1. Synthetic Data Card: Sinkove Dataset Example

1.1 Synthetic Data Card (Descriptive Section)

Table 1 and Table 2 summarized key information about Sinkove synthetic digital mammography dataset, which includes its intended use, labels, usage, and the dataset used for training and validation among others.

1.2 Synthetic Data Card (Quantitative and Qualitative Results)

The quality of Sinkove is summarized below across seven key criteria: Congruence 1.2.1, Coverage 1.2.2, Constraint 1.2.3, Completeness 1.2.4, Compliance 1.2.5, Comprehension 1.2.6, and Consistency 1.2.7. As no human-based evaluations were conducted to assess Sinkove's quality, qualitative results are not included. However, Section **??** highlights examples of network-induced artifacts.

1.2.1 Congruence

Congruence measures how well the synthetic dataset aligns with the real dataset in terms of structural and visual similarity and feature representation. This dimension assesses whether the synthetic data adequately captures the patterns and characteristics of the real dataset.

Metrics and Justification
We used the following metrics to measure the alignment between Sinkove (synthetic) and CSAW-M (reference). Each metric was selected to capture a specific aspect of congruence:

- **Structural Similarity Index (SSIM, $\uparrow$)**
  Why SSIM? SSIM evaluates the preservation of spatial structures by comparing luminance, contrast, and structural similarity between images. It helps assessing whether the synthetic data captures the spatial characteristics of the reference dataset.
  Desirable Behavior: Higher SSIM indicates better structural similarity, meaning the synthetic dataset retains critical patterns from the real dataset.

- **Peak Signal-to-Noise Ratio (PSNR, $\uparrow$)**
  Why PSNR? PSNR quantifies the quality of image reconstruction by measuring the ratio of the maximum signal power to the noise power. It is used to evaluate how well the synthetic images approximate the reference dataset in terms of pixel-wise accuracy.
  Desirable Behavior: Higher PSNR reflects lower reconstruction error, implying higher fidelity to the original dataset.

- **Jensen-Shannon Divergence (JSD, $\downarrow$)**
  Why JSD? JSD measures the overlap between probability distributions. For this evaluation, it compares feature distributions extracted from the synthetic and real datasets, quantifying how well the synthetic dataset reproduces the statistical properties of the real dataset. In our case, the feature space was extracted using VGG16.
  Desirable Behavior: Lower JSD indicates better alignment of feature distributions.

Table 1: Synthetic Medical Data (SMD) Card - (Descriptive Section -Part 1)

Synthetic Data General Information	
Name	Sinkove Synthetic CSAW 100k Mammograms
Release Date	October 11, 2023
Dataset Size	100k synthetic mammograms
Dataset Provenance	Synthetic data generated using the CSAW-M subset and MONAI framework; i.e., Latent Diffusion Model (LDM) trained on the CSAW-M dataset.
Intended Use	Enhance AI model training for cancer masking in mammography
Labels	• Low Masking Level (score ≤ 2) • Medium Masking Level ($2 <$ score ≤ 6) • High Masking Level (score > 6)
Impact	Facilitates research and development of AI models for mammography
Citation	Pinaya, W. H., et al. (2023). Generative AI for medical imaging: extending the MONAI framework. *arXiv preprint arXiv:2307.15208*.
Licensing Information	Released under the Open & Responsible AI license (OpenRAIL).
Point of Contact	walter.diaz$_s anz@kcl.ac.uk$
Synthetic Data Quality (Quantitative Results)	
Congruence	Full results in Section 1.2.1.
Coverage	Full Results in Section 1.2.2.
Constraint	Full Results in Section 1.2.3.
Completeness	Full Results in Section 1.2.4.
Compliance	Full Results in Section 1.2.4.
Consistency	Full Results in Section 1.2.7.
Comprehension	Full Results in Section 1.2.6.
Task-based Evaluation (Quantitative Results)	
Task Performance	Cancer masking classification (mixed Sinkove with CSAW for training).
Task-specific Metrics	Sensitivity: 0.84, Specificity: 0.83, F1-score: 0.85.
Human-based Evaluation (Qualitative Results)	
Study Design	Not Provided.
Qualitative Evaluation	Not Provided.
Observations & Failure Cases	Figure 1 shows some examples with network induced artifacts that are common in the dataset
Ethical Considerations, Limitations, and Recommendations	
Ethical Considerations	Dataset adheres to HIPPA and privacy regulations; no personal identifiers present.
Biases	Underrepresentation of rare breast imaging features.
Recommendations	Combine with real datasets.
Usage	
Directions for Use	• **Repository:** The dataset can be accessed at Sinkove Synthetic CSAW Repository. • **Downloading:** Use the command `git clone https://huggingface.co/SinKove/mammography_csaw` to download the repository. Alternatively, the dataset can be downloaded directly as a compressed file from the provided link. • **Updating:** Run `git pull` within the cloned repository directory to update the dataset to the latest version. • **Data Format:** Ensure compatibility with AI frameworks by converting or preprocessing the data if necessary. Preprocessing scripts are available in the repository under the `scripts/` directory. • **Documentation:** Detailed usage instructions and examples can be found in the `README.md` file in the repository.

