur la distance de deux lois de probabilité. In: Annales de l’ISUP. Vol. 6, pp. 183–198.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., Tao, D., 2019. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2427–2436.

Galloway, M.M., 1975. Texture analysis using gray level run lengths. Comput. Graph. Image Process. 4, 172–179.

Garrucho, L., Kushibar, K., Osuala, R., Diaz, O., Catanese, A., Del Riego, J., Bobowicz, M., Strand, F., Igual, L., Lekadir, K., 2023. High-resolution synthesis of high-density breast mammograms: application to improved fairness in deep learning based mass detection. Front. Oncol. 12, 1044496.

Garrucho, L., Kushibar, K., Reidel, C.-A., Joshi, S., Osuala, R., Tsirikoglou, A., Bobowicz, M., del Riego, J., Catanese, A., Gwoździewicz, K., Cosaka, M.-L., Abo-Elhoda, P.M., Tantawy, S.W., Sakrana, S.S., Shawky-Abdelfatah, N.O., Salem, A. M.A., Kozana, A., Divjak, E., Ivanac, G., Nikiforaki, K., Klontzas, M.E., García-Dosdá, R., Gulsun-Akpinar, M., Lafcı, O., Mann, R., Martín-Isla, C., Prior, F., Marias, K., Starmans, M. P.A., Strand, F., Díaz, O., Igual, L., Lekadir, K., 2025. A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. Sci. Data 12 , 453. https://doi.org/10.1038/s41597-025-04707-4

Gillies, R.J., Kinahan, P.E., Hricak, H., 2016. Radiomics: images are more than pictures, they are data. Radiology 278 , 563–577.

Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings. arXiv:1412.6572.

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., 2012. A kernel two-sample test. J. Mach. Learn. Res. 13, 723–773.

Guan, H., Liu, M., 2021. Domain adaptation for medical image analysis: a survey. IEEE Trans. Biomed. Eng. 69, 1173–1185.

Guo, B., Lu, D., Szumel, G., Gui, R., Wang, T., Konz, N., Mazurowski, M.A., 2024. The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study. arXiv:2409.04368.

Gupta, R., Kurc, T., Sharma, A., Almeida, J.S., Saltz, J., 2019. The emergence of pathomics. Curr. Pathobiol. Rep. 7, 73–84.

Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B., 2009. Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. 2, 147–171.

Halling-Brown, M.D., Warren, L.M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M.G., Wilkinson, L.S., Given-Wilson, R.M., McAvinchey, R., Young, K.C., 2020. Optimam mammography image database: a large-scale resource of mammography images and clinical data. Radiol.: Artifi. Intell. 3, e200103.

Hansen, C., Glinskis, S., Raju, A., Kornreich, M., Park, J., Pawar, J., Herzog, R., Zhang, L., Odry, B., 2024. Inpainting pathology in lumbar spine MRI with latent diffusion. arXiv:2406.02477.

Haralick, R.M., Shanmugam, K., Dinstein, I.H., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. , 610–621.

Hashmi, A. U.R., Almakky, I., Qazi, M.A., Sanjeev, S., Papineni, V.R., Mahapatra, D., Yaqub, M., 2024. Xreal: Realistic anatomy and pathology-aware x-ray generation via controllable diffusion model. arXiv:2403.09240.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural Inf. Process. Syst. 30.

Huang, E.P., O’Connor, J. P.B., McShane, L.M., Giger, M.L., Lambin, P., Kinahan, P.E., Siegel, E.L., Shankar, L.K., 2023. Criteria for the translation of radiomics into clinically useful tests. Nat. Rev. Clin. Oncol. 20, 69–82.

Huang, X., Liu, M.-Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 172–189.

Hölscher, D.L., Bouteldja, N., Joodaki, M., Russo, M.L., Lan, Y.-C., Sadr, A.V., Cheng, M., Tesar, V., Stillfried, S.V., Klinkhammer, B.M., Barratt, J., Floege, J., Roberts, I. S.D., Coppo, R., Costa, I.G., Bülow, R.D., Boor, P., 2023. Next-generation morphometry for pathomics-data mining in histopathology. Nat. Commun. 14 (1), 470. https://doi.org/10.1038/s41467-023-36173-0

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y., 2019. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI Press. https://doi.org/10.1609/aaai.v33i01.3301590

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S., 2024. Rethinking fid: towards a better evaluation metric for image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Jiang, Y., Edwards, A.V., Newstead, G.M., 2021. Artificial intelligence applied to breast MRI for improved diagnosis. Radiology 298 (1), 38–46.

