Uncertainty Quantification for Synthetic Medical Images

Koukoutegos K.[a,b], Sizikova E.[c], Bosmans H.[a,b], and Badano A.[c]

[a]UZ Leuven, Department of Radiology, Herestraat 49, 3000 Leuven, Belgium
[b]KU Leuven, Department of Imaging and Pathology - Medical Physics & Quality Assessment, Herestraat 49, 3000 Leuven, Belgium
[c]U.S. Food and Drug Administration, Center for Devices and Radiological Health, Office of Science and Engineering Laboratories, Silver Spring, 20993 Maryland, USA

ABSTRACT

Aim: The goal of this project is to review key uncertainty estimation classes in medical imaging, and summarize the directions of uncertainty measurement occurring as a result of using synthetic data during training. Finally, we present a case study summarizing uncertainty estimation for the example task of breast lesion segmentation.
Materials and Methods: We trained UNet-based segmentation models using varying proportions of patient and synthetic mammograms from the INbreast and M-SYNTH datasets. Three types of uncertainty—data, reader, and case—were quantified using Dice standard deviation across test splits, random seeds, and Monte Carlo Dropout.
Results and Conclusions: The uncertainty patterns varied depending on the composition of synthetic and patient data. Alignment of synthetic data with test distributions improved model confidence, while inconsistency across seeds highlighted the need for tuning the synthetic data amount. Some cases remained uncertain regardless of data volume, suggesting intrinsic segmentation difficulty. These results highlight the nuanced impact of synthetic data on model reliability.

Keywords: Breast Imaging, Neural Networks, Synthetic Data, Uncertainty Quantification

1. INTRODUCTION

The growing availability of synthetic medical imaging data presents new opportunities for training and evaluating AI models, particularly in settings where patient data is scarce or difficult to obtain. Synthetic data is increasingly used to supplement or, in some cases, replace patient data during training in medical imaging tasks. However, as its use becomes more widespread, it is critical to assess not only how synthetic data affects downstream performance, but also how it influences model uncertainty—especially in high-stakes applications such as clinical diagnosis.

To quantify this impact, researchers often compare performance metrics (e.g., mean accuracy or Dice score) across models trained with varying proportions of synthetic data. For example, prior studies[3,6,9] have examined three experimental paradigms: (1) limited patient

E-mail: elena.sizikova@fda.hhs.gov

data, where portions of patient data is used for training, (2) data replacement, where a portion of patient data is replaced by synthetic data and compared against a model trained on the full patient dataset (the oracle), and (3) data addition, where synthetic data is added incrementally to the patient dataset, and performance is evaluated on a fixed held-out test set. In each of the scenarios above, the size of training set and the proportions of patient and synthetic data are controlled to ensure fair comparisons. In addition to evaluating the effect of average performance in scenarios described above, models trained on a combination of synthetic and patient data may exhibit uncertainty due to variation in input images, as well as model/reader performance. In other words, while uncertainty estimation is a mature area of machine learning research, there is a lack of paradigm for evaluating uncertainty specifically due to the use of synthetic data during AI training in medical imaging. On the other hand, the use of synthetic data is increasing for this application, and understanding the variability and reliability of model predictions—i.e., the uncertainty—would offer a more nuanced and comprehensive evaluation of synthetic data utility.[8,11]

In this study, we approach this problem through the lens of probabilistic uncertainty quantification,[2] which uses probability theory to model and estimate uncertainty in AI predictions. In this framework, two broad types of uncertainty are typically distinguished. Aleatoric uncertainty captures inherent noise or randomness in the data generating process, such as variability in image quality or ambiguous labels, is difficult to measure and considered irreducible. In contrast, epistemic uncertainty arises from a lack of knowledge or limited data and reflects the model's uncertainty about its own parameters or structure and the uncertainty induced by the learning process; it is reducible with more data or better modeling.[1,4] To better understand how synthetic data influences different sources of epistemic uncertainty in medical image segmentation, we apply three commonly observed types of performance variability:

- Data uncertainty, emerging when models are trained on data from different source domains (e.g. patient and synthetic). This introduces additional variation in learned representations and reflects epistemic uncertainty arising from data distribution shifts.
- Reader uncertainty (also known as model or observer uncertainty), which quantifies the variability across different models trained on the same data but with different random initializations. This is another form of epistemic uncertainty, capturing sensitivity to the learning process itself.
- Case uncertainty (or subject uncertainty), capturing variability in predictions due to differences across individual test cases, which reflects epistemic uncertainty at the input level.

