MELBA Journal Machine Learning for Biomedical Imaging

Knowledge-based anomaly detection for identifying network-induced shape artifacts

Rucha **Deshpande** [1], Tahsin **Rahman** [1], Miguel **Lago** [1], Adarsh **Subbaswamy** [1], Jana G. **Delfino** [1], Ghada **Zamzmi** [1], Elim **Thompson** [1], Aldo **Badano** [1], Seyed **Kahaki** [1]*

1 Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, U. S. Food and Drug Administration, USA

Abstract

Synthetic data provides a promising approach to address data scarcity for training machine learning models; however, adoption without proper quality assessments may introduce artifacts, distortions, and unrealistic features that compromise model performance and downstream utility for medical imaging research and evaluation purposes. This work introduces a novel knowledge-based anomaly detection method for detecting network-induced shape artifacts in synthetic images. The introduced method utilizes a two-stage approach comprising (i) a novel feature extractor that constructs a specialized feature space by analyzing the per-image distribution of angle gradients along anatomical boundaries, and (ii) an isolation forest-based anomaly detector. This representation captures local shape variations while maintaining global anatomical correspondence regardless of size differences. The isolation forest is trained on patient data to learn normal anatomical shape characteristics and assigns anomaly scores to synthetic images. This method may be used to detect anatomically unrealistic images irrespective of the generative model used and provides interpretability through its knowledge-based design. We demonstrate the effectiveness of the method for identifying network-induced shape artifacts in two synthetic mammography datasets generated by different architectures: a latent diffusion model and StyleGAN2, trained on CSAW-M and VinDr-Mammo patient datasets respectively. Quantitative evaluation shows that the method successfully concentrates artifacts in the most anomalous partition (1st percentile), with AUC values of 0.97 (CSAW-syn) and 0.91 (VMLO-syn). Comparison with anomaly detection based on other standard shape descriptors was used to demonstrate the superiority of the proposed method at isolating subtle shape artifacts along with extreme ones. In addition, a reader study involving three imaging scientists confirmed that images identified by the method as containing network-induced shape artifacts were also flagged by human readers with mean agreement rates of 66% (CSAW-syn) and 68% (VMLO-syn) for the most anomalous partition, approximately 1.5-2 times higher than the least anomalous partition. Kendall-Tau correlations between algorithmic and human rankings were 0.45 and 0.43 for the two datasets, indicating reasonable agreement despite the challenging nature of subtle artifact detection. This method is a step forward in supporting quality assurance practices for synthetic data in research and development contexts, as it allows developers to evaluate synthetic images for known anatomic constraints and pinpoint and address specific issues to improve the overall quality of a synthetic dataset. This method is presented as a research tool for synthetic image quality evaluation. Its application in clinical or regulatory decision-making contexts would require additional validation appropriate to the intended use.

Keywords

Network-induced shape artifacts, digital mammography, shape analysis, anomaly detection, synthetic data evaluation

Article informations

https://doi.org/10.59275/j.melba.2026-eb33

Volume 2026, Received: 2026-02-05, Published 2026-09-29

*Corresponding author: Seyed.Kahaki@fda.hhs.gov

Special issue: Medical Imaging with Deep Learning (MIDL) 2025 special issue

Guest editors: Lisa Koch, Ronald M. Summers, Chen Chen, Yan Zhuang

1. Introduction

Deep learning has improved medical imaging, particularly for automated image analysis applications in medical imaging research. However, the development of robust deep learning models is hindered by limited access to large-scale, high-quality patient datasets. This limitation is further worsened by issues such as patient privacy concerns (Giuffrè and Shung, 2023), unavailability of data for rare diseases (Vrudhula et al., 2024), and low quality (or even the lack) of truth labels. The limited availability of high-quality datasets restricts the ability to train and validate deep learning models, leading to over-fitting, poor generalization, and reduced accuracy when applying the models to unseen real-world data (Yang et al., 2024).

Synthetic data, defined as "artificial data that is intended to mimic the properties and relationships seen in real patient data" (U.S. Food and Drug Administration, 2024), offers a promising solution to address data-related limitations in healthcare (Sizikova et al., 2024). By generating synthetic data with known truth labels, the costs associated with data collection and acquisition can be significantly reduced. Furthermore, synthetic data can be used to augment real patient datasets, particularly for patients with rare diseases and/or from underrepresented populations, thereby reducing dataset bias due to class imbalance and potentially improving performance generalizability of deep learning algorithms in medical imaging (Chen et al., 2021; Garcea et al., 2023).

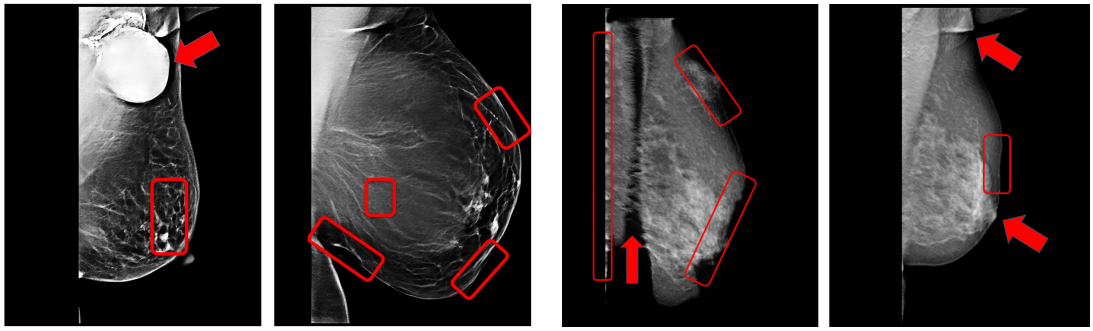
Various techniques have been developed to generate synthetic data, ranging from knowledge-based approaches (Badano et al., 2018; Kim et al., 2024) to advanced generative artificial intelligence (AI) methods (Kazerouni et al., 2023; Showrov et al., 2024). Based on physical modeling, knowledge-based approaches preserve physical constraints between attributes and physical findings (Sizikova et al., 2023), whereas generative AI approaches learn from large datasets and capture complex patterns and relationships in the data (Jeong et al., 2022; Koetzier et al., 2024). Despite these advancements, assessing the quality and clinical relevance of synthetic medical images, particularly those generated by generative AI models, remains a challenge (Deshpande et al., 2025a; Müller-Franzes et al., 2023).

A key challenge with synthetic data obtained from generative AI models is the occurrence of unrealistic features or artifacts, which can arise when models prioritize matching overall data distributions over preserving fine-grained, image-level details or adherence to clinical or anatomical constraints. Figure 1 shows examples of synthetic mammography images from datasets created by training different generative AI models on the CSAW-M (Pinaya et al., 2023a) and VinDr-Mammo (Pham et al., 2022) patient datasets respectively. The red annotations in these images highlight various AI-generated artifacts, including unnatural geometric patterns, synthetic noise within breast tissue, and artificial discontinuities along tissue boundaries. These artifacts demonstrate failure modes where the generative AI model inadequately capture the natural anatomical shapes and morphological structures in real patient mammography, possibly introducing systematic biases that could potentially affect the performance of downstream AI models in research settings, warranting further evaluation before clinical or regulatory application.

While network-induced shape artifacts in synthetic images are well-documented in the literature (Lee et al., 2023; Müller-Franzes et al., 2023; Deshpande et al., 2025a; Kelkar et al., 2023; Schwarz et al., 2021), automated methods for identifying such artifacts in individual images remain scarce (Deshpande et al., 2025a). Popular evaluation methods typically rely on dataset-wide metrics (Borji, 2022; Zamzmi et al., 2025), which summarize overall distribution alignment in a feature space. While useful for assessing general trends, these metrics overlook localized artifacts that may appear only in a fraction of the dataset, thus making it difficult to identify individual distorted images. This lack of granular assessment is particularly problematic because synthetic datasets often contain a vast number of images, which makes visual assessment of artifacts in individual images challenging.

While there is a potential need to evaluate the quality of individual synthetic images, such assessments in large synthetic datasets have unique challenges: (i) low-prevalence artifacts may be missed in visual spot-checking, which fails to provide a comprehensive evaluation of all images in the dataset, (ii) the types of artifacts induced by networks are often unknown and artifact labels typically unavailable, (iii) artifacts may vary across different generative models, and (iv) domain-specific factors, such as anatomy or imaging protocols, add further complexity to automated artifact detection. Thus, there is a need for domain-relevant methods that assess individual images for the presence of network-induced shape artifacts in synthetic datasets.

In this paper, we propose a method, outlined in Figure 2, for detecting network-induced shape artifacts using a knowledge-based feature space that captures the shape characteristics of the anatomy of interest. This feature space is constructed by analyzing the per-



(a) CSAW-M synthetic (b) VinDr-Mammo synthetic

Figure 1: Examples of synthetic mammography images generated from (a) CSAW-M and (b) VinDr-Mammo datasets. Red annotations are areas with unrealistic features that deviate from real patient mammographic anatomy.

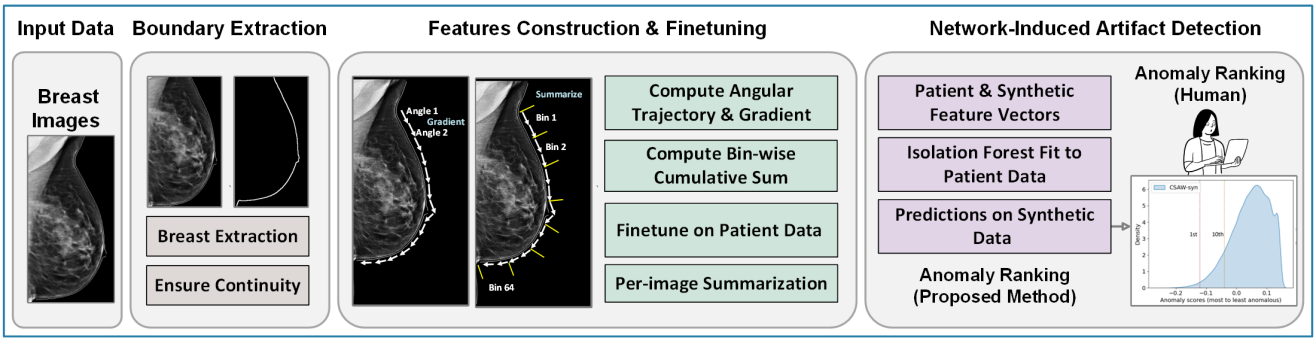


Figure 2: Overview of the proposed method for detecting network-induced shape artifacts.

image distribution of angle gradients along the boundary of the anatomical region of interest. Building on this representation, artifact detection is performed using an isolation forest (Liu et al., 2008). The isolation forest is trained on a patient dataset to capture the shape characteristics of real data and is then applied to the corresponding synthetic dataset. Each synthetic image is assigned an anomaly score, where highly negative scores indicate a higher likelihood of containing artifacts, while non-negative and high positive scores correspond to normal images. The proposed method (i) can identify images with unrealistic anatomical shapes, (ii) can greatly improve the efficiency of visual search by ranking a dataset of images, (iii) is model-agnostic due to the features being obtained from the generated images alone, and (iv) is interpretable due to its knowledge-based design. This method allows each image in a synthetic dataset to be evaluated for adherence to known anatomical constraints. As a result, developers can pinpoint and address specific issues and improve the overall quality of synthetic datasets in a targeted and efficient manner.