Kang, M., Chikontwe, P., Won, D., Luna, M., Park, S.H., 2023. Structure-preserving image translation for multi-source medical image domain adaptation. Pattern Recognit. 144, 109840.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T., 2020. Training generative adversarial networks with limited data. Adv. Neural Inf. Process. Syst. 33, 12104–12114.

Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.-H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., Baydar, B., Lachinov, D., Han, S., Pauli, J., Isensee, F., Perkonigg, M., Sathish, R., Rajan, R., Sheet, D., Dovletov, G., Speck, O., Nürnberger, A., Maier-Hein, K.H., Bozdağı Akar, G., Ünal, G., Dicle, O., Selver, M.A., 2021. Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. Med. Image Anal. 69, 101950. https://doi.org/https://doi.org/10.1016/j.media.2020.101950

Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A., 2020. Gans for medical image analysis. Artif. Intell. Med. 109, 101938.

Kiat, L.S., 2021. greentfrapp/lucent: Lucid library adapted for PyTorch. https://github.com/greentfrapp/lucent.

Kim, B., Kwon, G., Kim, K., Ye, J.C., 2024. Unpaired image-to-image translation via neural schrödinger bridge. In: The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=uQBW7ELXfO.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R., 2018. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. PMLR, pp. 2668–2677. https://proceedings.mlr.press/v80/kim18d.html.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings. arXiv:1412.6980.

Kocak, B., Akinci D'Antonoli, T., Mercaldo, N., Alberich-Bayarri, A., Baessler, B., Ambrosini, I., Andreychenko, A.E., Bakas, S., Beets-Tan, R. G.H., Bressem, K., Buvat, I., Cannella, R., Cappellini, L.A., Cavallo, A.U., Chepelev, L.L., Chu, L. C.H., Demircioglu, A., deSouza, N.M., Dietzel, M., Fanni, S.C., Fedorov, A., Fournier, L.S., Giannini, V., Girometti, R., Groot Lipman, K. B.W., Kalarakis, G., Kelly, B.S., Klontzas, M.E., Koh, D.-M., Kotter, E., Lee, H.Y., Maas, M., Marti-Bonmati, L., Müller, H., Obuchowski, N., Orlhac, F., Papanikolaou, N., Petrash, E., Pfaehler, E., Pinto dos Santos, D., Ponsiglione, A., Sabater, S., Sardanelli, F., Seeböck, P., Sijtsema, N.M., Stanzione, A., Traverso, A., Ugga, L., Vallières, M., van Dijk, L.V., van Griethuysen, J. J.M., van Hamersvelt, R.W., van Ooijen, P., Vernuccio, F., Wang, A., Williams, S., Witowski, J., Zhang, Z., Zwanenburg, A., Cuocolo, R., 2024. METhodological Radiomics score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII. Insight. Imaging 15 (1), 8. https://doi.org/10.1186/s13244-023-01572-w

Konz, N., Chen, Y., Dong, H., Mazurowski, M.A., 2024a. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention.

Konz, N., Chen, Y., Gu, H., Dong, H., Mazurowski, M.A., 2024b. Rethinking perceptual metrics for medical image translation. In: Medical Imaging with Deep Learning. https://openreview.net/forum?id=acNu3glOTh.

Konz, N., Godfrey, C., Shapiro, M., Tu, J., Kvinge, H., Brown, D., 2023. Attributing learned concepts in neural networks to training data. In: NeurIPS Workshop on Attributing Model Behavior at Scale. https://openreview.net/forum?id=ZP7hgFNXQy.

Konz, N., Mazurowski, M.A., 2024a. The effect of intrinsic dataset properties on generalization: unraveling learning differences between natural and medical images. In: The Twelfth International Conference on Learning Representations.

Konz, N., Mazurowski, M.A., 2024b. Reverse engineering breast mris: predicting acquisition parameters directly from images. In: Medical Imaging with Deep Learning. PMLR, pp. 829–845.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G.P.M., Granton, P., Zegers, C. M.L., Gillies, R., Boellard, R., Dekker, A., Aerts, H. J.W.L., 2012. Radiomics: extracting more information from medical images using advanced feature analysis. Eur. J. Cancer 48 (4), 441–446. https://doi.org/10.1016/j.ejca.2011.11.036

Lew, C.O., Harouni, M., Kirksey, E.R., Kang, E.J., Dong, H., Gu, H., Grimm, L.J., Walsh, R., Lowell, D.A., Mazurowski, M.A., 2024. A publicly available deep learning model and dataset for segmentation of breast, fibroglandular tissue, and vessels in breast MRI. Sci. Rep. 14 (1), 5383.

