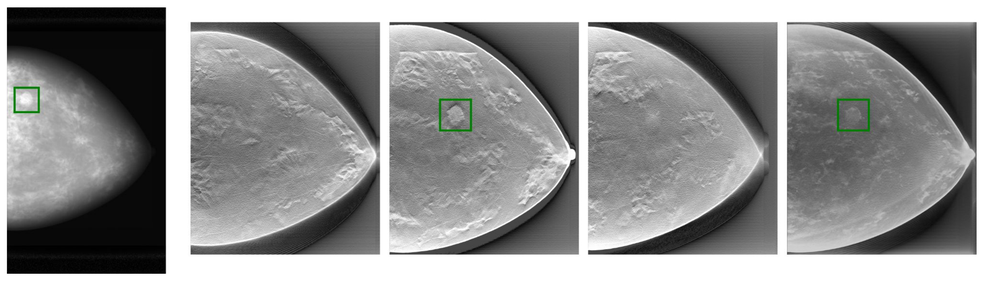
T-SYNTH: A Knowledge-Based Dataset of Synthetic Breast Images

Christopher Wiedeman†, Anastasiia Sarmakeeva†, Elena Sizikova*, Daniil Filienko, Miguel Lago, Jana G. Delfino, Aldo Badano

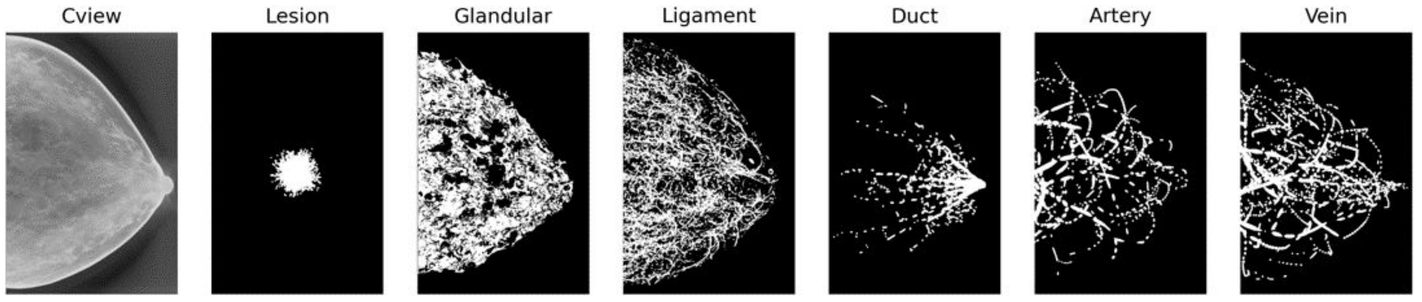
Office of Science and Engineering Labs, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD, USA

Motivation

Robust breast image dataset are required for developing computer-aided diagnosis for breast cancer. Data synthesized with deep generative models are limited by the demographic characteristics of the original training data and the ground-truth information captured in the imaging system [1,2]. Here, we release T-SYNTH, a large-scale open-source synthetic dataset of paired digital mammography (DM) and digital breast tomosynthesis (DBT) images generated with a knowledge-based (KB) model. T-SYNTH contains 9,000 images and is designed to complement the M-SYNTH DM dataset [6].



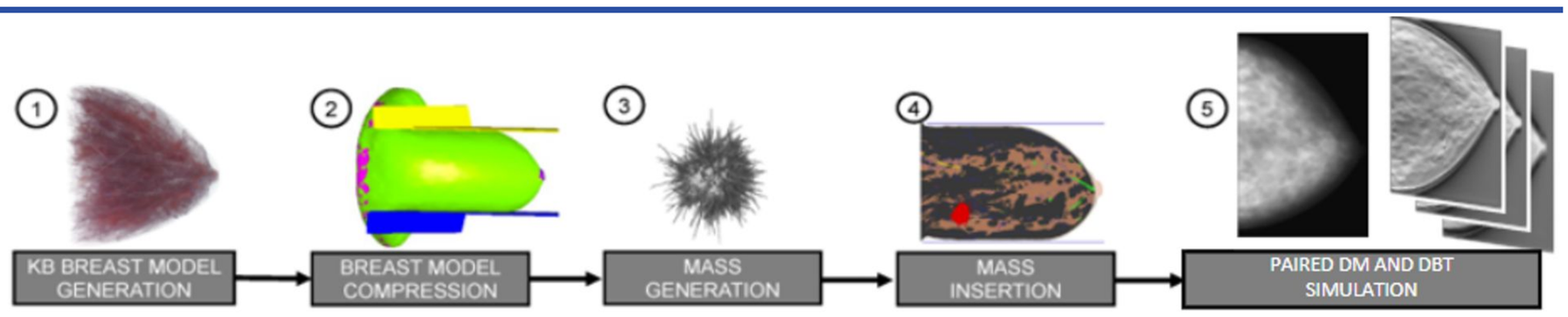
DM (left) with lesion (green), three slices of the corresponding DBT (middle), and cview image (right) for a T-SYNTH image



Example c-view image with the corresponding tissue masks available in T-SYNTH.