This work extends our prior conference paper presented at MIDL 2025 in several meaningful ways. First, we include a systematic comparison of our proposed feature space against standard radiomics-based shape descriptors from the PyRadiomics library (van Griethuysen et al., 2017), demonstrating the advantage of our method in detecting subtle artifacts beyond the most extreme cases. Second, we present an ablation study examining the effect of bin size on feature distributions and anomaly score rankings, providing empirical justification for our choice of $N = 64$. Third, we apply the trained isolation forest to the patient datasets themselves, analyzing the types of naturally occurring shape anomalies identified therein and contrasting them with network-induced artifacts found in the synthetic datasets. Fourth, we report Fréchet Radiomic Distance (FRD) scores (Konz et al., 2026) for both synthetic datasets as a complement to FID, providing a more medically relevant measure of distributional distance. Finally, to support reproducibility and enable better interpretation of dataset-specific findings, we have expanded the dataset descriptions to include details on the generative model architectures, training frameworks, and conditioning strategies for both synthetic datasets. Together, these additions constitute a meaningful extension of the conference work and provide a more complete characterization of the proposed method and its behavior across datasets

Table 1: Synthetic image datasets used in this study, including the number of generated images, training dataset, number of training images, and generative model employed. The Synkove dataset is publicly available and was used as is, the VMLO-syn dataset was generated for this work.

Synthetic Image Dataset Name	Generated Images	Name of training dataset	Training images	Generative model
CSAW-syn (SinKove)	100,000	CSAW-M	9,523	Latent diffusion U-Net
VMLO-syn	10,000	VinDr-Mammo (MLO view)	10,000	StyleGAN2

and experimental conditions.

2. Materials & Methods

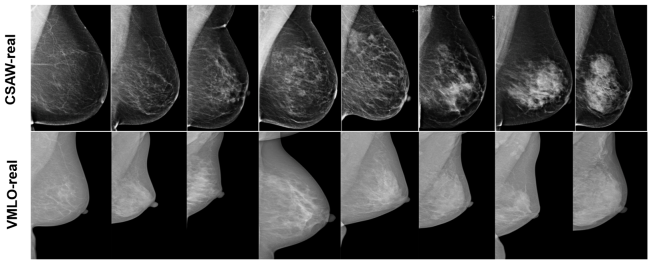
2.1 Datasets

While our proposed method is model agnostic and can be adapted to diverse anatomy, we demonstrate its utility using two synthetic digital mammography datasets in the mediolateral oblique (MLO) view. Each synthetic dataset was generated via a different generative model.

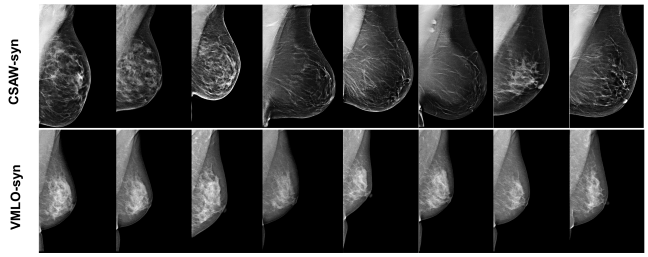
2.1.1 SinKove

The first synthetic dataset, SinKove, is a public dataset (Pinaya et al., 2023b) with 100,000 images generated by a latent diffusion model trained on the CSAW-M patient dataset (Sorkhei et al., 2021). The latent diffusion model comprises a variational autoencoder with KL-regularization to learn the latent representation from the patient data and then a diffusion U-Net to learn a generative model of the latents. The training of this model was performed using the NVIDIA (mention of commercial products does not constitute FDA endorsement; see Disclaimer) MONAI framework with NVIDIA GPUs. The trained model was then used to generate synthetic images which are not paired to the patient images and match the patient dataset at the distribution level. No demographic information or meta-data was used in the generation process. The only conditioning was masking level of cancer at 3 levels: low, medium and high.

CSAW-M (Sorkhei et al., 2021) is a publicly accessible mammography dataset of approximately 10,000 images drawn from the CSAW (Cohort of Screen-Aged Women) cohort of mammography screening data collected from women aged 40–74 in Sweden. The CSAW-M dataset is hosted on the SciLifeLab Data Repository, a Swedish life science data-sharing infrastructure, and is available for non-commercial use. Alongside the images, the dataset provides metadata, expert lesion annotations, clinical endpoints, density measures, and image acquisition parameters. It is partitioned into three subsets: a training set of 9,523 examples, a public test set of 497 examples, and a private test set of 475 examples. Only the training split was used to train the diffusion model for the Sinkove synthetic dataset.



(a) Patient data samples



(b) Synthetic data samples

Figure 3: Examples of images from (a) patient datasets: CSAW-real (upper) and VMLO-real (lower) and (b) synthetic datasets: CSAW-syn (upper) and VMLO-syn (lower).

Image selection focused on participants from Karolinska University Hospital, with the most recent mediolateral oblique (MLO) view chosen for each participant to maximize breast visualization. For cancer cases, the contralateral breast was used to prevent label contamination from tumor presence, whereas for non-cancer cases, breast side was selected at random. The available images were preprocessed using a standardized and released as 632×512, 8-bit PNG files ready for downstream analysis. Examples of CSAW-M images are shown in Figure 3.

These real and synthetic datasets are hereafter referred to as CSAW-real and CSAW-syn, respectively.

2.1.2 VMLO-syn

The second synthetic dataset, VMLO-syn, was generated using StyleGAN2 (Karras et al., 2020) trained on the VinDr-Mammo dataset (Pham et al., 2022). Only the MLO view subset (VMLO-real), which contains approximately 10,000 images, of the patient data was used for training resulting in a model that can generate MLO view images. We trained StyleGAN2 on NVIDIA A100 GPUs. Similar to the approach used for the Sinkove dataset, the images generated for VMLO-syn are not paired to the patient images and only match the patient dataset at the distribution level. The generation of the synthetic images was completely unconditioned.

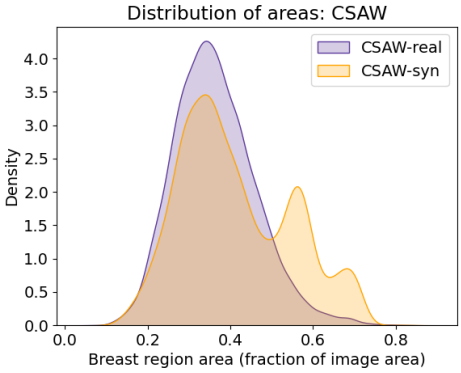
The VinDr-Mammo dataset (Pham et al., 2022) is a publicly available, anonymized, large-scale collection of mammography images developed to support research in computer-aided detection (CADe) and diagnosis (CADx) systems for breast cancer screening. It comprises 20,000 mammography images in DICOM format, drawn from 5,000 examinations performed between 2018 and 2020 at Hanoi Medical University Hospital (HMUH) and Hospital 108 (H108) in Vietnam. All images are full-field digital mammography (FFDM) acquisitions that were retrieved from each institution's Picture Archiving and Communication System (PACS), consistent with standard clinical acquisition workflows as described by the dataset authors. The dataset encompasses both screening and diagnostic exams, with images acquired using equipment from Siemens, IMS, and Planmed. To ensure patient privacy, all identifiable information was stripped from DICOM metadata and image annotations, and text appearing in image corners was additionally masked. The pseudonymization process was validated manually by human reviewers. Annotations are provided at both the breast level and the lesion level. Examples of VinDr-Mammo images are shown in Figure 3.

A summary of the synthetic dataset information is provided in Table 1.

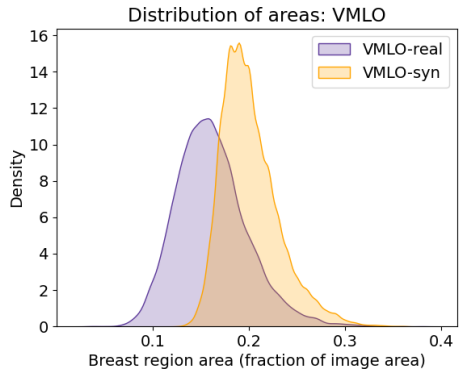
2.1.3 Pre-processing

For both datasets, pre-processing was performed as specified by the dataset authors (Pham et al., 2022; Sorkhei et al., 2021). For the VMLO-real dataset, an additional image resizing to 512×512 was performed to meet the input requirements of the generative model.

The patient and synthetic data distributions for breast area are shown below (Figure 4). Neither synthetic dataset exactly matches the breast area distribution in the corresponding patient dataset. The CSAW-syn dataset extrapolates beyond the real breast area distribution, and generates breasts of larger sizes than those in the corresponding training dataset. On the other hand, in VMLO-syn, the breast area distribution is shifted (biased) as compared to the corresponding patient dataset. This indicates that in both cases, 1) larger breast shapes than those in the patient dataset were generated and that 2) the prevalence of large breast shapes as seen in the patient dataset was greater than expected in the synthetic dataset. Thus, in both cases, the original distribution of breast area is clearly not maintained.