Li, F., Hu, Z., Chen, W., Kak, A., 2023a. Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 632–641.

Li, H., Zhu, Y., Burnside, E.S., Drukker, K., Hoadley, K.A., Fan, C., Conzen, S.D., Whitman, G.J., Sutton, E.J., Net, J.M., Ganott, M., Huang, E., Morris, E.A., Perou, C.M., Ji, Y., Giger, M.L., et al., 2016. MR Imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of mammaprint, oncotype DX, and PAM50 gene assays. Radiology 281 (2), 382–391. Publisher: Radiological Society of North America. https://doi.org/10.1148/radiol.2016152110

Li, Y., Shao, H.-C., Liang, X., Chen, L., Li, R., Jiang, S., Wang, J., Zhang, Y., 2023b. Zero-shot medical image translation via frequency-guided diffusion models. IEEE Trans. Med. Imaging 43 (8), 980–993 .

Liu, M., Maiti, P., Thomopoulos, S., Zhu, A., Chai, Y., Kim, H., Jahanshad, N., 2021. Style transfer using generative adversarial networks for multi-site mri harmonization. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. Springer, pp. 313–322.

Lopez, M.G., Posada, N., Moura, D.C., Pollán, R.R., Valiente, J. M.F., Ortega, C.S., Solar, M., Diaz-Herrero, G., Ramos, I., Loureiro, J., Fernandes, T.C., de Araújo, B.F., 2012. Bcdr: a breast cancer digital repository. In: 15th International Conference on Experimental Mechanics. Vol. 1215, pp. 113–120.

Mahendran, A., Vedaldi, A., 2015. Understanding deep image representations by inverting them. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5188–5196.

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M.A., Wiesenfarth, M., Kavur, A.E., Sudre, C.H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M.B., Cardoso, M.J., Cheplygina, V., Cimini, B.A., Collins, G.S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Haase, R., Hashimoto, D.A., Hoffman, M.M., Huisman, M., Jannin, P., Kahn, C.E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B.A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A.L., Mattson, P., Meijering, E., Menze, B., Moons, K. G.M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C.I., Shetty, S., van Smeden, M., Summers, R.M., Taha, A.A., Tiulpin, A., Tsaftaris, S.A., Van Calster, B., Varoquaux, G., Jäger, P.F., 2024. Metrics reloaded: recommendations for image analysis validation. Nat. Method. 21 (2), 195–212. https://doi.org/10.1038/s41592-023-02151-z

Maintainers, T., contributors, 2016. Torchvision: pytorch's computer vision library. https://github.com/pytorch/vision.

Mandrekar, J.N., 2010. Receiver operating characteristic curve in diagnostic test assessment. J. Thorac. Oncol. 5 (9), 1315–1316.

McNaughton, J., Fernandez, J., Holdsworth, S., Chong, B., Shim, V., Wang, A., 2023. Machine learning for medical image translation: a systematic review. Bioengineering 10 (9), 1078.

Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z.A., Yang, Y., 2026. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. Radiol.: Artifi. Intell. 2, e210315. https://doi.org/10.1148/ryai.210315

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, c., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S. M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.-C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Székely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K., et al., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34 (10), 1993–2024. https://ieeexplore.ieee.org/document/6975210. https://doi.org/10.1109/TMI.2014.2377694

Modanwal, G., Vellal, A., Buda, M., Mazurowski, M.A., 2020. Mri image harmonization using cycle-consistent generative adversarial network. In: Medical Imaging 2020: Computer-Aided Diagnosis. Vol. 11314. SPIE, pp. 259–264.

Mårtensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., Rektorova, I., Bonanni, L., Pardini, M., Kramberger, M.G., Taylor, J.-P., Hort, J., Snædal, J., Kulisevsky, J., Blanc, F., Antonini, A., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., Simmons, A., Aarsland, D., Westman, E., 2020. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. Med. Image Anal. 66, 101714. https://doi.org/10.1016/j.media.2020.101714

Olah, C., Mordvintsev, A., Schubert, L., 2017. Feature visualization. Distill 2 (11), e7.

