A knowledge-based method for detecting network-induced shape artifacts in synthetic images

Rucha Deshpande[1] RUCHA.DESHPANDE@FDA.HHS.GOV
Miguel Lago[1] MIGUEL.LAGO@FDA.HHS.GOV
Adarsh Subbaswamy[1] ADARSH.SUBBASWAMY@FDA.HHS.GOV
Seyed Kahaki[1] SEYED.KAHAKI@FDA.HHS.GOV
Jana Delfino[1] JANA.DELFINO@FDA.HHS.GOV
Aldo Badano[1] ALDO.BADANO@FDA.HHS.GOV
Ghada Zamzmi[1] ALZAMZMIGAA@FDA.HHS.GOV
[1] *Center for Devices and Radiological Health, U. S. Food and Drug Administration, U.S.A.*

Editors: Accepted for publication at MIDL 2025

Abstract

Synthetic data provides a promising solution to address data scarcity for training machine learning models; however, adopting it without proper quality assessments may introduce artifacts, distortions, and unrealistic features that compromise model performance and clinical utility. This work introduces a novel knowledge-based method for detecting network-induced shape artifacts in synthetic images. The method can 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 using two synthetic mammography datasets. A reader study further confirmed that images identified by the method as likely containing network-induced artifacts were also flagged by human readers. This method is a step forward in the responsible use of synthetic data by ensuring that synthetic images adhere to realistic anatomical and shape constraints.

Keywords: synthetic data evaluation, network-induced artifacts, mammography

1. Introduction

Deep learning has transformed medical imaging, particularly for automated image analysis applications and clinical decision-making. However, the development of robust deep learning models is hindered by limited access to large-scale, high-quality patient datasets. In this context, synthetic data has emerged as a promising solution for augmenting patient datasets while safeguarding patient privacy. 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). Despite these advancements, assessing the quality and clinical relevance of synthetic medical images, particularly those generated by deep generative models, remains a challenge (Deshpande et al., 2025; 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 arise when models prioritize matching overall data distributions over preserving fine-grained, image-level details. This can lead to distortions, unnatural shapes, and other irregularities that compromise the reliability and

clinical utility of downstream models. Although the issue of network-induced artifacts in synthetic images has been documented in literature (Lee et al., 2023; Müller-Franzes et al., 2023; Deshpande et al., 2025; Kelkar et al., 2023; Schwarz et al., 2021), automated methods for identifying such artifacts in individual images remain scarce (Deshpande et al., 2025). Popular evaluation methods typically rely on dataset-wide metrics (Borji, 2022), 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.

Assessing individual image quality in large synthetic datasets has 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 artifacts in synthetic datasets.

In this paper, we propose a method for detecting network-induced 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-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, (iii) is model-agnostic, 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.

2. Materials & Methods

2.1. Datasets

We used two synthetic digital mammography datasets in the mediolateral oblique (MLO) view to demonstrate the proposed method. Each synthetic dataset was generated via a different generative model.

The first synthetic dataset, Sinkove, is a public dataset (Pinaya et al., 2023) with 100,000 images generated by a latent diffusion model trained on the CSAW-M patient dataset (Sorkhei et al., 2021), which consists of about 10,000 images sized 632×512. These real and synthetic datasets are hereafter referred to as CSAW-real and CSAW-syn, respectively.

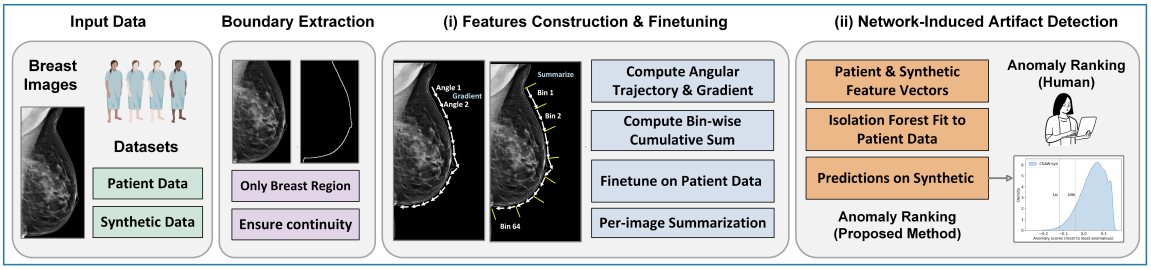


Figure 1: Overview of the proposed method for detecting network-induced shape artifacts.