(a) CSAW



(b) VMLO

Figure 4: Distribution of breast areas as a fraction of the total image area. In case of CSAW-syn, breast area is extrapolated beyond the real distribution whereas in VMLO-syn, the breast area distribution appears shifted towards larger breast areas.

2.2 Network-induced Shape Artifact Detection

Our method (Figure 2) consists of boundary extraction, a novel feature construction and fine-tuning algorithm, and artifact detector.

2.2.1 Boundary Extraction and Feature Space Construction

The algorithm has two main stages: (1) boundary extraction and (2) feature construction and finetuning.

Boundary extraction and tracking To generate a feature representation of the anatomy of interest, the region of interest (breast in this use case), is segmented via thresholding and morphological operations. Then, a set of boundary pixels, $\mathcal{P} = \{p_{r,c} \in \mathcal{N}^2\}$, is extracted from the segmented mask, where r, c respectively indicate the row and column indices of a boundary pixel p; straight edges of the imaging window are excluded from $\mathcal{P}$. Disjointed boundary sections are then connected via morphological opening to ensure robustness when the anatomy of interest exceeds the field of view.

The boundary extraction process involves three steps: image pre-processing, removal of unwanted edges, and ensuring boundary connectivity.

1. Image pre-processing consists of the following sub-steps:
   (a) Gaussian filtering with $\sigma = 0.5$ to reduce noise
   (b) Thresholding (set to 0) : This threshold may vary based on the dataset. In case of background noise, it may be chosen adaptively or as a percentile of the intensity distribution of an image.
   (c) Morphological cleaning: This involved filling holes and morphological opening to obtain a compact object especially in images with great variance in anatomical contrast.
   (d) Masking: Obtain the largest connected object, i.e., the image mask. Note that a different segmentation strategy could also be employed instead of the steps above.
   (e) Boundary extraction: Obtain the boundary of the mask by subtracting the mask from its dilated version. This is followed by skeletonizing the boundary.
2. Removal of straight edges corresponding to the imaging window is required for mammography, since the anatomy of interest is bound by the straight edges of an imaging window. It is performed as follows:
   (a) The top horizontal edge and the vertical side edge are both eliminated by identifying the co-ordinates corresponding to the mode of the boundary co-ordinates in this region.
   (b) Any small floating edges are removed.
3. The last step is to ensure connectivity of the extracted border. If multiple skeletons are still present, binary morphological closing and skeletonization are recursively performed for increasingly greater breaks in skeletons. Eventually, the largest fully connected skeleton is retained.

Feature construction and fine-tuning The boundary points in $\mathcal{P}$ are ordered by tracking from the top-left and along the anatomical curvature, resulting in a vector $\mathbf{b} = (b_1, b_2, ..., b_K)$. The process starts at the top-left boundary point ($b_1 = \min_{r,c} P$) and follows two rules: (i) b_{k+1} lies within at most the 5×5 neighborhood of the current point b_k, and (ii) it has the smallest angular gradient relative to b_k and b_{k-1}. The tracking terminates when no boundary point is found in the neighborhood, with optional truncation to exclude chest wall regions.

The angular trajectory $\mathbf{a} = (a_1, a_2, ..., a_{K-1})$ is computed from the ordered boundary points as $a_k = \angle(b_k, b_{k+1}) = \tan^{-1}\frac{b_{k+1,r} - b_{k,r}}{b_{k+1,c} - b_{k,c}}$. The angular gradient vector $\mathbf{a}'$ is then calculated to capture changes in anatomical shape.

Due to the varying number of boundary points for different breast sizes and shapes, a binning procedure needs to be performed on the angular gradient vector. To normalize for variations in size, $\mathbf{a}'$ is binned into N equal bins (chosen empirically) and summarized using the cumulative sum (cusum) of angle gradients within each bin. This representation aggregates local shape variations while maintaining global correspondence between bins and anatomical regions, regardless of size differences. Each image, patient or synthetic, is represented as a N-dimensional vector in this feature space, which preserves the details of shape for downstream analysis. Note that increasing the bin count captures more local effects, while reducing the bin count emphasizes more global effects. We empirically selected $N = 64$ as an optimal compromise between the two factors. However, this parameter can be refined by users based on their prior knowledge of the artifact sizes they aim to capture, providing flexibility to tailor the method to different use cases.

The feature representations from the previous step are summarized per image to capture typical shape variation rates in anatomical shape. For example, the sharp variation associated with the presence of a nipple is expected to occur only once in an image and any greater rate of occurrence might be indicative of an artifact. The 64-d bin-wise representation is then mapped to a 16-dimensional vector (per-image) as follows. Histogram bin boundaries are determined from the 1st (p_1) and 99th (p_{99}) percentiles of the

patient data distribution. For a target dimensionality of $B = 16$ bins, 15 equally spaced internal bin edges are generated between p_1 and p_{99}. The complete set of bin edges is defined as

$$\mathcal{E} = [-\infty, e_1, e_2, \ldots, e_{15}, +\infty], \quad (1)$$

where the outer bins capture values below the 1st percentile and above the 99th percentile, respectively. Once computed, these bin edges remain fixed.

The same edge set $\mathcal{E}$ is then applied to both the real and synthetic datasets. For an image with angle-gradient values $\mathbf{a}' = [a_1, \ldots, a_N]$, a histogram is constructed by counting the number of values falling into each bin,

$$h_b = \sum_{k=1}^{M} \mathbf{1}(a_k \in \text{Bin}_b), \qquad b = 1, \ldots, 16, \quad (2)$$

where $\mathbf{1}(\cdot)$ is the indicator function.

To remove dependence on the number of samples per image, the histogram is normalized by its total count,

$$\tilde{h}_b = \frac{h_b}{\sum_{j=1}^{16} h_j}. \quad (3)$$

The final feature representation is therefore a normalized 16-dimensional distribution vector,

$$\tilde{\mathbf{h}} = [\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_{16}], \quad (4)$$

Because the bin boundaries are learned solely from the patient dataset and subsequently applied to both datasets, the resulting feature vectors are expressed in a common reference space, enabling meaningful comparison between real and synthetic distributions.

Since the length of the breast trajectory is strongly correlated with breast area (Pearson's R = 0.9), and we have observed synthetic data do not necessarily maintain the same distribution of area as the paired patient data, binning and per-image summarization obtain a feature vector of the same length (N) from each image ensuring robustness to the effects of varying area.

2.2.2 Comparison with radiomics shape features

To investigate the ability of our proposed feature space to detect shape anomalies compared to other shape descriptors, we evaluated our anomaly detection pipeline using a standard set of shape descriptors from the PyRadiomics library (van Griethuysen et al., 2017) which included Elongation, Major Axis Length, Maximum Diameter, Mesh Surface, Minor Axis Length, Perimeter, Perimeter Surface Ratio, Pixel Surface, and Sphericity. Our masking pipeline was used to extract the mask of the breast region for each image and the feature vector was computed on the mask using the default feature extractor module in Pyradiomics. The extracted features were scaled to zero mean and unit standard deviation.

2.2.3 Artifact Detection

For detecting network-induced shape artifacts, we employ an isolation forest (iForest), an established unsupervised anomaly detection algorithm (Liu et al., 2008). Although iForest has been used in various applications (Al Farizi et al., 2021), its application towards identifying network-induced artifacts in synthetic medical images is novel.

Specifically, from the patient dataset (X) as represented in the feature space described in Section 2.2.1, a subsample of a dataset (X') is selected. Next, X' is recursively partitioned over a random subset of its features to construct a tree T until each observation is isolated. Several such isolation trees are constructed and together they constitute an isolation forest. The number of trees (100) and the subsampling size (256) are set to the default values (Liu et al., 2008) but can be adjusted as desired for different applications. The isolation forest yields an anomaly score for each observation based on the average number of partitions required to isolate it across the forest. This score is determined by factors such as the path length to isolate an observation, the average path length over the isolation trees, and the subsampling size. Negative scores indicate outliers, while positive and near-zero scores correspond to inliers. The isolation forest trained on the patient dataset is then employed for prediction on the synthetic dataset. Thus, each synthetic image receives a score, and a rank based on this score.

The ability of the proposed method to detect shape artifacts in patient data was explored by using the fitted isolation forests to make predictions on the corresponding patient datasets.

2.2.4 Ablation study

An ablation study was performed to investigate the impact of the bin size on the distribution of anomaly scores computed on the synthetic datasets. Feature extraction was performed using bin sizes of N=128 and N=256 after which anomaly detection was performed on the extracted features using the isolation forest approach outlined above. We evaluated Spearman rank correlations between the anomaly ranks obtained using these values of N and our selected N=64. The

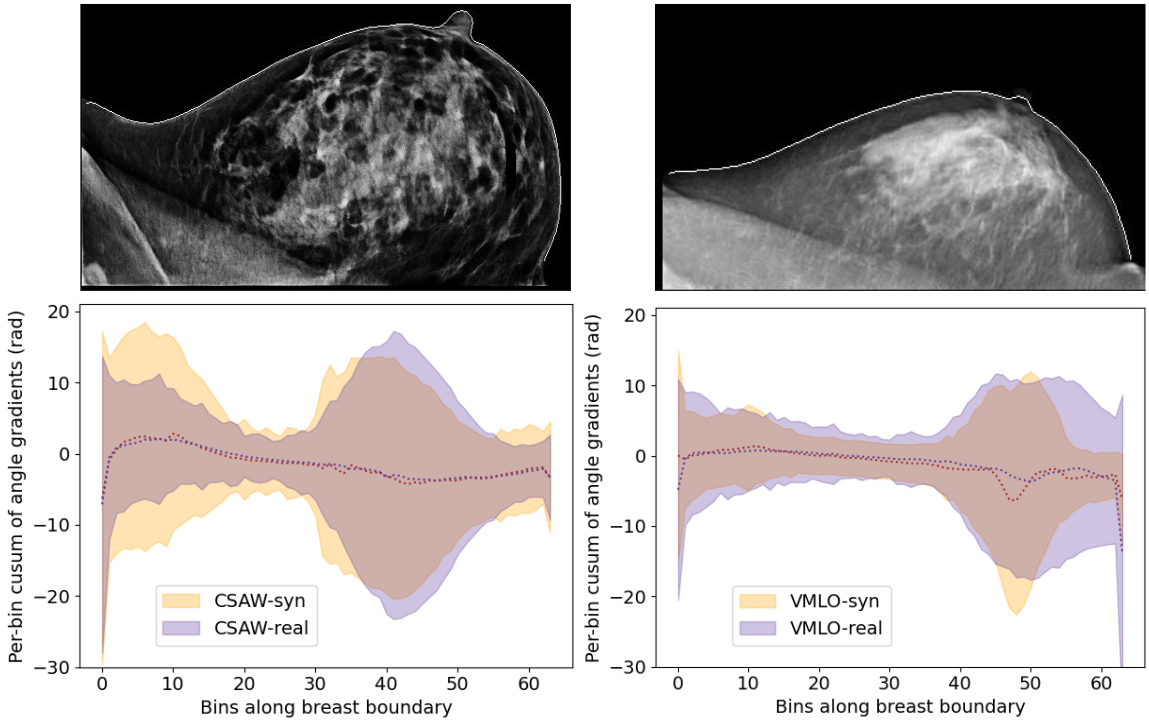


Figure 5: Distributions of bin-wise cumulative sum of angle gradients (blue dotted line: patient data mean, red dotted line: synthetic data mean, shaded region: one std) show substantial but incomplete overlap between patient and synthetic datasets. Left-to-right bins represent breast shape from top to bottom.

bins-size can be selected by the user in a task-specific manner.