By relating these performance variability types to the probabilistic concept of epistemic uncertainty, we aim to provide a deeper understanding of how synthetic data affects model reliability and generalization, through empirical experiments on breast lesion segmentation.

2. MATERIALS AND METHODS

2.1 Proposed framework

Given dataset pairs (**X,Y**), where **X** denotes mammograms and **Y** lesion segmentation maps, we train UNet[7] segmentation models $\mathcal{M} : f(\mathbf{X}) = \mathbf{Y}$. In our study, we use both patient $\mathcal{D}_{patient} = (\mathbf{X}_{patient}, \mathbf{Y}_{patient})$ and synthetic $\mathcal{D}_{synthetic} = (\mathbf{X}_{synthetic}, \mathbf{Y}_{synthetic})$ mammographic images. To estimate the segmentation uncertainty, varying combinations of (**X,Y**) pairs are created, using different percentages of *patient* and *synthetic* data respectively. These dataset settings can be expressed as:

$$\mathcal{D}_{p,q} = (p\mathcal{D}_{patient}) \cup (q\mathcal{D}_{synthetic}) \tag{1}$$

$$\mathcal{D}_{p,q} = (p\mathbf{X}_{patient}, p\mathbf{Y}_{patient}) \cup (q\mathbf{X}_{synthetic}, q\mathbf{Y}_{synthetic}) \tag{2}$$

$$\mathcal{D}_{p,q} = ((p\mathbf{X}_{patient} \cup q\mathbf{X}_{synthetic}), (p\mathbf{Y}_{patient} \cup q\mathbf{Y}_{synthetic})) \tag{3}$$

where p and q denote the proportions of patient and synthetic data. For each dataset setting $\mathcal{D}_{pi,qi} = (pi\mathcal{D}_{\text{patient}}, qi\mathcal{D}_{\text{synthetic}})$, a model $\mathcal{M}_{pi,qi}$ is trained. This allows us to benchmark the uncertainty of different models $\mathcal{M}_{pi,qi}$ against the baseline $\mathcal{M}_{100,0}$, which is trained solely on patient data. By comparing different uncertainty values, we can assess the robustness and reliability of the segmentation performance across varying dataset compositions. The proposed framework is depicted in Fig. 1.

2.2 Data uncertainty

A simplistic approach to derive the uncertainty across an entire dataset is by calculating the standard deviation of the Dice score σ_{Dice} across the test split of each $\mathcal{D}_{pi,qi}$. Given a set of $\mathcal{N}$ Dice coefficients, $\sigma_{\text{Dice}_{pi,qi}}$ is the uncertainty of each model $\mathcal{M}_{pi,qi}$. Comparison of the uncertanties for each *pi* and *qi* values can identify the best *patient* and *synthetic* percentages that make the model most confident.

2.3 Reader uncertainty

To quantify the uncertainty inherent to the model's parameters, we trained different realizations of UNet models using different starting seed points for the parameter space. Let $Dice_{q1}$, $Dice_{q2}$, ... , $Dice_{qi}$ be the average Dice scores for the entire test set, when each of $q1, q2, ..., qi \in Q$ seed points is used to initialize training. The uncertainties of each model trained using Q are $\sigma_{\text{Dice}_{q1}}$, $\sigma_{\text{Dice}_{q2}}$, ..., $\sigma_{\text{Dice}_{qi}}$.

2.4 Case uncertainty

To estimate the uncertainty of an individual case, we incorporate Monte Carlo Dropout[10] during model evaluation. The uncertainty for each case is quantified by calculating the standard deviation of the Dice across different evaluations. Let $\mathbf{Y}_{k1}, \mathbf{Y}_{k2}, ..., \mathbf{Y}_{kj}$ represent the predicted outputs for the k^{th} test case using MC Dropout $\mathcal{C}$ times, i.e. $j = 1, 2, ..., \mathcal{C}$. This results in a distribution of Dice scores for the k^{th} case, with $Dice_{kj} \in \{Dice_{k1}, Dice_{k2}, ..., Dice_{kC}\}$. The uncertainty of the prediction can be quantified by calculating the standard deviation of this distribution, σ_{Dice_k}, which represents how much uncertain the model is when predicting case k.