The second synthetic dataset, VMLO-syn, was generated using StyleGAN2 (Karras et al., 2020) trained on the VinDr-Mammo (VMLO) dataset (Pham et al., 2022), which contains approximately 10,000 images. We trained StyleGAN2 on Nvidia A100 GPUs. For both datasets, pre-processing was performed as specified by the dataset authors (Pham et al., 2022; Sorkhei et al., 2021). For the VMLO dataset, an additional image resizing to 512×512 was performed to meet the input requirements of the generative model. Additional information about both patient datasets and example images can be found in Appendix A, whereas examples of synthetic images are provided in the Results section.

2.2. Network-induced Shape Artifact Detection

Our method (Figure 1) consists of (i) a novel feature extractor, and (ii) an artifact detector.

2.2.1. Feature Space Construction

Boundary extraction and tracking. To generate a feature representation of the anatomy of interest, the breast region is first segmented via thresholding and morphological operations (details in Appendix H). Specifically, 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 $\mathcal{P}$. Disjoint boundary sections are then connected via morphological opening to ensure robustness when the anatomy of interest exceeds the field of view. 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. To normalize for variations in size, $\mathbf{a}'$ is binned into 64 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 64-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 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.

Per-image summarization of features. 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. The bin edges of the 16-d vector are determined from the approximate range (1-99 percentiles) of cusum values in the *patient* data distribution and the extreme bins are kept open. Each per-image feature vector from the previous step is binned into a 16-d vector accordingly for a given dataset. Extreme bins are eliminated for robustness and the resulting patient and synthetic feature vectors serve as inputs for artifact detection.

2.2.2. Artifact Detection

For detecting network-induced artifacts, we employ 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). 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 proposed method demonstrates robustness to variations in breast area, as evidenced by the experiments in Appendix B.

3. Results

We present three sets of results from the proposed method–dataset-level, image-level, and reader study results–with additional results provided in Appendix C.

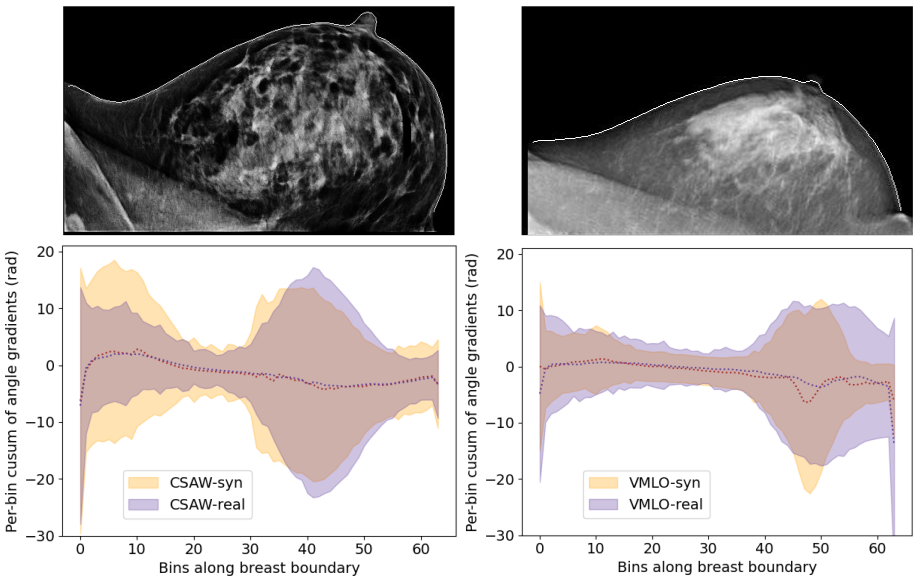


Figure 2: Distributions of bin-wise cumulative sum of angle gradients (dotted line: mean, shaded: one std) show substantial but incomplete overlap between patient and synthetic datasets. Left-to-right bins represent breast shape from top to bottom.