Orton, M.R., Hann, E., Doran, S.J., Shepherd, S. T.C., Ap Dafydd, D., Spencer, C.E., López, J.I., Albarrán-Artahona, V., Comito, F., Warren, H., Shur, J., Messiou, C., Larkin, J., Turajlic, S., Koh, D.-M., on behalf of the TRACERx RenalConsortium, 2023. Interpretability of radiomics models is improved when using feature group selection strategies for predicting molecular and clinical targets in clear-cell renal cell carcinoma: insights from the TRACERx renal study. Cancer Imaging 23 (1), 76. https://doi.org/10.1186/s40644-023-00594-3

Osuala, R., Kushibar, K., Garrucho, L., Linardos, A., Szafranowska, Z., Klein, S., Glocker, B., Diaz, O., Lekadir, K., 2023a. Data synthesis and adversarial networks: a review and meta-analysis in cancer imaging. Med. Image Anal. 84, 102704.

Osuala, R., Lang, D.M., Verma, P., Joshi, S., Tsirikoglou, A., Skorupko, G., Kushibar, K., Garrucho, L., Pinaya, W. H.L., Diaz, O., Schnabel, J.A., Lekadir, K., 2024. Towards learning contrast kinetics with multi-condition latent diffusion models. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2024. Springer Nature Switzerland, Cham, pp. 713–723.

Osuala, R., Skorupko, G., Lazrak, N., Garrucho, L., García, E., Joshi, S., Jouide, S., Rutherford, M., Prior, F., Kushibar, K., Díaz, O., Lekadir, K., 2023b. medigan: a Python library of pretrained generative models for medical image synthesis. J. Med. Imaging 10 (6), 061403. https://doi.org/10.1117/1.JMI.10.6.061403

Park, K., Choe, Y.J., Veitch, V., 2024. The linear representation hypothesis and the geometry of large language models. In: Forty-first International Conference on Machine Learning.

Park, T., Efros, A.A., Zhang, R., Zhu, J.-Y., 2020. Contrastive learning for unpaired image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer, pp. 319–345.

Paschali, M., Chen, Z., Blankemeier, L., Varma, M., Youssef, A., Bluethgen, C., Langlotz, C., Gatidis, S., Chaudhari, A., 2025. Foundation models in radiology: what, how, why, and why not. Radiology 314 (2), e240597.

Pérez-García, F., Sparks, R., Ourselin, S., 2021. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput. Method. Program. Biomed., 106236https://doi.org/10.1016/j.cmpb.2021.106236

Phan, V. M.H., Liao, Z., Verjans, J.W., To, M.-S., 2023. Structure-preserving synthesis: MaskGAN for unpaired MR-CT translation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 56–65.

Pinaya, W. H.L., Tudosiu, P.-D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J., 2022. Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models. Springer, pp. 117–126.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning. PMLR, pp. 8748–8763. https://proceedings.mlr.press/v139/radford21a.html.

Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y., 2018. MURA Dataset: towards radiologist-level abnormality detection in musculoskeletal radiographs https://openreview.net/forum?id=r1Q98pjiG.

Reinke, A., Tizabi, M.D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A.E., Rädsch, T., Sudre, C.H., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Buettner, F., Cardoso, M.J., Cheplygina, V., Chen, J., Christodoulou, E., Cimini, B.A., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Glocker, B., Godau, P., Hashimoto, D.A., Hoffman, M.M., Huisman, M., Isensee, F., Jannin, P., Kahn, C.E., Kainmueller, D., Kainz, B., Karargyris, A., Kleesiek, J., Kofler, F., Kooi, T., Kopp-Schneider, A., Kozubek, M., Kreshuk, A., Kurc, T., Landman, B.A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A.L., Meijering, E., Menze, B., Moons, K. G.M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rafelski, S.M., Rajpoot, N., Reyes, M., Riegler, M.A., Rieke, N., Saez-Rodriguez, J., Sánchez, C.I., Shetty, S., Summers, R.M., Taha, A.A., Tiulpin, A., Tsaftaris, S.A., Van Calster, B., Varoquaux, G., Yaniv, Z.R., Jäger, P.F., Maier-Hein, L., 2024. Understanding metric-related pitfalls in image analysis validation. Nat. Method. 21 (2), 182–194. https://doi.org/10.1038/s41592-023-02150-0

Reiss, T., Hoshen, Y., 2023. Mean-shifted contrastive loss for anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37, pp. 2155–2162.