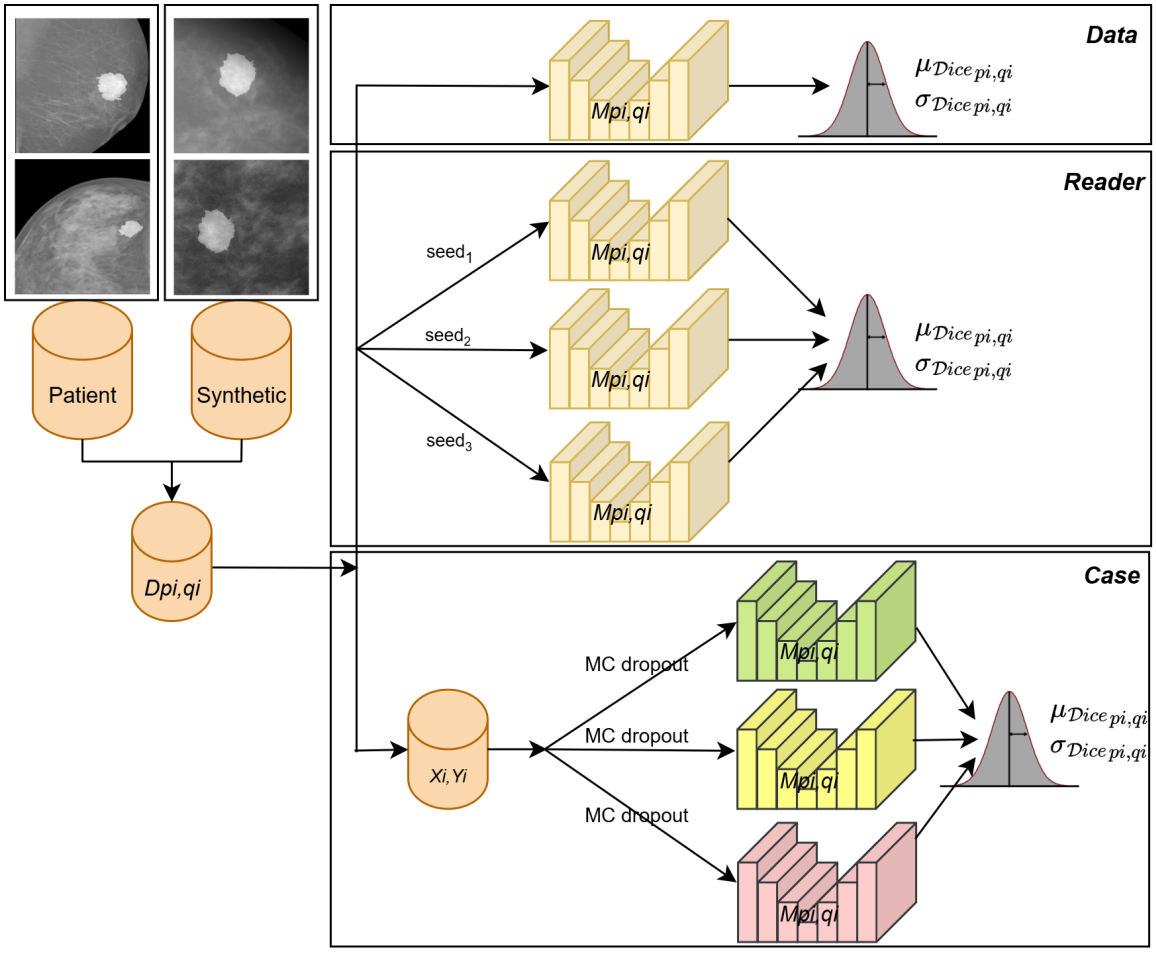


Figure 1: Sample uncertainty quantification framework. For each combination of patient and synthetic data $\mathcal{D}_{pi,qi}$, a UNet $\mathcal{M}_{pi,qi}$ is trained. Data uncertainty is captured from a single UNet, on the entire test set. Reader uncertainty is calculated using different random initial points during the UNet training. Case uncertainty is calculated using Monte Carlo Dropout, predicting the same case using different realizations of the same model. Segmentation uncertainty is quantified as the standard deviation of Dice similarity coefficient σ_{Dice}.

2.5 Datasets

In this study, two distinct datasets were utilized, in order to assess and quantify the uncertainty of our proposed settings. INbreast dataset[5] consists of patient mammographic cases, while M-SYNTH[9] comprises synthetic data.