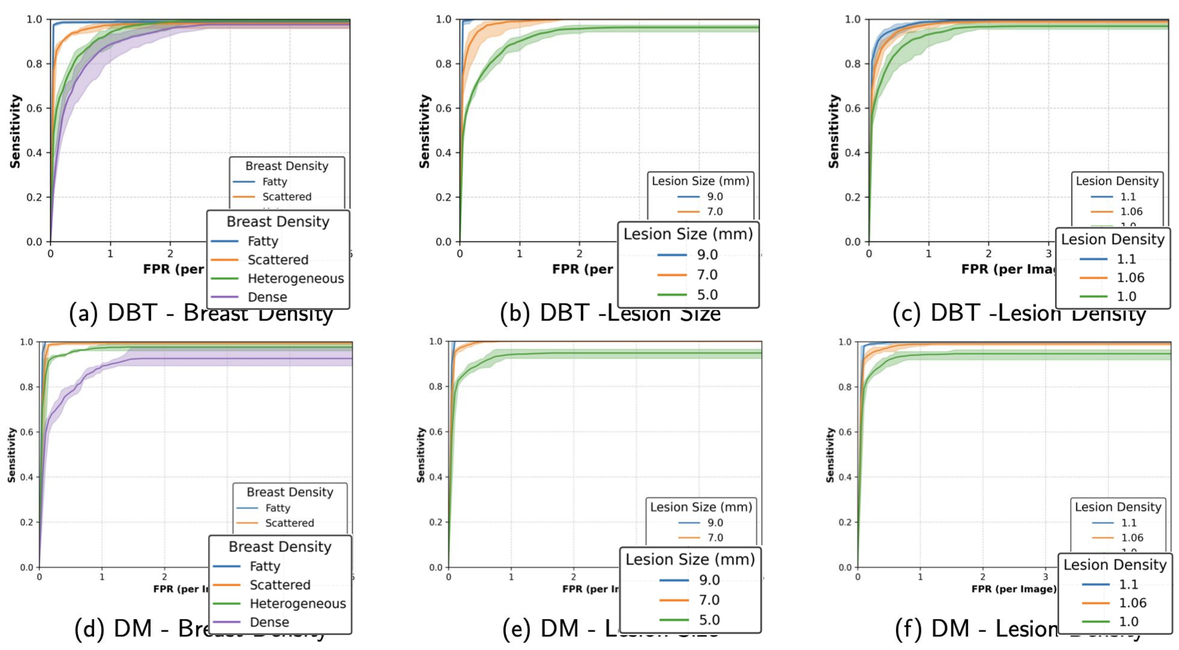
Methods

Using a KB simulation, 150 positive DBT and DM images were generated for each combination of the factors: breast density (fatty, scattered, heterogeneous, dense), lesion diameter (5, 7, 9mm), and lesion density (1.0, 1.06, 1.1x glandular density). Corresponding lesion-absent images were also synthesized. Faster R-CNN [3] was trained for lesion detection and evaluated with free-response ROC curves, with results aggregated across five trials.



A knowledge-based simulation pipeline for generating T-SYNTH. 1-2: The open source Virtual Imaging Clinical Trials for Regulatory Evaluation (VICTRE) was used to generate breast tissue models and simulate compression [5, 6]. 3-4: Lesion growth is modeled with advection-reaction-diffusion equations and accounts for surrounding tissue stiffness [5, 7]. 5: For each phantom, DM and DBT images were simulated with MC-GPU, designed to replicate Siemens Mammomat Inspiration 5. C-view images were approximated from DBT volumes using the method from Klein, et al [8].