3.1. Dataset-Level Results

Figure 2 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 2, 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-M) and Vietnamese (VinDr-Mammo) 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. Further quantitative results are provided in Appendix C. Notably, in the VMLO-syn dataset, the bin-wise means (dotted lines in Figure 2) are clearly distinct from the corresponding patient dataset, indicating a bias in the synthetic dataset toward a specific breast shape.

The Fréchet Inception Distance (FID) scores for the two synthetic datasets are 22 (CSAW-syn) and 43 (VMLO-syn), indicating reasonable visual quality overall, even though some images contain artifacts (Figure 3 and Figure 4). Note that although FID scores may vary based on the generative model and its optimization (Lucic et al., 2018), in case of

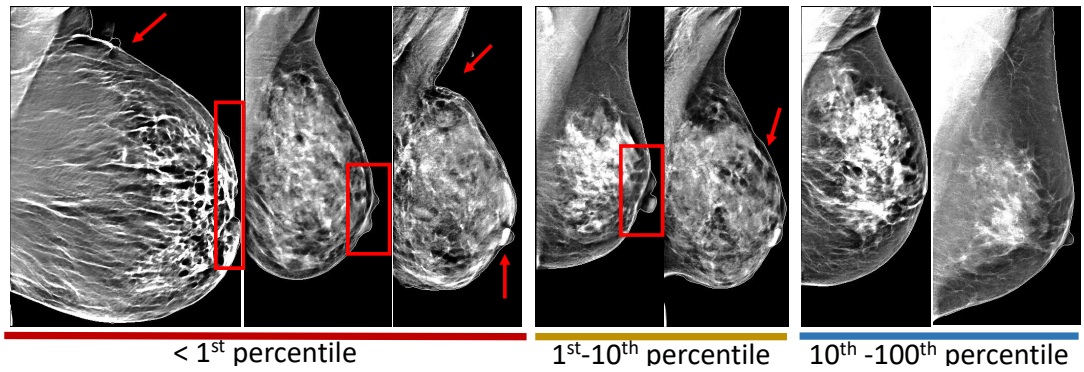


Figure 3: Most to least (L-R) artifact images from CSAW-syn as ranked by our method. Annotations in red are for display only.

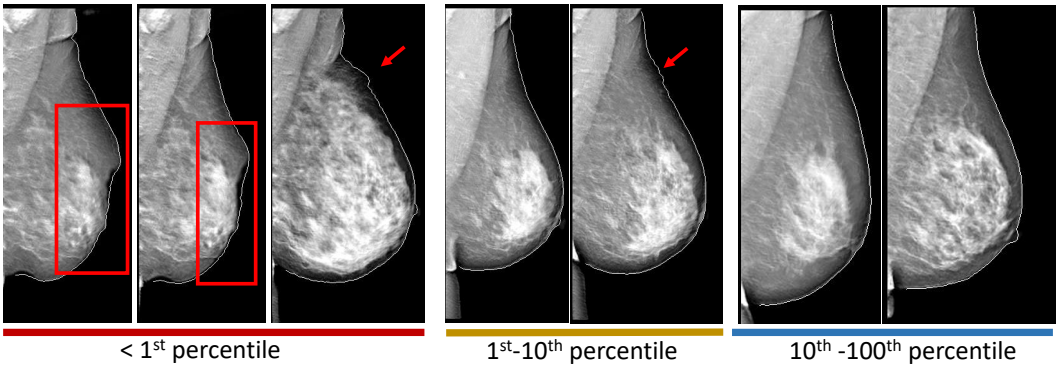


Figure 4: Most to least (L-R) artifact images from VMLO-syn as ranked by our method. Annotations in red are for display only.

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). The FID scores of our synthetic datasets also lie within this range. 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.

3.2. Image-Level Results

The distributions of anomaly scores obtained using the proposed method are shown in Figure 8 in Appendix D. 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) cor-

responds to images with minimal or no visible artifacts. The categorization of anomaly scores into three partitions is a reasonable starting point when the prevalence of anomalies in a dataset is unknown. These partitions can be further customized and fine-tuned given prior knowledge of the task or the approximate rate of images with artifacts.