2.5.1 INbreast

The INbreast dataset is a large publicly available collection of full field X-ray digital mammography (FFDM) images. It includes 410 mammograms, representing both mediolateral oblique (MLO) and craniocaudal (CC) views from 115 patients. Among these cases, lesions are present in 105 of them, with a total of 113 identified lesions (some cases contain more than one). Each mammogram in the dataset is accompanied by ground truth annotations

and BIRADS scores, providing detailed information on the presence, location, and characteristics of the lesions. We used a split of $[0.65, 0.15, 0.25]$ for train, validation, and test sets for all experiments.

2.5.2 M-SYNTH

M-SYNTH is a collection of synthetic digital mammography (DM) images generated to simulate various clinical scenarios. This dataset includes images representing four different breast fibroglandular density distributions of dense, heterogeneously dense, scattered, and fatty, imaged using Monte Carlo X-ray simulations. Lesions are characterized by three different mass radii (5 mm, 7 mm, and 9 mm) and three different mass densities (1.0, 1.06, and 1.1, representing the ratio of the radiodensity of the mass to that of fibroglandular tissue). Additionally, the images were generated with five relative dose levels (20%, 40%, 60%, 80%, and 100%) of the clinically recommended dose for each density. The M-SYNTH dataset is publicly available and includes ground truth annotations for all lesions. In our experiments we made use of fatty breasts, with a lesion density of 1.0, lesion size of 5 mm, and 100% of the clinically recommended dose.

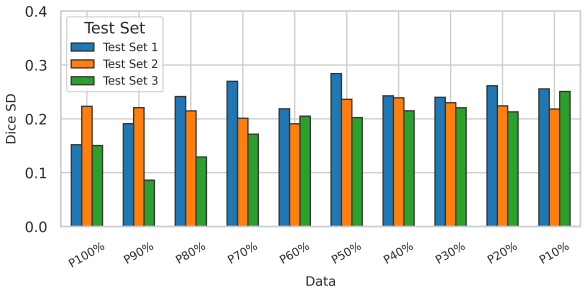
3. RESULTS

3.1 Data Uncertainty

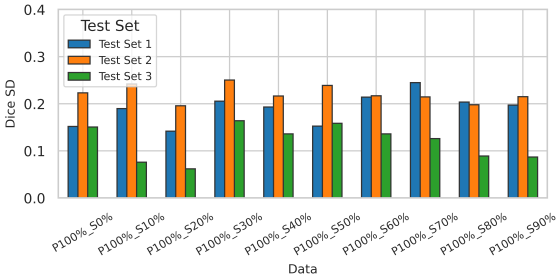
Segmentation uncertainty analysis was conducted across three different test set splits. The model behavior varied depending on the combination of patient and synthetic data availability, as depicted in Fig. 2. When patient data was decreasing, the uncertainty increased consistently across all test splits, ranging between 0.08 and 0.28. In contrast, when the amount of patient data was kept constant with increasing synthetic data, the uncertainty varied across test splits: it increased in split 1, remained stable in split 2, and decreased in split 3, with observed ranges of 0.14–0.24 for split 1, 0.19–0.25 for split 2, and 0.06–0.15 for split 3. When the amount of patient data decreased while increasing synthetic data, the uncertainty depended on the specific test split, increasing in splits 1 and 3 while remaining stable in split 2. The corresponding uncertainty ranges were 0.07–0.24 (split 1), 0.17–0.24 (split 2), and 0.05–0.23 (split 3).

3.2 Reader Uncertainty

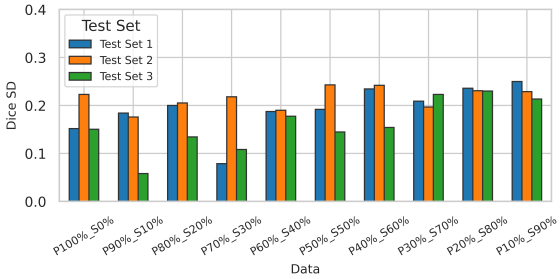
Segmentation uncertainty was further assessed with respect to different random initializations (seeds 0, 1, and 2) during the model training (Fig. 3). When patient data was decreasing, the uncertainty increased consistently across all seeds, with ranges of 0.15–0.28 for seed 0, 0.13–0.26 for seed 1, and 0.19–0.28 for seed 2. When the amount of patient data was kept constant and synthetic data was increased, the uncertainty fluctuated across all seeds — increasing, then decreasing, and increasing again — with ranges of 0.14–0.24 for seed 0, 0.17–0.24 for seed 1, and 0.18–0.26 for seed 2. In the scenario where patient data decreased and synthetic data increased, the uncertainty either increased or remained relatively stable depending on the seed, ranging from 0.07–0.24 (seed 0), 0.15–0.28 (seed 1), and 0.18–0.26 (seed 2).