The code for feature extraction and anomaly detection using the proposed method is available on Github (Deshpande et al., 2025b).

3. Results

We start by quantitatively comparing the synthetic datasets with their patient data counterparts. We then present results from the feature extraction process, as well as three sets of results from the proposed shape artifact detection method: dataset-level, image-level, and reader study results.

The Fréchet Inception Distance (FID) scores (Heusel et al., 2017) for the two synthetic datasets are 22 (CSAW-syn) and 43 (VMLO-syn) when compared to their paired patient datasets, CSAW-real and VMLO-real respectively. The Inception Scores (IS) (Salimans et al., 2016) for the patient datasets are 3.00 for VMLO-real and 2.30 for CSAW-real wheras the IS for the synthetic datasets are 2.17 for CSAW-syn and 2.16 for VMLO-syn. FID and IS scores lie within ranges reported in published literature for synthetic medical images, though these metrics do not constitute a clinical image quality assessment. FID scores may vary based on the generative model and its optimization (Lucic et al., 2018), and in case of synthetic medical images and mammography images, these values have been reported to lie in the range of 10-50 (Saragih et al., 2024; Oyelade et al., 2022; Rai et al., 2024).

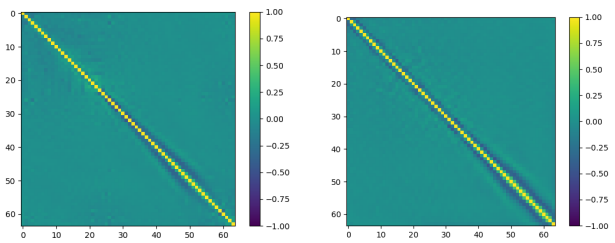
We also computed Fréchet Radiomic Distance (FRD) scores (Konz et al., 2026) for the datasets in which distributional distance is computed based on extracted radiomic texture features. The FRD scores for the two synthetic datasets are 27.9 (CSAW-syn) and 62.7 (VMLO-syn) when compared to their paired patient datasets. Being a distance metric like FID, lower FRD scores indicate greater similarity between the synthetic and the real data. For synthetic Breast MRI images, FRD scores have been reported to be within the 20-50 range (Konz et al., 2026), although FRD primarily captures texture features, which may or may not be correlated with shape similarity between two datasets. Like the FID metrics obtained of the VMLO-syn and CSAW-syn, the FRD scores also seem to indicate greater distributional difference for the VMLO-syn compared to the CSAW-syn.

However, these findings illustrate the limitations of dataset-level metrics, such as FID, in reliably detecting shape artifacts in individual images, emphasizing the need for more granular evaluation methods.

3.1 Feature Extraction Results

Figure 5 shows the distribution of the bin-wise cusum of angle gradients in synthetic and real datasets. The dotted lines represent the bin-wise mean, while the shaded regions indicate one standard deviation for each bin. The bins along the X-axis correspond to different sections of the breast boundary, arranged sequentially from top to bottom of the breast region. These distributions reflect the typical breast shape for a dataset. Specifically, the early bins correspond to the chest wall, followed by low-variance bins representing the breast region up to the nipple. High standard deviation in the right half indicates the nipple region, while final low-variance bins correspond to the lower breast boundary.

In Figure 5, we observe that the distributions of the two *patient* datasets are different, despite exhibiting similar trends across bins. The differences may arise from the distinct patient populations in the Swedish (CSAW-real) and Vietnamese (VMLO-real) datasets. Next, the patient and synthetic datasets show substantial, but not perfect overlap for both CSAW-syn and VMLO-syn, suggesting that breast shape is not entirely preserved in the synthetic datasets and that anomalous images may be present. Notably, in the VMLO-syn dataset, the bin-wise means (dotted lines in Figure 5) are clearly distinct from the corresponding patient dataset, indicating a bias in the synthetic dataset toward a specific breast shape. Despite the observed difference in mean between the VMLO-syn and VMLO-real, particularly between bins 40 and 60, the mean values for the VMLO-syn are still within the standard deviation envelope of the real data. As a result, the shape bias in the VMLO-syn images is unlikely to contribute to their isolation as anomalous samples when processed by the detection pipeline.



(a) CSAW-real (b) VMLO-real

Figure 6: Cross-correlation matrices of patient images in the proposed feature space indicate only moderate correlation among some neighboring feature dimensions, and are largely diagonal otherwise.

3.1.1 Feature similarity between patient and synthetic datasets

For the proposed 64-dimensional feature space, the cross-correlation of the feature vectors of the patient datasets are shown in Figure 6. For each patient dataset, 64-dimensional shape features were extracted per-image after which the cross-correlation matrix was generated by computing the correlation of each feature with every other feature across the dataset. Note that neighboring components demonstrate high correlations as they are constrained by the breast shape, but the level of correlation drops rapidly for bins that are farther apart. Thus, conventional dimensionality reduction methods (e.g., principal component analysis) do not provide substantial improvement in representational efficiency at this stage.

We quantified the differences between patient and synthetic distributions as follows. Approximately 9,000 samples were selected from both patient and synthetic datasets. For each patient sample, we calculated the Mahalanobis distance relative to the mean and covariance of the patient dataset, creating a distribution of distances. The same procedure was applied to all synthetic samples, using the mean and covariance of the patient dataset. To compare these distributions, we performed a Kolmogorov-Smirnov (KS) test, yielding the following KS-statistics: 0.068 (p-value < 0.0001) for CSAW, and 0.13 (p-value < 0.0001) for VMLO. This indicates that patient and synthetic distributions have distinct distributions and anomalous images are present in this feature space.

3.2 Dataset-Level Results

The distributions of anomaly scores for both synthetic datasets are shown in Figure 7. Anomaly scores from the isolation forest are bounded between -1 and 1 on the X-axis. Decreasing scores along the negative axis (from 0 to -1) correspond to increasingly anomalous images. In contrast, positive values and values close to zero signify normal images or those with minimal network-induced shape artifacts. Both distributions are left-tailed, suggesting that only a small fraction of the synthetic dataset contains highly anomalous shapes. Thresholds corresponding to the 1^{st} and 10^{th} percentiles of the most anomalous images are marked in Figure 7. Three quantile partitions are defined in the distribution of artifact scores, categorized as follows: P1 (1^{st} percentile) represents the images with the most obvious or pronounced network-induced shape artifacts, P2 (1^{st}–10^{th} percentile) includes images with moderate network-induced shape artifacts, and P3 (10^{th}–100^{th} percentile) corresponds to images with

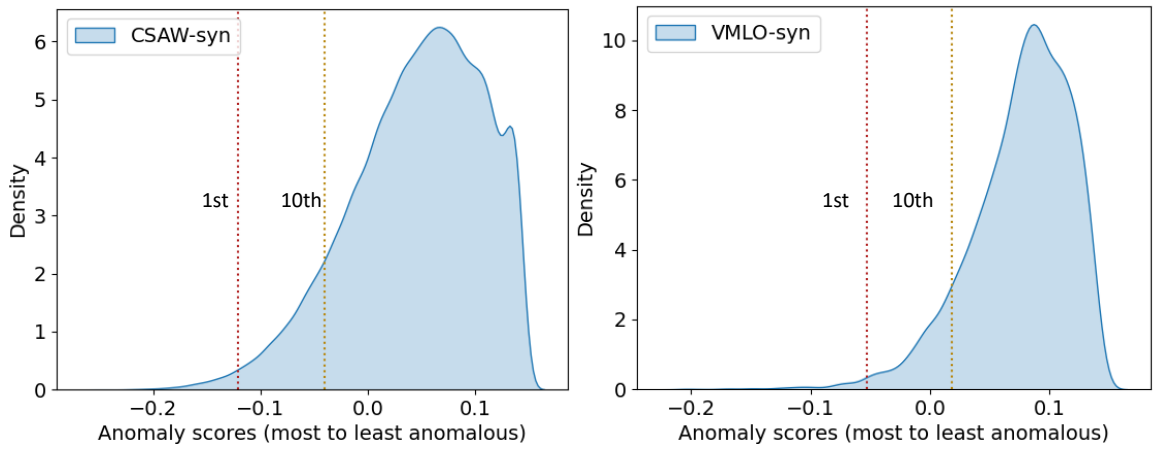


Figure 7: Distribution of anomaly scores for the two synthetic datasets. 1^{st} and 10^{th} percentile are marked.