Figure 3 and Figure 4 present examples of 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 exhibits 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 4), alongside minor local artifacts (1st-10th percentile) like non-smooth or angular boundaries. Artifacts typically observed in P1 were not observed in patient datasets, and thus, originated from the generative process itself. Some artifacts in P2 resembled the most anomalous images in the patient dataset, as discovered when the proposed method was applied to patient data; (examples of natural shape artifacts shown in Appendix J). 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.

3.3. Results from the Reader Study

A 2-Alternative Forced Choice (2-AFC) reader study was conducted separately for each dataset using the SimplePhy tool (Lago, 2021). Three imaging scientists participated as readers. Additional details about the reader study are provided in Appendix E.

Thirty images from each 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 (435 trials) were shown to the readers as illustrated in Figure 9 in Appendix E. Readers were instructed to selected the image with stronger shape artifacts from each side-by-side comparison.

Results are summarized in Table 1 as the percentage of trials (mean$\pm$std) in which an image was identified as having the most pronounced network-induced shape artifacts, reported for each partition. For both datasets, the first partition ($\leq$1%) was consistently identified as containing the most network-induced shape artifacts by all readers, with mean values approximately 1.5-2 times higher than the last partition (10-100%). This indicates that shape artifacts were effectively concentrated in the first partition, greatly improving search efficiency compared to random visual searches (details in Appendix F). Further, the similar mean values for the first partition ($\approx$66%) suggest that all readers consistently found artifacts in this partition to be most distinctive compared to other partitions.

Note that the mean values are expected to be bounded between 17% and 83% assuming that images within a partition are highly similar with each other in terms of the presence/level of visible artifacts (details in Appendix G). Additionally, mean values decreased

Table 1: 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**			**VMLO**		
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%

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 dataset, reflecting a lower fraction of shape artifacts in VMLO compared to CSAW. This is supported by the score distributions (Figure 8 in Appendix D), where the second partition in CSAW contains negative scores, while that in VMLO straddles zero.

Kendall-Tau correlations between reader rankings and the algorithm rankings over all images in the reader study were 0.45 for CSAW (reader values: 0.43, 0.51, 0.40) and 0.43 for VMLO (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 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 found to be 0.97 (CSAW) and 0.91 (VMLO).

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 negatively affect downstream task performance. 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. It provides rank-

ings instead of binary decisions on artifact presence, which can improve the efficiency of visual search for artifacts. 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. We hope that our method will assist annotators and domain experts by providing the first step in obtaining labels for a dataset, which can then be used to develop semi-supervised or supervised anomaly detection methods, or to clean synthetic datasets before using them for training or testing AI models.

While we provide an example with breast imaging, the proposed method can be generalized to other anatomies and imaging modalities where the region of interest can be segmented, as further elaborated in Appendix I. It is particularly valuable in scenarios where anatomy cannot be described adequately by area-based features or conventional compactness and convexity features. In addition to mammograms, other examples include 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.

A limitation of this work is that 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. 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 (refer Appendix J) 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, the proposed method provides a practical tool for isolating individual images with network-induced shape artifacts. This is valuable for evaluating large synthetic datasets where visual inspection of individual images is impractical. As the method is agnostic to the generative process, it is applicable to emerging generative model architectures and scenarios where the nature of artifacts is unknown.

Acknowledgments

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.

Rucha Deshpande acknowledges 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. Rucha Deshpande acknowledges the FDA's CDRH HPC for the computational resources used in this publication.

We thank Frank Samuelson and Nicholas Petrick for providing valuable feedback on our work during the development process.

Conflict of Interest

We have no conflicts of interest to declare.

Code Availability

The code will be made available at https://github.com/DIDSR/ShapeCheck

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.

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, 2025.

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.

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.

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.

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.

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.

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.

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 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*, 2023.

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.

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.

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.

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.

Appendix A. Details and image example of patient datasets

Table 2 summarizes patient datasets, with additional details provided below.