(a)

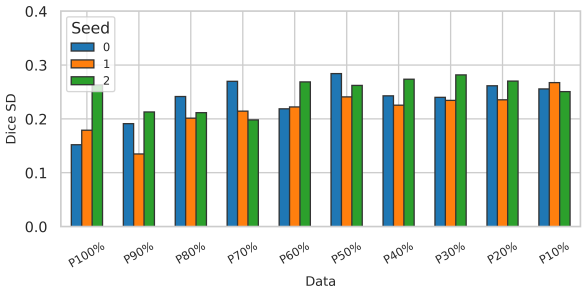


(b)

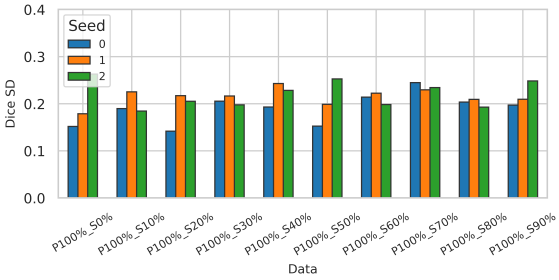


(c)

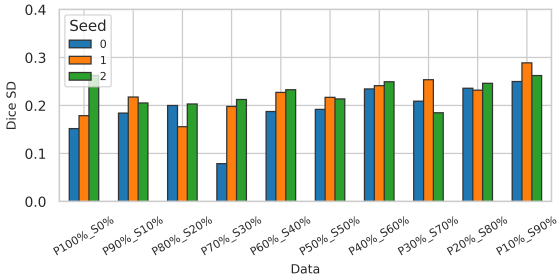
Figure 2: Data uncertainty. (a) Patient data, (b) and (c) varying combinations of patient and synthetic.



(a)



(b)



(c)

Figure 3: Reader uncertainty. (a) Patient data, (b) and (c) varying combinations of patient and synthetic.

3.3 Case Uncertainty

Segmentation uncertainty was also evaluated across individual test cases, using MC Dropout. When patient data was decreased, the uncertainty increased slightly, with average values ranging between 0.19 and 0.25. When the amount of patient data remained constant and synthetic data increased, the uncertainty first decreased and then increased again, with average values ranging between 0.08 and 0.22. In the case of decreasing patient data combined with increasing synthetic data, the uncertainty showed a slight decrease followed by a minor increase, ranging between 0.06 and 0.25 on average.