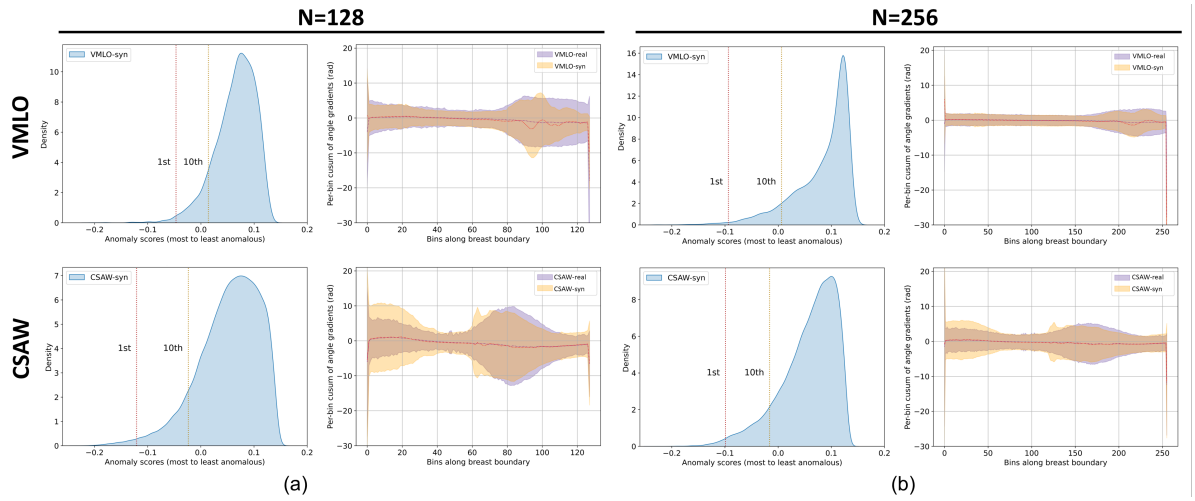


Figure 8: Impact of bin size on anomaly scores and distributions of bin-wise cumulative sum of angle gradients for (a) N=128 (b) N=256. 1^{st} and 10^{th} percentiles are marked in the anomaly score distribution plots.

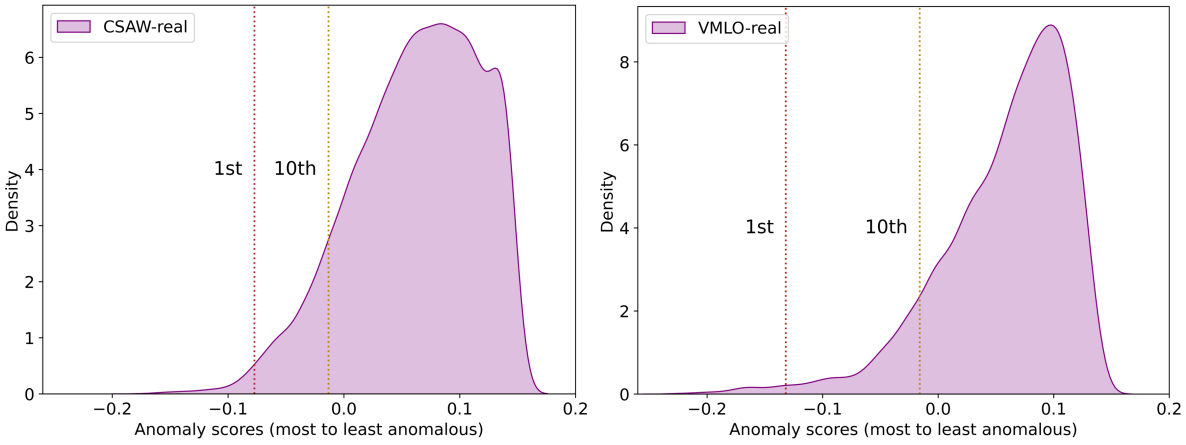


Figure 9: Distribution of anomaly scores for the two patient datasets. 1^{st} and 10^{th} percentile are marked.

minimal or no visible shape artifacts. The 10^{th} percentile boundary of CSAW-syn being below zero and the one for VMLO-syn being above zero likely indicates that the VMLO-syn dataset has fewer strongly anomalous images compared to CSAW-syn. The categorization of anomaly scores into three partitions is a reasonable starting point when the prevalence of anomalies in a dataset is unknown. This allows us to

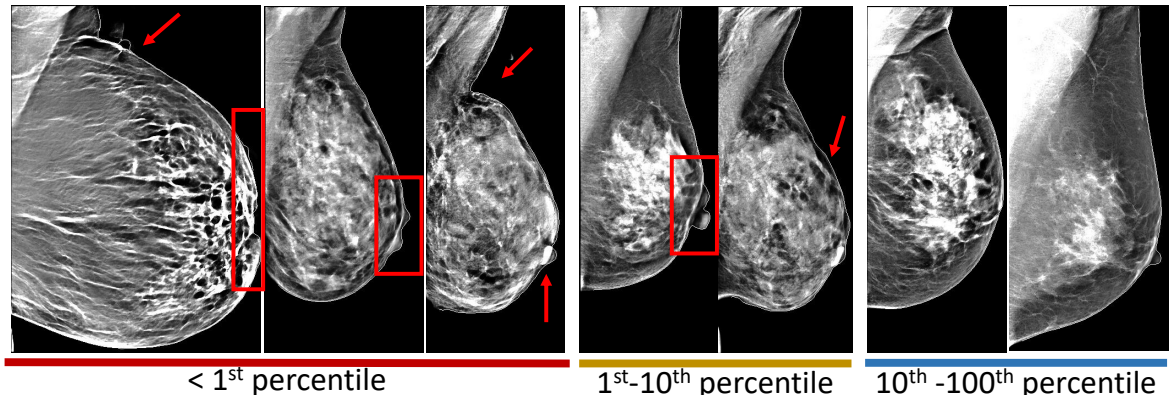


Figure 10: Most to least (L-R) artifact images from CSAW-syn as ranked by our method. Annotations in red are for display only.

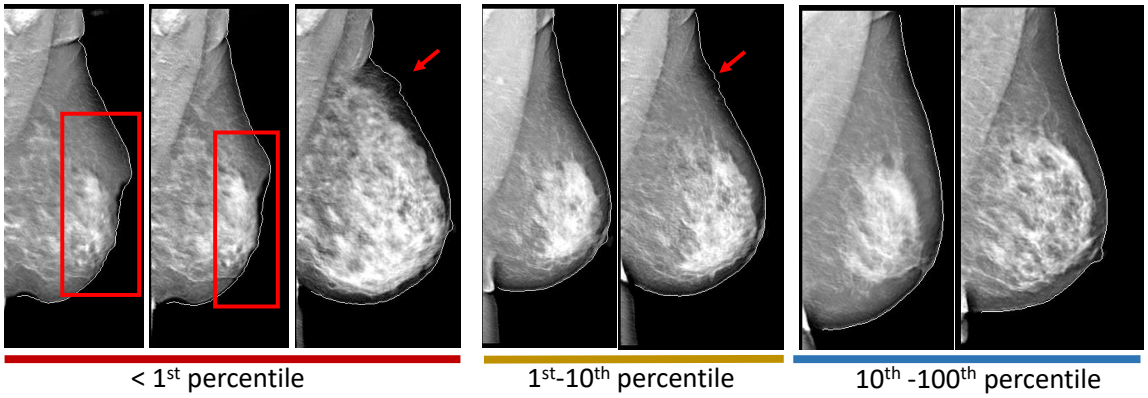


Figure 11: Most to least (L-R) artifact images from VMLO-syn as ranked by our method. Annotations in red are for display only.

sample images from the partitions and manually validate the presence of artifacts. The left-tailed nature of these distributions means that the 1^{st} and 10^{th} percentiles may typically correspond to extreme and moderate artifacts, but these are not fixed bounds. These partitions can be further customized and fine-tuned given prior knowledge of the task or the approximate rate of images with artifacts.

The impact of varying bin sizes on bin-wise feature distribution and anomaly detection is shown in Figure 8. The 1^{st} and 10^{th} percentile thresholds for both datasets are largely consistent for each dataset for bin sizes N=128 and N=256 compared to the chosen N=64 and the shape of the distributions broadly mirror that of N=64. Increasing the bin size effectively captures shape features with higher local resolution for detecting errors. For both datasets, Spearman's rank correlations with our chosen N=64 decreased with increasing N (VMLO-syn: $\rho = 0.62$ for N=128 and $\rho = 0.44$ for N=256; CSAW-syn: $\rho = 0.80$ for N=128 and $\rho = 0.61$ for N=256), suggesting that larger bin sizes shift the method toward capturing finer-grained local features, which alters image rankings. This is consistent with the intended design flexibility: users may select N based on the scale of artifacts they aim to detect.

To assess the robustness of our proposed method to shift in breast area distributions, each patient-synthetic dataset pair was partitioned according to the quartiles of the breast area distribution in the patient dataset. Distinct partitions in the synthetic dataset were created by matching the partition thresholds determined from the patient dataset. The proposed method was then individually employed on each matched patient-synthetic partition to obtain anomaly scores. An overall anomaly ranking was obtained over the entire dataset based on the anomaly scores from all partitions. It was observed that this global ranking obtained over the area-wise application of the method had a strong correlation (Spearman's $\rho = 0.93$ for both datasets) with the rankings obtained from employing the method on the entire dataset at once.

The distributions of anomaly scores for both patient datasets are shown in Figure 9. Like their synthetic data anomaly score distribution counterparts, both distributions are left-tailed. The distribution for CSAW-real is broader, compared to the more compact

VMLO-real, suggesting a wider variety of shapes.

The results from our anomaly detection experiment with radiomics shape descriptors are seen in Figure 13. The distributions of anomaly scores for CSAW-syn and VMLO-syn in Figure 13 (a) and (b) are left tailed and overall similarly structured as the ones obtained using the proposed method. The distribution obtained for CSAW-syn has a sharper and higher peak compared to the one obtained using the proposed method and shows indication of bi-modality with a small peak between the 1^{st} and 10^{th} percentile thresholds.

3.3 Image-Level Results

Figure 10 and Figure 11 present examples of hand-selected images from each partition as ranked by the proposed method for CSAW-syn and VMLO-syn, respectively. As shown in the figures, images in P1 ($\leq 1\%$) have visible shape artifacts. In P2 (1%-10%), CSAW-syn shows clear artifacts, while VMLO-syn exhibits minor distortions. In P3 (10%-100%), both datasets display well-formed breast shape without visible artifacts.

The figures also highlight that each synthetic dataset seems to exhibit different types of artifacts. In CSAW-syn, artifacts are primarily local, such as multiple nipples, poorly formed nipple regions, and sharp chest wall angles. In contrast, VMLO-syn shows more global artifacts, with visibly malformed breast shape (below the 1st percentile in Figure 11), alongside minor local artifacts (1st-10th percentile) like non-smooth or angular boundaries. These differences may arise from a variety of sources including the type of generative model used to generate the dataset (diffusion model for CSAW-syn and StyleGAN2 for VMLO-syn), the source patient dataset, or even the training process itself. In this work, we focus on the ability to detect these artifacts, regardless of the source.