A.1. CSAW-M Patient Dataset

CSAW-M (Sorkhei et al., 2021) is a publicly available mammography dataset designed for non-commercial use, hosted by the SciLifeLab Data Repository, a Swedish infrastructure for sharing life science data. It provides screening mammograms accompanied by metadata, including expert masking potential assessments, clinical endpoints, density measures, and image acquisition parameters. Unlike other mammography datasets, CSAW-M focuses on masking potential rather than tumor detection, with images selected to represent diverse breast shapes and densities. The dataset comprises a training set (9,523 examples), a public test set (497 examples), and a private test set (475 examples) derived from the CSAW cohort, a collection of millions of mammograms from screening participants aged 40–74 collected between 2008 and 2015.

Images in CSAW-M were curated from participants at Karolinska University Hospital, using the most recent mediolateral oblique (MLO) view for optimal breast visualization. To avoid contamination from tumor presence, contralateral breast images were used for cancer cases, while random breast sides were selected for non-cancer cases. Images were preprocessed to ensure uniformity, including resizing, intensity scaling, zero-padding, and removal of text annotations. This resulted in 632×512, 8-bit PNG images suitable for analysis. Examples of CSAW-M images are shown in Figure 5.

A.2. VinDr-Mammo Patient Dataset

The VinDr-Mammo dataset (Pham et al., 2022) is a publicly available, comprehensive collection of mammography images created to advance research in computer-aided detection (CADe) and diagnosis (CADx) systems for breast cancer screening. The dataset consists of 20,000 mammography images in DICOM format, sourced from 5,000 exams conducted between 2018 and 2020 at Hanoi Medical University Hospital (HMUH) and Hospital 108 (H108) in Vietnam. It includes both screening and diagnostic exams. Images were captured using equipment from Siemens, IMS, and Planmed. To protect patient privacy, all identifiable information was removed from DICOM metadata and image annotations, with additional masking applied to textual information in the image corners. The pseudonymization process underwent manual validation by human reviewers. The dataset offers both breast-level assessments and lesion-level annotations. The dataset was divided into training (80%) and testing (20%) subsets using an iterative stratification algorithm to ensure balanced representation of key attributes, such as BI-RADS categories, breast density levels, and finding types. Examples of CSAW-M images are shown in Figure 5.

Table 2: Summary of patient digital mammography datasets

Dataset	Cases	Images	Density	View	Origin	Year	Acquisition
CSAW-M	10,020	10,020	Yes	MLO	Sweden	2021	FFDM
VinDr-Mammo	5,000	20,000	Yes	MLO, CC	Vietnam	2022	FFDM

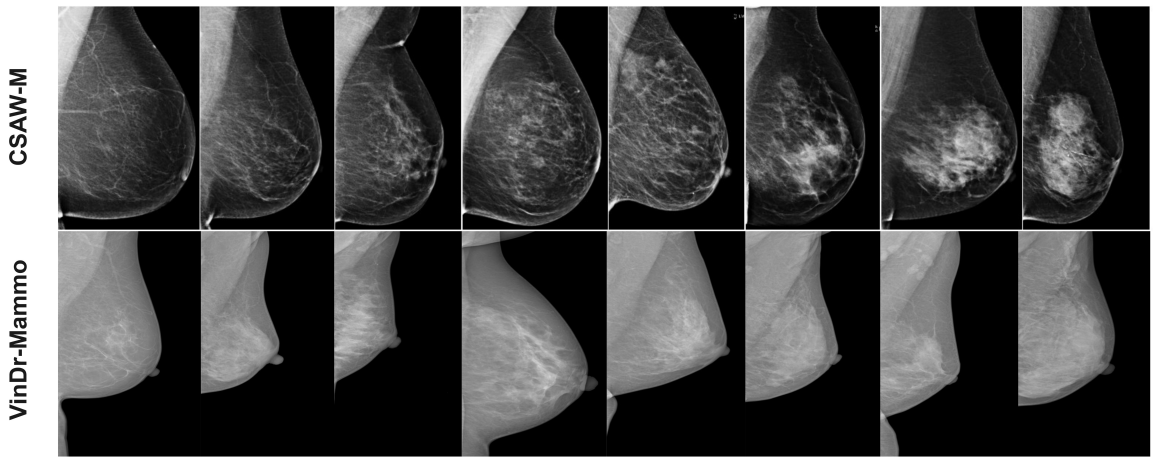
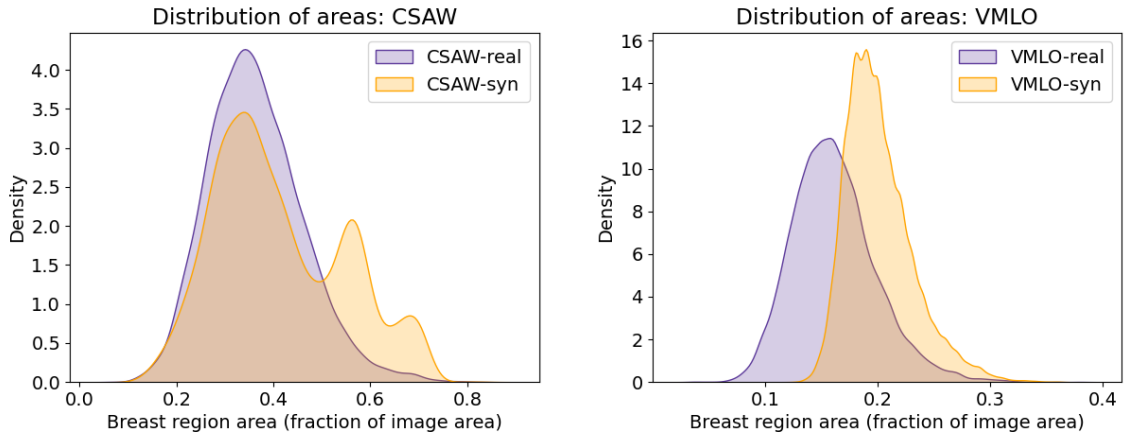