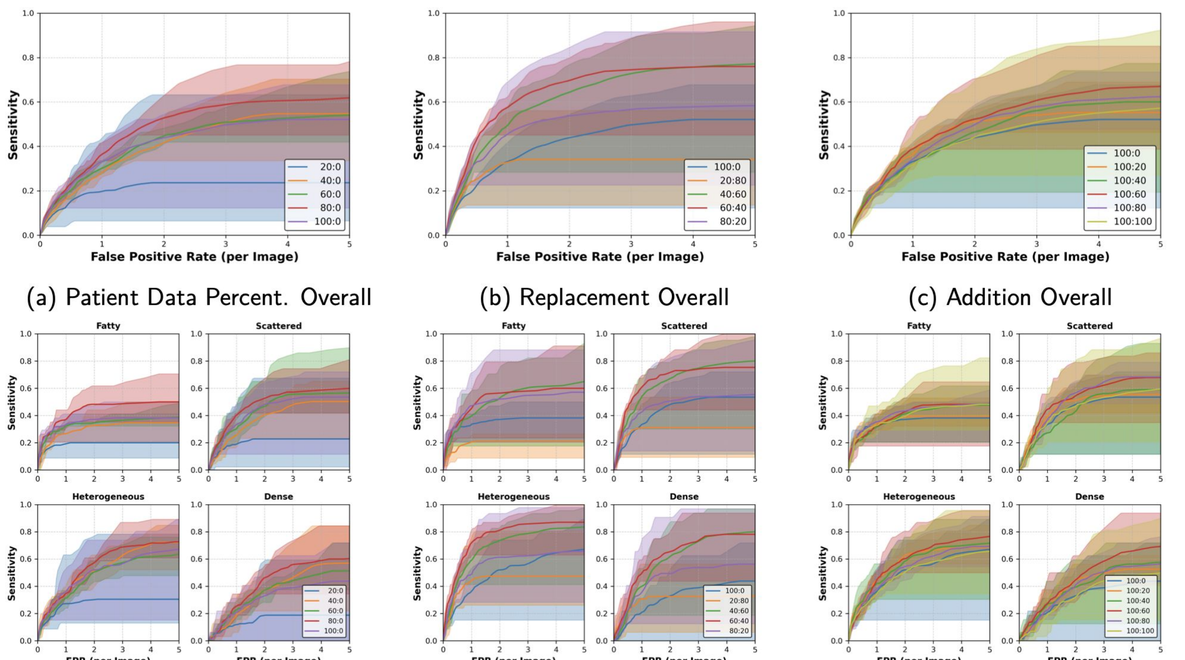
Results



(a) DBT - Breast Density (b) DBT -Lesion Size (c) DBT -Lesion Density

(d) DM - Breast Density (e) DM - Lesion Size (f) DM - Lesion Density

*FROC curves for models trained and evaluated on synthetic data (**Synthetic Subgroup Analysis**). Shaded region represents minimum and maximum bounds. Both DBT (top) and DM (bottom) shows expected trends (performance increases with decreased breast density, increased lesion size, and increased lesion density.*



(a) Patient Data Percent. Overall (b) Replacement Overall (c) Addition Overall

(d) Patient Data Percent. Subgroup (e) Replacement Subgroup (f) Addition Subgroup

*FROC curves on patient test set for models trained with a combination of patient and synthetic data (**Patient Data Augmentation**). (a) model has only a fraction of patient training data available, (b) a proportion of patient data is replaced with synthetic data, (c) a proportion of synthetic data is added to the full available patient training set. The addition of T-SYNTH data improves performance, particularly for scattered and dense breast examples.*

Contributions

- T-SYNTH is a large synthetic breast image dataset with paired DM and DBT images and a balanced representation of patient subgroups.
- The KB approach used for simulation T-SYNTH allows release of detailed pixel-level segmentation masks for lesions and other tissues
- Evaluation when training and testing a detection model on T-SYNTH yields expected performance differences among subgroups, increasing our confidence in the model's efficacy.
- All data and pretrained models and training code are publicly available at: **https://huggingface.co/datasets/didsr/tsynth** **https://github.com/DIDSR/tsynth-release**

*Contact: Elena.Sizikova@fda.hhs.gov †: Equal Contribution

References

1. Aldo Badano, et al. The stochastic digital human is now enrolling for in silico imaging trials–methods and tools for generating digital cohorts. *Progress in Biomedical Engineering*, 2023.
2. Travis Zack, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 2024.
3. Shaoqing Ren, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. 2015.
4. Jiwoong J Jeong, et al. The EMory BrEast imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiology: Artificial Intelligence*, 2023.
5. Aldo Badano, et al. Evaluation of digital breast 12 tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial. *JAMA Network Open*, 2018.
6. Elena Sizikova, et al. 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*, 2024.
7. Aunnasha Sengupta, et al. In situ tumor model for longitudinal in silico imaging trials. *Physics in Medicine & Biology*, 2024.
8. Devi Klein, et al.. A 2D synthesized image improves the 3D search for foveated visual systems. I*EEE Transactions on Medical Imaging*, 2023.