The most anomalous images identified within the patient data highlighted cases of imaging issues such as inadequate tissue coverage and missing nipples, as well as atypical breast outlines which can occur due to surgical scars. Examples of these natural shape artifacts are shown in Figure 12.

For the synthetic datasets, while some artifacts in P2 resembled the most anomalous images in the patient datasets, artifacts typically observed in P1 were not observed in patient data, and thus, originated from the generative process itself. This confirms that the method can identify artifacts not present in the training dataset. It is important to note that the method can detect these network-induced artifacts without requiring prior knowledge of the anatomy nor any provided labels.

Figures 13 (c) and (d) highlight hand selected sample images from the different anomaly rank partitions obtained using the radiomics shape descriptors. For both datasets, images in P1 ($\leq 1\%$) have visible shape artifacts. However, unlike the proposed method, the incidence of clearly observable network induced artifacts drops sharply in P2 (1%-10%), where samples from both CSAW-syn and VMLO-syn exhibit minor distortions or mostly distortion-free images. In P3 (10%-100%), both datasets display well-formed breast shape without visible artifacts. This indicates that while other general-purpose shape descriptors may be able to isolate the more extreme shape artifacts, the proposed method may be able to additionally isolate subtle ones leading to a more useful ranking of the images in the overall dataset.

3.4 Results from the Reader Study

A 2-Alternative Forced Choice (2-AFC) reader study was conducted separately for each synthetic dataset using the SimplePhy tool (Lago, 2021). Three non-clinical medical imaging researchers participated as readers. Readers were presented with pairs of images and asked to select the one they considered to have the most anomalous shape. An example screenshot from the study interface is shown in Figure 14. The viewing settings were carefully calibrated for each reader to ensure consistency, and each experiment was completed in a single reading session.

Thirty images from each synthetic dataset were selected for the reader study, with ten images drawn from each of the three partitions: $\leq$1%, 1-10%, and 10-100%. This sampling ensures sufficient representation of images that were ranked most anomalous by the algorithm (tail of the anomaly score distribution). All image-pair combinations of the thirty selected images, including image-pairs within the same partition, which resulted in a total of 435 trials, were shown to the readers as illustrated in Figure 14. Readers were instructed to select the image with stronger shape artifacts from each side-by-side comparison.

Results are summarized in Table 2 as the percentage of trials (mean$\pm$std) in which an image was identified as having the most pronounced shape artifacts, reported for each partition. For both datasets, the first partition ($\leq$1%) was consistently identified as containing the most shape artifacts by all readers, with mean values approximately 1.5-2 times higher than the last partition (10-100%). This indicates that shape artifacts were effectively concentrated in the first partition, greatly improving search efficiency compared

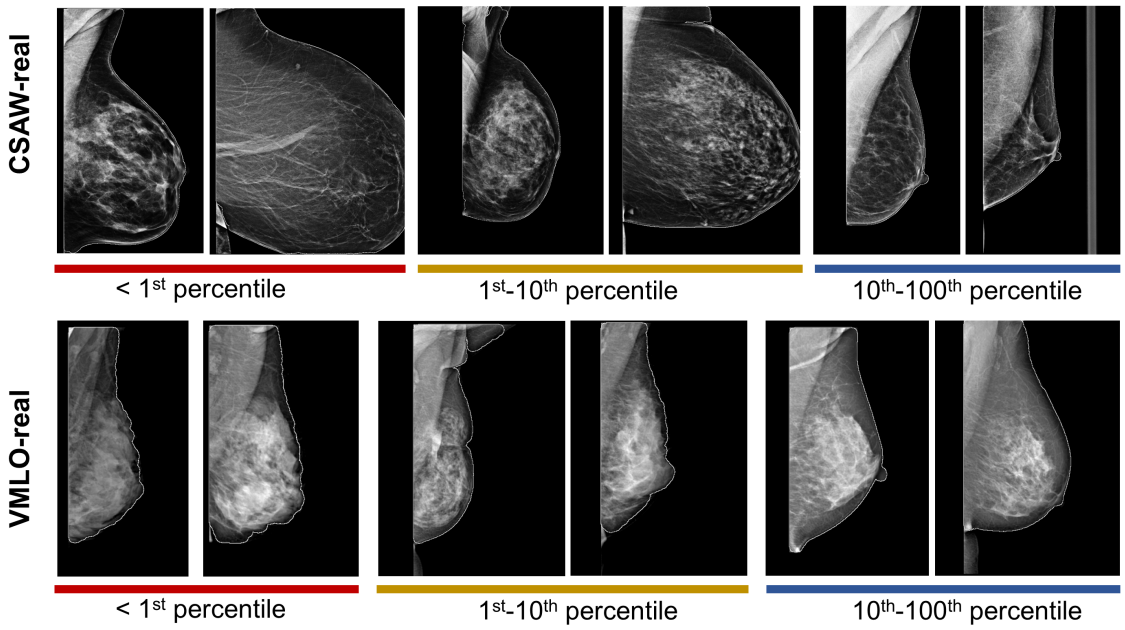


Figure 12: Examples of images identified as anomalous when the proposed method is applied to patient datasets. The first row shows CSAW-real, and second row shows VMLO-real. Practically observed, yet unusual phenomena, such as inadequate tissue coverage, skin folds, missing nipples, etc., can be observed in the identified images.

to random visual searches. Further, the similar mean values for the first partition (≈66%) suggest that all readers consistently found artifacts in this partition to be most distinctive compared to other partitions.

Irrespective of the presence of artifacts, if images lying in the same partition are highly similar, a consistent *intra-partition* ranking cannot be obtained. In our reader study, 10 images are chosen from each partition. Thus, each image is compared with 9 images from the same partition and 29 images over all three partitions. If intra-partition similarity is extremely high, an image is chosen 50% of the times (equal probability in a 2-AFC) in all its intra-partition trials. This translates to its selection at a rate of about 17% (1/6) on average over the experiment. This is computed as follows:

$$\text{Image chosen (\%)} = P_{\text{random}} \times \frac{n_{\text{image}}}{N_{\text{total}}} \tag{5}$$

where:

- P_{random}: the intra-partition probability of random selection of an image.
- n_{image}: the intra-partition trial count for the specific image.
- N_{total}: the total number of trials conducted in the experiment for the specific image.

In our case:

$$\text{Image chosen (\%)} = \frac{1}{2} \times \frac{9}{29} \tag{6}$$

Thus, an image will be chosen at least 17% of the times and not more than $100 - 17 = 83\%$ of the times on average if intra-partition similarity is extremely high. The 17% selection rate will be lower (and the corresponding upper-bound will be higher) if images with varying degrees of artifacts are present in the same partition.

Additionally, mean values decreased monotonically across the three partitions for all readers in both datasets, confirming that successive partitions contained fewer shape artifacts. This is in contrast to the equal mean value (≈50%) for all partitions expected if images were chosen at random. Further, the second partition is only slightly higher than the third partitions but less so for the VMLO-syn dataset, reflecting a lower fraction of shape artifacts in VMLO-syn compared to CSAW-syn. This is supported by the score distributions shown in Figure 7, where the second partition in CSAW-syn contains negative scores, while that in VMLO-syn straddles zero.

Kendall-Tau correlations (Kendall, 1938) between reader rankings and the algorithm rankings over all images in the reader study were 0.45 for CSAW-syn (reader values: 0.43, 0.51, 0.40) and 0.43 for VMLO-syn (reader values: 0.33, 0.42, 0.55), indicating reasonable agreement between the two. Note that high Kendall-Tau values are *not* expected due to low visual distinguishability between images with close rankings.