Figure 5: Examples of images from patient datasets: CSAW-M and VinDr-Mammo.

Appendix B. Robustness to the effects of breast area

The patient and synthetic data distributions for breast area are shown below (Figure 6). 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. Although the length of the breast trajectory is strongly correlated with area (Pearson's R = 0.9), binning and per-image summarization ensures robustness to the effects of area.

To assess the robustness of our proposed method to 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.



CSAW-reals VMLO-reals

Figure 6: 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.

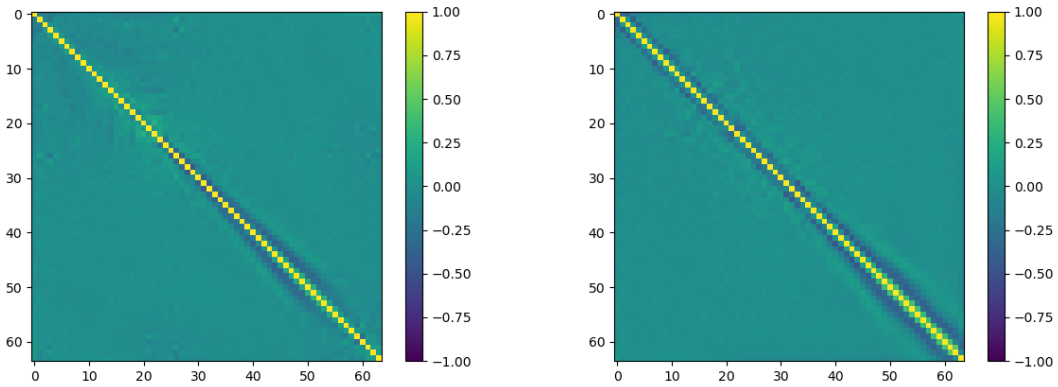
Appendix C. Additional results for the dataset-level similarity between patient and synthetic datasets

For the proposed 64-dimensional feature space, the cross-correlation of the feature vectors of the patient datasets are shown in Figure 7. Note that neighboring components demonstrate correlations as they constrain each other according to the possibilities of breast shape. 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 = 8.27e-19) for CSAW, and 0.13 (p value = 2.52e-66) for VMLO. This indicates that patient and synthetic distributions have distinct distributions and anomalous images are present in this feature space.