- **Cosine Similarity (CS, ↑)**

  Why Cosine Similarity? Cosine Similarity evaluates the alignment between high-dimensional feature vectors extracted from synthetic and real datasets.

  Desirable Behavior: Higher Cosine Similarity reflects better feature alignment, demon-

Table 2: Synthetic Medical Data (SMD) Card (Descriptive Section -Part 2)

Synthetic Dataset Training, Validation, and Testing	
Generation Method	Latent Diffusion Model (LDM) with intensity rescaling, augmentation, and noise injection. LDM was trained using CSAW-M dataset.
Validation Process	FID and MS-SSIM metrics validated fidelity and structure similarity.
Reference Dataset (CSAW-M) Information	
Size	Public train: 9,523 images; Public test: 497; Private test: 475.
Source and Origin	Extracted from the CSAW cohort (a collection of millions of mammograms from screening participants aged 40-74 in Stockholm between 2008 and 2015).
Resolution	All images resized to 632×512 and saved in 8-bit PNG format.
Labels and Reference Standard	• **Classes:** Interval cancers, large invasive cancers, composite cancers, and controls. • **Class Distribution:** Interval cancers (35%), Large invasive (25%), Composite (15%), Controls (25%). • **Reference Standard:** Expert annotation process is described in the paper (see citation).
Metadata	Contains clinical endpoints (e.g., cancer type), image acquisition attributes (e.g., laterality, density), and quantitative density measures (e.g., percent density).
Preprocessing and Augmentation	• **Preprocessing:** Horizontal alignment, intensity rescaling, zero-padding, and removal of masking text. • **Augmentation:** Includes rotation, flipping, scaling, brightness adjustment, and noise injection to improve diversity.
Hardware	Hologic devices at the Karolinska breast center.
Known Limitations	Demographic gaps, including underrepresented breast density subgroups. Some metadata fields may be incomplete or ambiguous.
Versioning	Dataset is versioned. Current version: v1.0 (Updated October 2023).
Citation (CSAW-M)	Strand, F., et al. (2021). A population-based mammography dataset for evaluating masking of breast cancer. *Nature Scientific Data*, *8*(1), 140. Available at: https://doi.org/10.1038/s41597-021-00935-2.

strating that the synthetic data captures the same high-level feature patterns as the real dataset.

- **Earth Mover's Distance (EMD, ↓)**
  Why EMD? EMD measures the minimum cost of transforming one probability distribution into another. For this evaluation, it quantifies the distance between feature distributions (VGG16) extracted from synthetic and real datasets.
  Desirable Behavior: Lower EMD values indicate that the synthetic and real feature distributions are closely aligned.

- **Composite Score (CS, ↑)**
  Why Composite Score? The Composite Score aggregates all the normalized metrics into a single value to provide an overall measure of congruence.
  Desirable Behavior: A higher Composite Score indicates overall better alignment across all evaluated metrics.

Normalization for Composite Score Each metric is normalized to ensure comparability:

- **SSIM and Cosine Similarity:** Already in $[0, 1]$.