Finally, the AUC (area under the curve) values

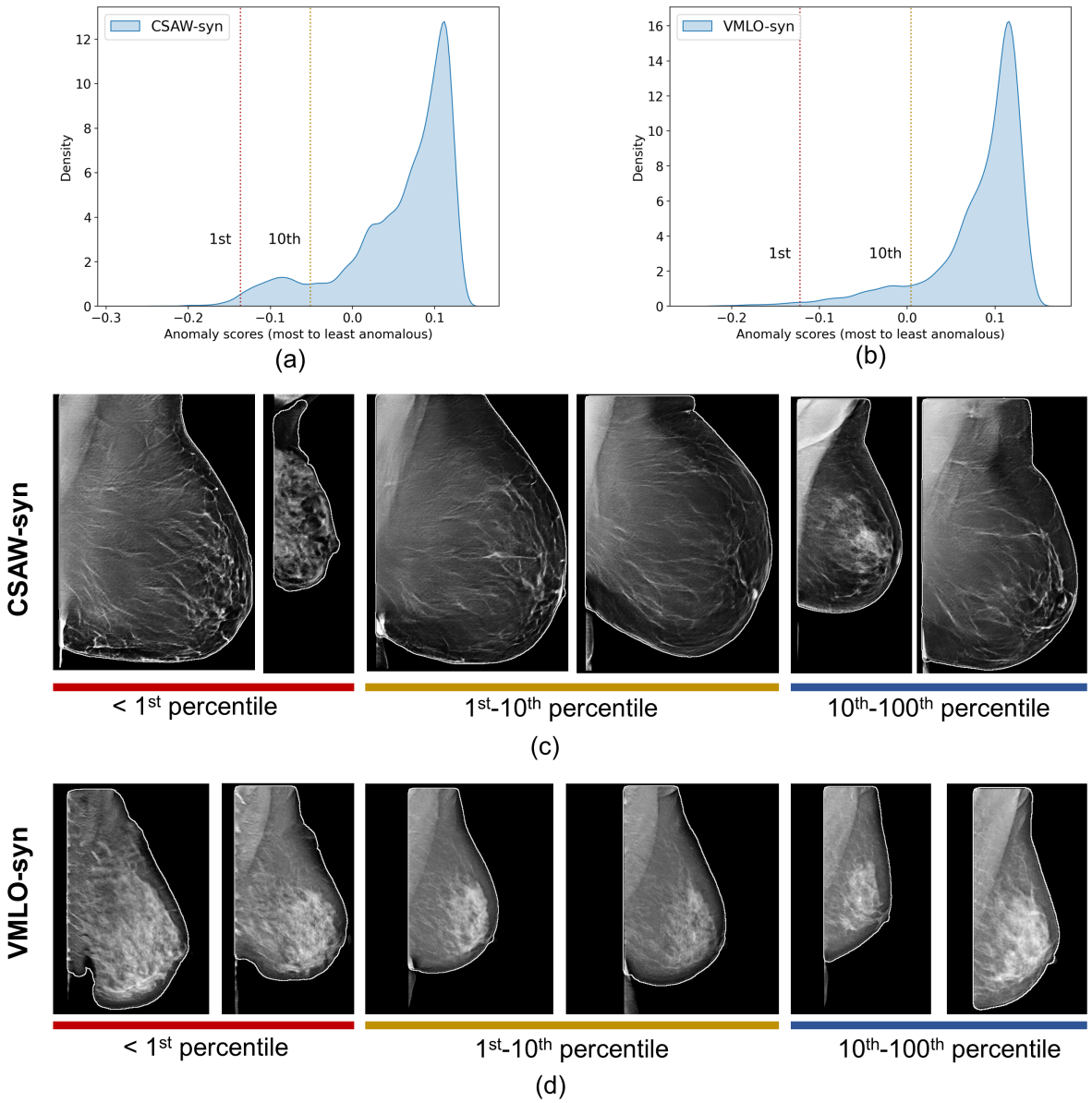
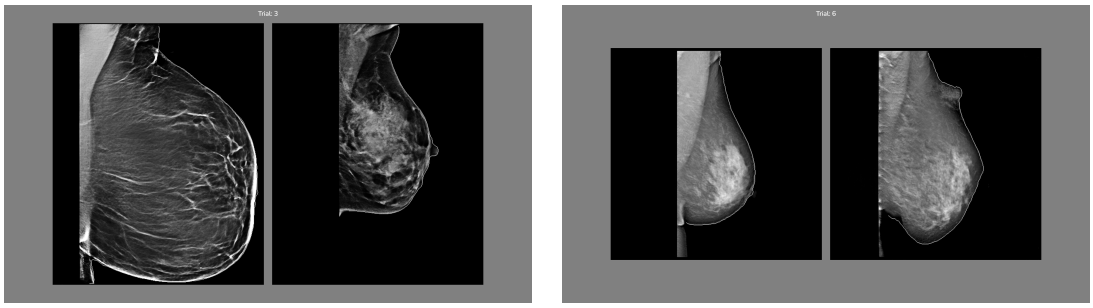


Figure 13: Results for anomaly detection using radiomics shape descriptors. Distribution of anomaly scores for (a) CSAW-syn (b) VMLO-syn. Most to least (L-R) artifact images from (c) CSAW-syn and (d) VMLO-syn as ranked through anomaly detection with radiomics shape descriptors.

were computed according to (Liu et al., 2008; Hand and Till, 2001) based on the anomaly rankings within the set of the images employed in the reader study, where true anomalies were defined based on the mean values from the reader study. The resulting AUC was 0.97 (CSAW) and 0.91 (VMLO), where true anomalies were defined based on reader consensus.

4. Discussion and Conclusion

We propose a knowledge-based method to detect network-induced shape artifacts, and demonstrate its use on synthetic mammograms. While similar knowledge-based approaches for characterizing shapes have been used in tasks such as kidney stone classification (Duan et al., 2013) and nipple localization in mammograms (Zhou et al., 2004), this is the first application of such an approach to synthetic data for detecting network-induced artifacts. Breast shape is associated with breast density, architectural distortions, and demographic features (Gaur et al., 2013; del Carmen et al., 2007), and unrealistic breast shape may potentially affect downstream model performance; however, the clinical impact was not assessed in this study and warrants future investigation. In addition, our method may potentially be adapted to other medical image modalities and shapes using domain knowledge, and hence highly versatile.



(a) CSAW-syn dataset pair (b) VMLO-syn dataset pair

Figure 14: Examples from the reader study conducted for the CSAW-syn and VMLO-syn datasets. Readers were tasked with evaluating pairs of images and selecting the image with more pronounced abnormal shapes. (a) The left panel shows an example from the CSAW-syn dataset with a P1 image on the left and a P1 image on the right. (b) The right panel displays an example from the VMLO-syn dataset with a P2 image on the left and a P1 image on the right.

Although research on characterizing breast tissue is extensive (Gastounioti et al., 2016), methods for evaluating breast shape fidelity are limited. Most shape variation in mammography occurs along the breast curvature, while the straight edges of the imaging window show minimal variation. Conventional shape features like compactness or area often miss anatomical inaccuracies along the curvature, making shape artifacts in synthetic mammograms difficult to detect. Our proposed method addresses these challenges effectively.

Our approach is designed to be widely applicable to synthetic datasets for detecting network-induced shape artifacts, even in the absence of artifact labels. Compared other shape descriptors, such as the radiomics shape descriptors, our method provides greater control over the local-global tradeoff in identifying anomalies. It provides rankings instead of binary decisions on artifact presence, which can improve the efficiency of visual search for artifacts, reducing the burden on expert readers. These rankings can effectively separate images with artifacts from normal images and users can adjust the threshold post-hoc according to their specific requirements, enabling flexible and tailored detection for research and quality-assurance evaluation purposes. Validation for any specific clinical or regulatory submission context would require additional testing appropriate to that intended use.

Consider a synthetic dataset of 10,000 images of which 5% contain artifacts. A random sample of 100 images would be expected to yield 5 images with artifacts on average. That is, the rate of artifact discovery is 1 in 20 ($= 0.05$). Practically, the small sample size (100) may result in even fewer, or no artifacts being discovered. However, if the proposed method is employed with an accuracy of 70% (highest mean values for both datasets were about 67%), the same sample would contain about 70 images with artifacts. The new rate of artifact discovery would be 7 in 10 images ($= 0.7$). Thus, the rate of artifact discovery would be improved 14 times (new rate/ old rate $= 0.7/0.05$) over random spot checking. More generally, the rate improves as a factor of $\mathrm{accuracy}\ (\%)/\mathrm{artifact\ prevalence}\ (\%)$, and may differ as the two factors vary.

In research settings, this method may serve as a starting point for dataset annotation and quality review. Its use in preparing datasets for specific clinical or regulatory applications would require additional validation appropriate to those contexts.

While we provide an example with breast imaging, the proposed method may be generalized to other anatomies and imaging modalities where the region of interest can be segmented, such as lungs, abdomen, brain, or any other anatomical region of interest. This broad applicability is due to the reliance of the method on the characteristic shape of the anatomy. It is particularly valuable in scenarios where anatomy cannot be described adequately by area-based features or conventional compactness and convexity features.

In lung imaging via chest radiographs, for example, the method may be applied to detect artifacts in the shape and contour of the lungs from synthetic data. By analyzing the boundary of the lung, we can extract the angle gradients along the lung border. These gradients capture the curvature and shape of the lung, which are necessary for detecting any anomalies caused by synthetic artifacts. In this context, the method can potentially identify unusual shapes that do not align with the expected lung morphology, such as irregularities in lung shape or distortions in the pleural surface, which may indicate network-induced artifacts.

Similarly, in abdominal computed tomography (CT), the method can potentially be applied to detect syn-

Table 2: Results from the reader study demonstrate that the images ranked worst (P1) by our method were also chosen as the worst by all readers. Highest values in bold.

Dataset	CSAW-syn			VMLO-syn		
Reader/Partition	P1	P2	P3	P1	P2	P3
Reader 1	**65±9%**	48±12%	37±9%	**63±9%**	45±9%	41±12%
Reader 2	**68±7%**	47±9%	34±8%	**67±9%**	42±11%	41±17%
Reader 3	**64±7%**	47±13%	39±7%	**73±8%**	42±8%	34±14%
Mean of means	**66%**	47%	37%	**68%**	43%	39%

thetic artifacts in shape for various organs such as the liver and the kidneys. By analyzing the contours of these organs and calculating the angle gradients along their boundaries, we can detect unnatural shapes, such as deformed or overly smoothed organ outlines, which might result from flaws in the synthetic data generation process. The method could flag these artifacts by identifying discrepancies in the natural shape of abdominal organs. Another example could be characterizing brain hemorrhages in computed tomography, which have characteristic shapes based on type, and assessing synthetic tissue lesions, which may have characteristic shapes based on the presence of malignancy. The proposed method can effectively capture such shape variations specific to anatomical structures. In the future, we plan to extend our work to synthetic datasets from other medical imaging modalities in a domain-relevant manner.

In the scenarios outlined, additional domain knowledge required to adapt our method is related to the masking or organ segmentation step. This highlights a limitation of this work since the method relies on a continuous boundary from a prior segmentation, and we did not analyze the impact of different segmentation methods. In low-quality synthetic datasets, network-induced background artifacts may negatively impact segmentation methods. While this was not observed in our work, it may be relevant for other synthetic datasets. To identify shape artifacts when constant thresholding may be unreliable for segmentation due to background artifacts in synthetic data, an adaptive threshold can be selected for segmentation. This threshold could be chosen as a percentile from the pixel intensity distribution in an image, in a user-informed manner, to ensure consistency with the patient data. Moreover, the task of medical image segmentation is rapidly becoming more accessible due to the rise in availability of powerful general-purpose medical image segmentation foundation models (Ma et al., 2024; Zhang et al., 2025) reducing the impact of this hurdle. Once the object boundary is obtained, the feature-space construction through boundary tracking and subsequent anomaly detection is domain agnostic.

A further consideration is the size of shape artifacts. While our method is agnostic to artifact size, prior knowledge about artifact scale or type can be incorporated into the algorithm through bin counts or relative bin locations. Another limitation of our work is that the clinical impact of the identified artifacts was not studied. In the future, a reader study conducted by radiologists on artifacts in synthetic data as well as natural shape artifacts (see for instance Figure 12) could provide valuable clinical insights. Additionally, the present work focuses on breast shape rather than breast tissue. We are currently developing a complementary method for assessing artifacts in breast tissue, which will also require validation by clinical experts. Other future research directions include further exploring the localization of detected artifacts within flagged images, as well as identifying the origin of shape artifacts in relation to the dataset, the model and its optimization.