Appendix D. Distribution of anomaly scores

The distributions of anomaly scores obtained using the proposed method are shown in Figure 8 for both synthetic datasets. 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 artifacts. Both



CSAW-reals VMLO-reals

Figure 7: 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.

distributions are left-tailed, suggesting that only a small fraction of the synthetic dataset contains highly anomalous images. Thresholds corresponding to the 1^{st} and 10^{th} percentiles of the most anomalous images are marked in Figure 8. The corresponding images for the three quantile partitions (most to least anomalous: $\leq$1%, 1-10%, 10-100%) are presented in Figure 3 (CSAW) and Figure 4 (VMLO).

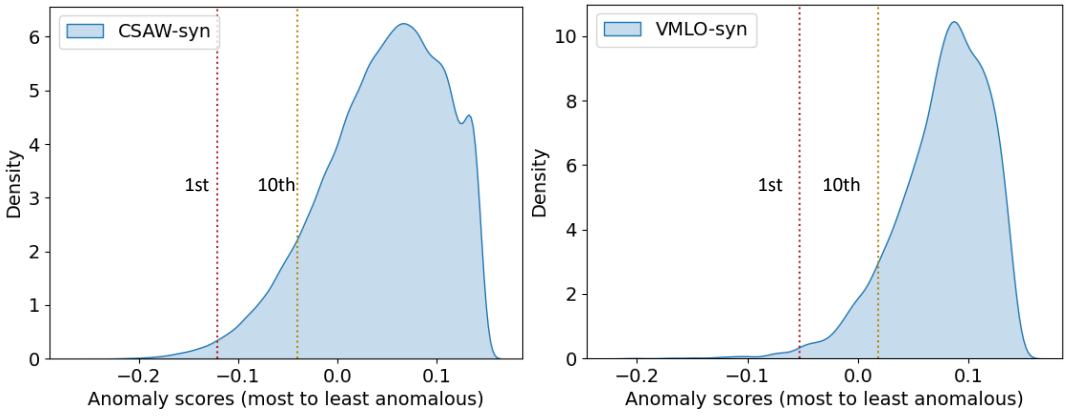
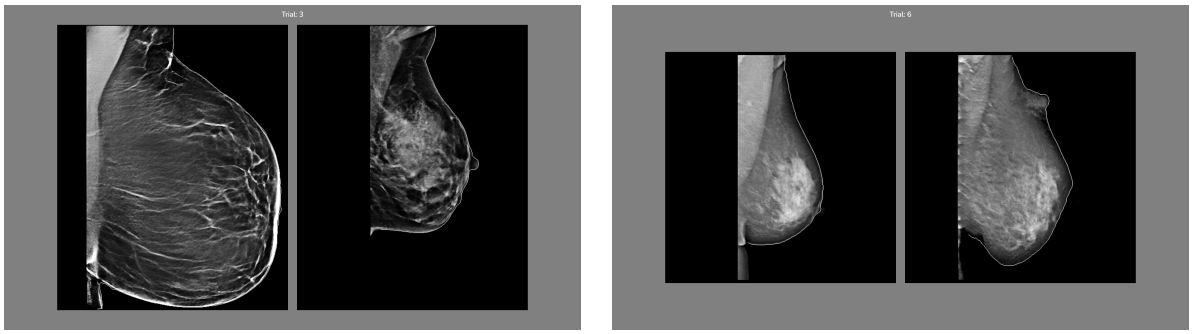


Figure 8: Distribution of anomaly scores for the two synthetic datasets. 1^{st} and 10^{th} percentile are marked.

Appendix E. Details about the reader study setup

The reader study was designed as a two-alternative forced-choice (2-AFC) experiment, where 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 9. The viewing settings were carefully calibrated for each reader to ensure consistency, and each experiment was completed in a single reading session.



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

Figure 9: 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. The left panel shows an example from the CSAW-syn dataset, while the right panel displays an example from the VMLO-syn dataset.

Appendix F. Efficiency of search computation

The proposed method can improve the efficiency of visual search. 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 a 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 accuracy (%)/artifact prevalence (%), and may differ as the two factors vary. The efficiency of artifact search is especially important when creating artifact labels for a dataset to reduce the burden of expert readers. Furthermore, it also enables efficient data cleaning and improves the quality of input data for downstream tasks.