Reynaud, H., Gomez, A., Leeson, P., Meng, Q., Kainz, B., 2025. Echoflow: A foundation model for cardiac ultrasound image and video generation. arXiv:2503.22357.

Rifi, A.L., Dufait, I., El Aisati, C., De Ridder, M., Barbé, K., 2023. Interpretability and repeatability of radiomic features: applied on in vivo tumor models. IEEE Trans. Instrum. Meas. 72, 1–7.

Rister, B., Yi, D., Shivakumar, K., Nobashi, T., Rubin, D.L., 2020. Ct-org, a new dataset for multiple organ segmentation in computed tomography. Sci. Data 7 (1), 381.

Rodenacker, K., Bengtsson, E., 2003. A feature set for cytometry on digitized microscopic images. Anal. Cell. Pathol. 25 (1), 1–36.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Saha, A., Harowicz, M.R., Grimm, L.J., Kim, C.E., Ghate, S.V., Walsh, R., Mazurowski, M.A., 2018a. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. Br. J. Cancer 119 (4), 508–516.

Saha, A., Harowicz, M.R., Grimm, L.J., Kim, C.E., Ghate, S.V., Walsh, R., Mazurowski, M.A., 2018b. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. Br. J. Cancer 119 (4), 508–516.

Salahuddin, Z., Woodruff, H.C., Chatterjee, A., Lambin, P., 2022. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. Comput. Biol. Med. 140, 105111.

Salimans, T., Karpathy, A., Chen, X., Kingma, D.P., 2017. PixelCNN++: improving the pixelCNN with discretized logistic mixture likelihood and other modifications. In: International Conference on Learning Representations. https://openreview.net/forum?id=BJrFC6ceg.

Saxena, S., Teli, M.N., 2021. Comparison and analysis of image-to-image generative adversarial networks: a survey. arXiv:2112.12625.

Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. f-anoGAN: fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. 54, 30–44.

Shi, L., Zhang, Y., Wong, I. H.M., Lo, C. T.K., Wong, T. T.W., 2023. MulhiST: multiple histological staining for thick biological samples via unsupervised image-to-image translation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 735–744.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings. arXiv:1409.1556.

Singh, A., Sengupta, S., Lakshminarayanan, V., 2020. Explainable deep learning models in medical image analysis. J. Imaging 6 (6), 52.

Sonn, G.A., Natarajan, S., Margolis, D. J.A., MacAiran, M., Lieu, P., Huang, J., Dorey, F.J., Marks, L.S., 2013. Targeted biopsy in the detection of prostate cancer using an office based magnetic resonance ultrasound fusion device. J. Urol. 189 (1), 86–92.

Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., Batmanghelich, K., 2022. Hierarchical amortized GAN for 3d high resolution medical image synthesis. IEEE J. Biomed. Health Inform. 26 (8), 3966–3975.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Thibault, G., Fertil, B., Navarro, C.L., Pereira, S., Cau, P., Lévy, N., Sequeira, J., Mari, J.-L., 2009. Texture indexes and gray level size zone matrix. application to cell nuclei classification. In: 10th International Conference on Pattern Recognition and Information Processing. https://api.semanticscholar.org/CorpusID:18989026.

Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., Saarakkala, S., 2018. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-Based approach. Sci. Rep. 8 (1), 1727. https://doi.org/10.1038/s41598-018-20132-7

Tschuchnig, M.E., Gadermayr, M., 2022. Anomaly detection in medical imaging-a mini review. In: Data Science–Analytics and Applications: Proceedings of the 4th International Data Science Conference–iDSC2021. Springer, pp. 33–38.

Van Griethuysen, J. J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G.H., Fillion-Robin, J.-C., Pieper, S., Aerts, H.J., 2017. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 77 (21), e104–e107.

Vorontsov, E., Molchanov, P., Gazda, M., Beckham, C., Kautz, J., Kadoury, S., 2022. Towards annotation-efficient segmentation via image-to-image translation. Med. Image Anal. 82, 102624.

Wagner, M.W., Namdar, K., Biswas, A., Monah, S., Khalvati, F., Ertl-Wagner, B.B., 2021. Radiomics, machine learning, and artificial intelligence–what the neuroradiologist needs to know. Neuroradiology, 1–11.