In conclusion, we have developed a knowledge-based method that efficiently identifies network-induced shape artifacts in synthetic medical images. Being agnostic to the generative process, it is applicable to emerging model architectures and scenarios where artifact characteristics are unknown, making it a useful research tool for evaluating synthetic datasets at scale. Its suitability for operational deployment in specific regulatory or clinical workflows would depend on additional task-specific validation.

Acknowledgments

Rucha Deshpande and Tahsin Rahman acknowledge funding by appointment to the Research Participation Program at the Center for Devices and Radiological Health administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US Department of Energy and the US Food and Drug Administration.

Disclaimer

This article reflects the views of the authors and does not represent the views or policy of the U.S. Food

and Drug Administration, the Department of Health and Human Services, or the U.S. Government. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

Ethical Standards

This work used only publicly available, de-identified datasets. No new human subjects research or animal studies were conducted. All applicable data use terms were observed.

Conflicts of Interest

We have no conflicts of interest to declare. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services. This is a contribution of the U.S. Food and Drug Administration and is not subject to copyright.

Data availability

Both patient datasets used in this work are public datasets. One synthetic dataset (SinKove) is also public. The VMLO-syn dataset was generated in-house and is not publicly available at this time. Inquiries may be directed to the corresponding author, subject to applicable FDA data-sharing policies.

References

Wahid Salman Al Farizi, Indriana Hidayah, and Muhammad Nur Rizal. Isolation forest based anomaly detection: A systematic literature review. In *2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, pages 118–122. IEEE, 2021.

Aldo Badano, Christian G Graff, Andreu Badal, Diksha Sharma, Rongping Zeng, Frank W Samuelson, Stephen J Glick, and Kyle J Myers. Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial. *JAMA Network Open*, 1(7):e185474–e185474, 2018.

Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022.

Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew F K Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.*, 5(6):493–497, June 2021.

Marcela G del Carmen, Elkan F Halpern, Daniel B Kopans, Beverly Moy, Richard H Moore, Paul E Goss, and Kevin S Hughes. Mammographic breast density and race. *American Journal of Roentgenology*, 188(4):1147–1150, 2007.

Rucha Deshpande, Varun A Kelkar, Dimitrios Gotsis, Prabhat Kc, Rongping Zeng, Kyle J Myers, Frank J Brooks, and Mark A Anastasio. Report on the aapm grand challenge on deep generative modeling for learning medical image statistics. *Medical Physics*, 52(1):4–20, 2025a.

Rucha Deshpande, Tahsin Rahma, Elim Thompson, Ghada Zamzmi, and Seyed Kahaki. Shapecheck, 2025b. URL `https://github.com/DIDSR/ShapeCheck`. Accessed 2025-10-24.

Xinhui Duan, Mingliang Qu, Jia Wang, James Trevathan, Terri Vrtiska, James C Williams, Amy Krambeck, John Lieske, and Cynthia McCollough. Differentiation of calcium oxalate monohydrate and calcium oxalate dihydrate stones using quantitative morphological information from micro-computerized and clinical computerized tomography. *The Journal of Urology*, 189(6):2350–2356, 2013.

Fabio Garcea, Alessio Serra, Fabrizio Lamberti, and Lia Morra. Data augmentation for medical imaging: A systematic literature review. *Comput. Biol. Med.*, 152(106391):106391, January 2023.

Aimilia Gastounioti, Emily F Conant, and Despina Kontos. Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. *Breast Cancer Research*, 18:1–12, 2016.

Shantanu Gaur, Vandana Dialani, Priscilla J Slanetz, and Ronald L Eisenberg. Architectural distortion of the breast. *American Journal of Roentgenology*, 201(5):W662–W670, 2013.

Mauro Giuffrè and Dennis L Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digit. Med.*, 6(1):186, October 2023.

David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45: 171–186, 2001.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jiwoong J Jeong, Amara Tariq, Tobiloba Adejumo, Hari Trivedi, Judy W Gichoya, and Imon Banerjee. Systematic review of generative adversarial networks (GANs) for medical image classification and segmentation. *J. Digit. Imaging*, 35(2):137–152, April 2022.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.

Varun A Kelkar, Dimitrios S Gotsis, Frank J Brooks, KC Prabhat, Kyle J Myers, Rongping Zeng, and Mark A Anastasio. Assessing the ability of generative adversarial networks to learn canonical medical image statistics. *IEEE Transactions on Medical Imaging*, 42(6):1799–1808, 2023.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Andrea Kim, Niloufar Saharkhiz, Elena Sizikova, Miguel Lago, Berkman Sahiner, Jana Delfino, and Aldo Badano. S-synth: Knowledge-based, synthetic generation of skin images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 734–744. Springer, 2024.

Lennart R Koetzier, Jie Wu, Domenico Mastrodicasa, Aline Lutz, Matthew Chung, W Adam Koszek, Jayanth Pratap, Akshay S Chaudhari, Pranav Rajpurkar, Matthew P Lungren, and Martin J Willemink. Generating synthetic data for medical imaging. *Radiology*, 312(3):e232471, September 2024.

Nicholas Konz, Richard Osuala, Preeti Verma, Yuwen Chen, Hanxue Gu, Haoyu Dong, Yaqian Chen, Andrew Marshall, Lidia Garrucho, Kaisar Kushibar, et al. Fréchet radiomic distance (frd): A versatile metric for comparing medical imaging datasets. *Medical Image Analysis*, page 103943, 2026.

Miguel A Lago. Simplephy: An open-source tool for quick online perception experiments. *Behavior Research Methods*, pages 1–8, 2021.

Juhun Lee, Tamerlan Mustafaev, and Robert M Nishikawa. Impact of gan artifacts for simulating mammograms on identifying mammographically occult cancer. *Journal of medical imaging*, 10(5): 054503–054503, 2023.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31, 2018.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1): 12098, 2023.

Olaide N Oyelade, Absalom E Ezugwu, Mubarak S Almutairi, Apu Kumar Saha, Laith Abualigah, and Haruna Chiroma. A generative adversarial network for synthetization of regions of interest based on digital mammograms. *Scientific Reports*, 12(1): 6166, 2022.

Hieu Huy Pham, Hieu Nguyen Trung, and Ha Quy Nguyen. Vindr-mammo: A large-scale benchmark dataset for computer-aided detection and diagnosis in full-field digital mammography. *Sci Data*, 2022.

Walter H L Pinaya, Mark S Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F da Costa, Ashay Patel, Hyungjin Chung, Can

Zhao, Wei Peng, Zelong Liu, Xueyan Mei, Oeslle Lucena, Jong Chul Ye, Sotirios A Tsaftaris, Prerna Dogra, Andrew Feng, Marc Modat, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Generative AI for medical imaging: extending the MONAI framework. 2023a. URL `https://huggingface.co/datasets/SinKove/synthetic_mammography_csaw`.

Walter HL Pinaya, Mark S Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F Da Costa, Ashay Patel, et al. Generative ai for medical imaging: extending the monai framework. *arXiv preprint arXiv:2307.15208*, 2023b.

Hari Mohan Rai, Serhii Dashkevych, and Joon Yoo. Next-generation diagnostics: the impact of synthetic data generation on the detection of breast cancer from ultrasound imaging. *Mathematics*, 12(18): 2808, 2024.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Daniel G Saragih, Atsuhiro Hibi, and Pascal N Tyrrell. Using diffusion models to generate synthetic labeled data for medical image segmentation. *International journal of computer assisted radiology and surgery*, 19(8):1615–1625, 2024.

Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*, 34:18126–18136, 2021.

Atif Ahmed Showrov, Md Tarek Aziz, Hadiur Rahman Nabil, Jamin Rahman Jim, Md Mohsin Kabir, MF Mridha, Nobuyoshi Asai, and Jungpil Shin. Generative adversarial networks (gans) in medical imaging: Advancements, applications and challenges. *IEEE Access*, 2024.

E. Sizikova, N. Saharkhiz, D. Sharma, M. Lago, B. Sahiner, J. G. Delfino, and A. Badano. Knowledge-based in silico models and dataset for the comparative evaluation of mammography ai for a range of breast characteristics, lesion conspicuities and doses. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Elena Sizikova, Andreu Badal, Jana G Delfino, Miguel Lago, Brandon Nelson, Niloufar Saharkhiz, Berkman Sahiner, Ghada Zamzmi, and Aldo Badano. Synthetic data in radiological imaging: current state and future outlook. *BJR Artif. Intell.*, 1(1), March 2024.

Moein Sorkhei, Yue Liu, Hossein Azizpour, Edward Azavedo, Karin Dembrower, Dimitra Ntoula, Athanasios Zouzos, Fredrik Strand, and Kevin Smith. Csaw-m: An ordinal classification dataset for benchmarking mammographic masking of cancer. 2021.

U.S. Food and Drug Administration. Ai/ml, September 2024. URL `https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/fda-digital-health-and-artificial\-intelligence-glossary-educational-resource`. Last modified September 26, 2024.

Joost J. M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G. H. Beets-Tan, Jean-Christophe Fillon-Robin, Steve Pieper, and Hugo J. W. L. Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21): e104–e107, 2017. .

Amey Vrudhula, Alan C Kwan, David Ouyang, and Susan Cheng. Machine learning and bias in medical imaging: Opportunities and challenges. *Circ. Cardiovasc. Imaging*, 17(2):e015495, February 2024.

Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging AI in real-world generalization. *Nat. Med.*, 30(10):2838–2848, October 2024.

Ghada Zamzmi, Adarsh Subbaswamy, Elena Sizikova, Edward Margerrison, Jana G Delfino, and Aldo Badano. Scorecard for synthetic medical data evaluation. *Commun Eng*, 4(1):130, July 2025.

Siqi Zhang, Qizhe Zhang, Shanghang Zhang, Xiaohong Liu, Jingkun Yue, Ming Lu, Huihuan Xu, Jiaxin Yao, Xiaobao Wei, Jiajun Cao, et al. A generalist foundation model and database for open-world medical image segmentation. *Nature Biomedical Engineering*, pages 1–16, 2025.

Chuan Zhou, Heang-Ping Chan, Chintana Paramagul, Marilyn A Roubidoux, Berkman Sahiner, Labomir M Hadjiiski, and Nicholas Petrick. Computerized nipple identification for multiple image analysis in

computer-aided diagnosis. *Medical Physics*, 31(10): 2871–2882, 2004.