Appendix G. Computation of bounds for reader study results

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:

$$\text{Image chosen } (\%) = P_{\text{random}} \times \frac{n_{\text{image}}}{N_{\text{total}}} \tag{1}$$

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:

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

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.

Appendix H. Details for the boundary extraction process

The boundary extraction process involves three steps:

1. Image pre-processing.
   (a) Gaussian filtering ($\sigma = 0.5$)
   (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 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. Skeletonize the boundary.
2. Remove straight edges corresponding to the imaging window.
   Note: This step may not be required for imaging modalities other than mammography, if the anatomy of interest is not bound by the straight edges of an imaging window.

(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. Ensure connectivity of the extracted border.

(a) Check: 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.

Appendix I. Extension of the proposed method to other biomedical applications

The core methodology can be extended easily to other anatomies, such as lungs, abdomen, brain, or any other anatomical region of interest, as long as the segmentation mask is available. This broad applicability is due to the reliance of the method on the characteristic shape of the anatomy.

Medical examples: In lung imaging via chest radiographs, the method can 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 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 be applied to detect synthetic 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.

Toy problem: Consider a toy problem where we assume that the task is to generate 3-pointed stars by training on a dataset of 3-pointed stars. Within an image, the sharp angular change associated with the point of the star is represented via its angular gradients. This sharp change occurs exactly 3 times, as captured in the distribution of these angular gradients. If a generative model trained on this dataset of 3-point stars, occasionally generates 5-point stars, i.e. a shape artifact, the angular gradient corresponding to the point of a star will have a clearly different distribution within an image. This difference in distribution can be flagged by an isolation forest and thus, the image is identified as a shape artifact.

Summary: In all these applications, the method works by constructing a knowledge-based feature space from the patient data, specific to the anatomical region being analyzed. By computing the distribution of angle gradients along the anatomical boundaries, the method can detect unnatural shape or deviation from expected anatomical structures, which may be indicative of synthetic artifacts.

Appendix J. Examples of natural artifacts in patient data

Natural artifacts from post-surgical breasts, thin breasts and skin folds are sometimes present in the patient datasets. Some examples of natural artifacts are shown in the figure below.

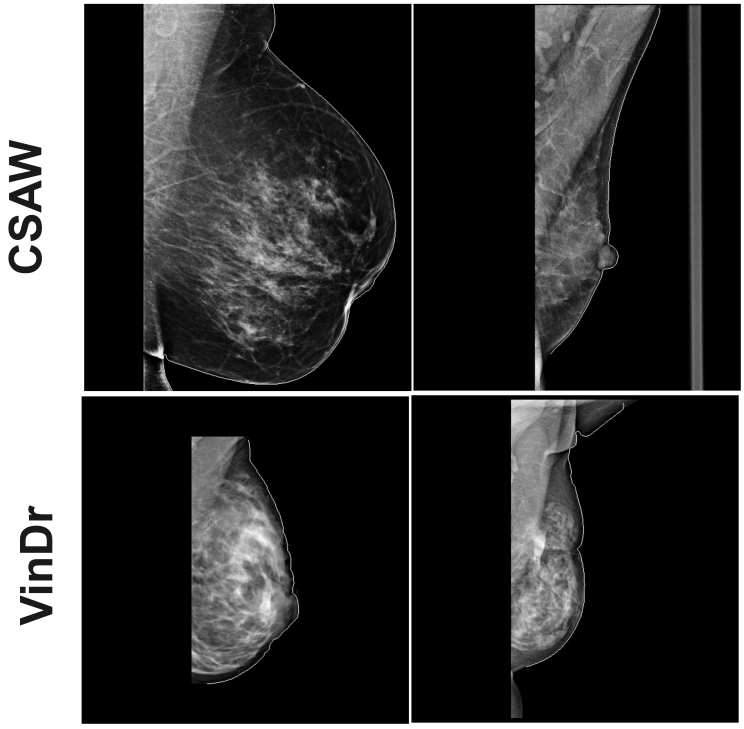


Figure 10: Examples of natural variations in breast shape from patient datasets: first row shows CSAW patient data, and second row shows VinDr patient data.