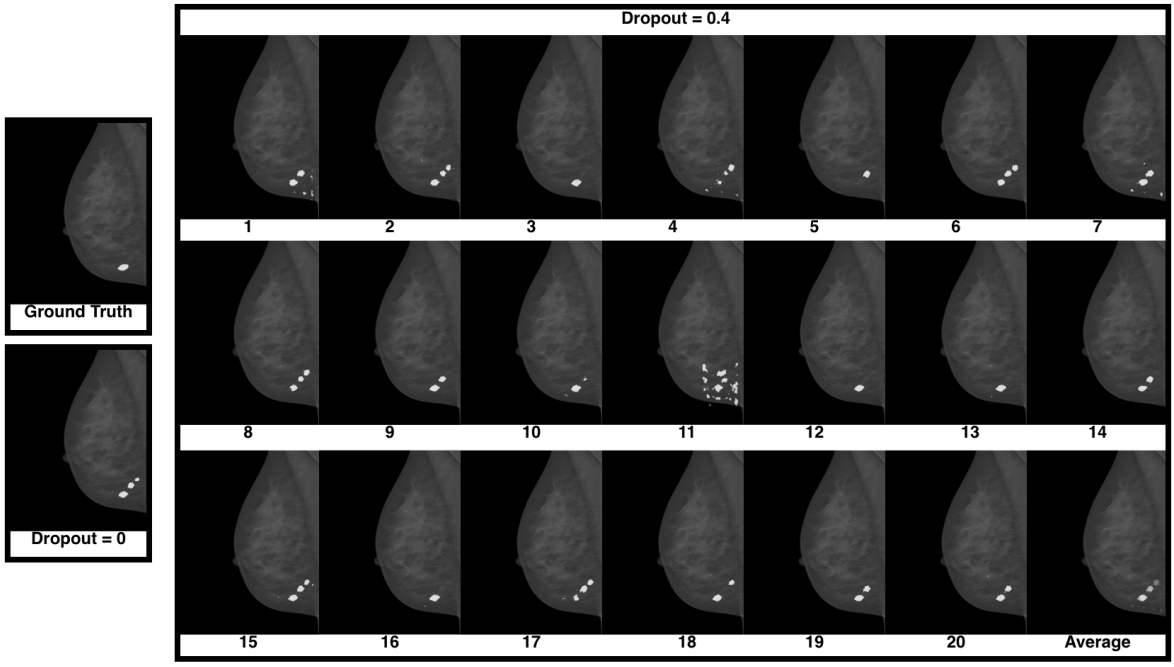


Figure 4: Test image inference using Monte Carlo dropout 20 times. Interestingly, the prediction when dropout = 0 is less accurate compared to some individual MC dropout samples (with dropout = 0.4).

4. LIMITATIONS

Our study considered only epistemic uncertainty; capturing aleatoric noise requires heteroscedastic output heads, which requires tailored modelling and is not directly applicable to most deep learning segmentation models. The limited INBreast patient cohort reduces our power to reliably estimate Dice variance and constrains further generalization. Additionally, the small number of M-SYNTH lesions may have biased our learning process as the synthetic proportion increases, potentially affecting model performance.

5. CONCLUSION

This study provides an initial investigation into how synthetic medical imaging data influences uncertainty in AI-based breast lesion segmentation. By distinguishing between data,

reader, and case uncertainty, we observed that the impact of synthetic data is highly context-dependent. First, the variability in uncertainty across different test splits suggests that the usefulness of synthetic cases may depend on how well they align with the target data distribution. Second, the oscillatory or inconsistent uncertainty trends across different random seeds highlight that there is no universally optimal amount of synthetic data; careful tuning is required to strike the right balance for model confidence. Last, persistent uncertainty in certain cases, even with abundant patient data, suggests inherent difficulty in segmenting specific lesions. While synthetic data may aid in these situations, its effectiveness again depends on identifying an appropriate inclusion threshold. These findings underscore the need for more nuanced strategies when incorporating synthetic data into medical AI training pipelines.

6. 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.

REFERENCES

1. S. C. Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering System Safety*, 54(2):217–223, 1996. Treatment of Aleatory and Epistemic Uncertainty.
2. L. Huang, S. Ruan, Y. Xing, and M. Feng. A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. *Medical Image Analysis*, 97:103223, 2024.
3. A. Kim, N. Saharkhiz, E. Sizikova, M. Lago, B. Sahiner, J. Delfino, and A. 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.
4. A. D. Kiureghian and O. Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009. Risk Acceptance and Risk Communication.
5. I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso. Inbreast: Toward a full-field digital mammographic database. *Academic Radiology*, 19(2):236–248, 2012.
6. P. Osorio, G. Jimenez-Perez, J. Montalt-Tordera, J. Hooge, G. Duran-Ballester, S. Singh, M. Radbruch, U. Bach, S. Schroeder, K. Siudak, et al. Latent diffusion models with image-derived annotations for enhanced ai-assisted cancer diagnosis in histopathology. *Diagnostics*, 14(13):1442, 2024.
7. Philipp, B. T. R. Olaf, and Fischer. U-net: Convolutional networks for biomedical image segmentation. pages 234–241. Springer International Publishing, 2015.
8. E. Sizikova, A. Badal, J. G. Delfino, M. Lago, B. Nelson, N. Saharkhiz, B. Sahiner, G. Zamzmi, and A. Badano. Synthetic data in radiological imaging: Current state and future outlook. *BJR—Artificial Intelligence*, page ubae007, 2024.
9. E. Sizikova, N. Saharkhiz, D. Sharma, M. Lago, B. Sahiner, J. 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. *Advances in Neural Information Processing Systems*, 36:37401–37412, 2023.

10. N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting, 2014.
11. G. Zamzmi, A. Subbaswamy, E. Sizikova, E. Margerrison, J. G. Delfino, and A. Badano. Scorecard for synthetic medical data evaluation. *Communications Engineering*, 4(1):130, 2025.