ScoreCard

1. General Information

NAME	VERSION HISTORY	RELEASE DATE
Sinkove Synthetic CSAW 100k Mammograms		October 11, 2023
DATASET SIZE	**DATASET MODALITY**	**DATASET PROVENANCE**
100k synthetic mammogram	Digital mammography	Synthetic data generated using the CSAW-M subset and MONAI framework
DATASET INTENDED USE	**DATASET LABELS**	**ATTRIBUTION AND LICENSING**
Enhance AI model training for cancer masking in mammography	"Low_Masking_Level": "score ≤ 2"; "Medium_Masking_Level": "2 < score ≤ 6"; "High_Masking_Level": "score > 6"	Released under the Open & Responsible AI license (OpenRAIL).
POINT OF CONTACT		
walter.diaz_sanz@kcl.ac.uk		

2. Data Quality Evaluation (7Cs Quantitative Results)

CONGRUENCE	COVERAGE	CONSTRAINT
See Section 1 for tables and plots	See Section 2 for tables and plots	See Section 3 for tables and plots
COMPLETENESS	**COMPLIANCE**	**COMPREHENSION**
See Section 4 for tables and plots	See Section 5 for tables and plots	See Section 6 for tables and plots
CONSISTENCY		
See Section 7 for tables and plots		

3. Task-based Evaluation (Quantitative Results)

TASK PERFORMANCE	TASK-SPECIFIC METRICS	PERFORMANCE BENCHMARK
Cancer masking classification (mixed Sinkove with CSAW for training).	Sensitivity: 0.84; Specificity: 0.81	

4. Human-based Evaluation (Qualitative Results)

STUDY DESIGN	READER STUDY RESULTS	OBSERVATIONS & FAILURE CASES
Not Provided	Not Provided	Not Provided

5. Ethical, Legal, and Practical Considerations

PRIVACY & ANONYMIZATION	BIASES	LIMITATIONS
Dataset adheres to HIPAA and privacy regulations; no personal identifiers present.	Underrepresentation of rare breast imaging features	Not provided

RECOMMENDATIONS

Combine with real datasets

6. Synthetic Dataset Usage

REPOSITORY ACCESS	PREPROCESSING REQUIREMENTS	USER DOCUMENTATION
https://huggingface.co/SinKove/synthetic_mammography_csaw	Not provided	Detailed usage instructions are in the `README.md` file

INTENDED AUDIENCE

Medical imaging researchers, AI developers.

7. Synthetic Dataset Training & Validation Process

GENERATION METHOD	VALIDATION & TESTING PROCESS
Latent Diffusion Model (LDM) with intensity rescaling, augmentation, and noise injection.	FID validated fidelity and structural similarity.

8. Reference Dataset General Information

PURPOSE	ORIGIN & SOURCE	DATASET SIZE
	Extracted from the CSAW cohort (2008-2015, Stockholm).	Public_Train: 9523; Public_Test: 497, Private_Test: 475

CLINICAL POPULATION	ACQUISITION DEVICES	REFERENCE STANDARD

Not Provided

Hologic devices at the Karolinska breast center

Expert annotation process described in the paper.

METADATA

Clinical endpoints; cancer type; image acquisition attributes; percent density

GROUND TRUTH LABELS

PREPROCESSING

632×512, 8-bit PNG format

KNOWN LIMITATIONS

Demographic gaps; incomplete metadata fields