Wang, Z., Wang, C., Peng, L., Lin, K., Xue, Y., Chen, X., Bao, L., Liu, C., Zhang, J., Xie, Y., 2024a. Radiomic and deep learning analysis of dermoscopic images for skin lesion pattern decoding. Sci. Rep. 14 (1), 19781.

Wang, Z., Yang, Y., Chen, Y., Yuan, T., Sermesant, M., Delingette, H., Wu, O., 2024b. Mutual information guided diffusion for zero-shot cross-modality medical image translation. IEEE Trans. Med. Imaging 43 (8), 2825–2838.

Wasserthal, J., Breit, H.-C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M., 2023. Totalsegmentator: robust segmentation of 104 anatomic structures in CT images. Radiol.: Artifi. Intell. 5 (5). https://doi.org/10.1148/ryai.230024

Wolterink, J.M., Dinkla, A.M., Savenije, M. H.F., Seevinck, P.R., van den Berg, C. A.T., Išgum, I., 2017. Deep MR to CT synthesis using unpaired data. In: Simulation and Synthesis in Medical Imaging: Second International Workshop, SASHIMI 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 10, 2017, Proceedings 2. Springer, pp. 14–23.

Woodland, M., Castelo, A., Taie, M.A., Silva, J. A.M., Eltaher, M., Mohn, F., Shieh, A., Kundu, S., Yung, J.P., Patel, A.B., Brock, K.K., 2024. Feature extraction for generative medical imaging evaluation: new evidence against an evolving trend. In: International Conference on Medical Image Computing and Computer-Assisted Intervention.

Wu, Y., Liu, F., Yilmaz, R., Konermann, H., Walter, P., Stegmaier, J., 2025. A pragmatic note on evaluating generative models with fr\'echet inception distance for retinal image synthesis. arXiv:2502.17160.

Yang, J., Dvornek, N.C., Zhang, F., Chapiro, J., Lin, M., Duncan, J.S., 2019. Unsupervised domain adaptation via disentangled representations: application to cross-modality liver segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. Springer, pp. 255–263.

Yao, L., Prosky, J., Covington, B., Lyman, K., 2019. A strong baseline for domain adaptation and generalization in medical imaging. In: Medical Imaging with Deep Learning. https://openreview.net/forum?id=S1gvm2E-t4.

Ye, J.-Y., Fang, P., Peng, Z.-P., Huang, X.-T., Xie, J.-Z., Yin, X.-Y., 2024. A radiomics-based interpretable model to predict the pathological grade of pancreatic neuroendocrine tumors. Eur. Radiol. 34 (3), 1994–2005.

Yi, X., Walia, E., Babyn, P., 2019. Generative adversarial network in medical imaging: a review. Med. Image Anal. 58, 101552.

Yip, S. S.F., Aerts, H.J., 2016. Applications and limitations of radiomics. Phys. Med. Biol. 61 (13), R150.

Zhang, Z., Yang, L., Zheng, Y., 2018. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9242–9251.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232.

Zwanenburg, A., Leger, S., Vallières, M., Löck, S., 2016. Image biomarker standardisation initiative. arXiv:1612.07003.

Zwanenburg, A., Vallières, M., Abdalah, M.A., Aerts, H. J.W.L., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R.J., Boellaard, R., Bogowicz, M., Boldrini, L., Buvat, I., Cook, G. J.R., Davatzikos, C., Depeursinge, A., Desseroit, M.-C., Dinapoli, N., Dinh, C.V., Echegaray, S., El Naqa, I., Fedorov, A.Y., Gatta, R., Gillies, R.J., Goh, V., Götz, M., Guckenberger, M., Ha, S.M., Hatt, M., Isensee, F., Lambin, P., Leger, S., Leijenaar, R. T.H., Lenkowicz, J., Lippert, F., Losnegård, A., Maier-Hein, K.H., Morin, O., Müller, H., Napel, S., Nioche, C., Orlhac, F., Pati, S., Pfaehler, E. A.G., Rahmim, A., Rao, A. U.K., Scherer, J., Siddique, M.M., Sijtsema, N.M., Socarras Fernandez, J., Spezi, E., Steenbakkers, R. J.H.M., Tanadini-Lang, S., Thorwarth, D., Troost, E. G.C., Upadhaya, T., Valentini, V., van Dijk, L.V., van Griethuysen, J., van Velden, F. H.P., Whybra, P., Richter, C., Löck, S., et al., 2020. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 295 (2), 328–338. Publisher: Radiological Society of North America. https://doi.org/10.1148/radiol.2020191145