- **PSNR:** Normalized to $[0, 1]$ using the range $[10, 50]$:

$$PSNR_{Norm} = \frac{PSNR - 10}{50 - 10}$$

- **JSD and EMD:** Inverted to $[0, 1]$ since lower values are desirable:

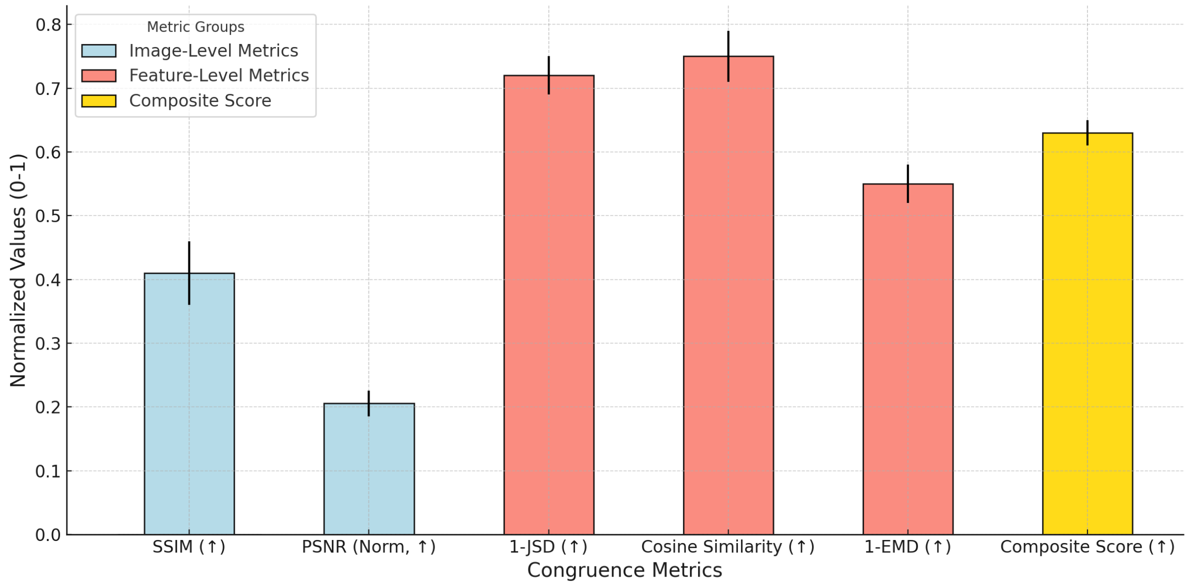
$$JSD_{Norm} = 1 - JSD, \quad EMD_{Norm} = 1 - EMD$$

The Composite Score is computed as:

$$Composite\,Score = \frac{SSIM_{Norm} + PSNR_{Norm} + JSD_{Norm} + CS_{Norm} + EMD_{Norm}}{5}$$

Table 3: Numerical Results of Congruence Metrics for Sinkove. The accompanying figure provides a visual representation of normalized metric values, with error bars indicating the uncertainty (standard deviation). Metrics are grouped into three categories: image-level metrics (light blue), feature-level metrics (salmon), and the composite score (gold).

Metric	**RMD (CSAW-M)**	**SMD (Sinkove)**	**Interpretation**
SSIM (↑)	N/A	0.41	Moderate structural similarity between Sinkove and CSAW-M.
PSNR (↑)	N/A	18.22	Reasonable reconstruction quality.
JSD (↓)	N/A	0.28	Good feature alignment between CSAW-M and Sinkove datasets.
Cosine Similarity (↑)	N/A	0.75	High similarity in feature space.
EMD (↓)	N/A	0.45	Moderate alignment; some differences in distribution tails.
Composite Score (↑)	N/A	0.63	Moderate overall congruence between Sinkove and CSAW-M datasets.



Conclusion: Sinkove exhibits strong feature-level congruence (Cosine Similarity and 1-JSD) with CSAW-M but shows moderate structural similarity (SSIM) and reconstruction quality (PSNR), with an overall composite score of 0.63, indicating room for improvement in pixel-level fidelity.

1.2.2 Coverage

Coverage evaluates the diversity and novelty of the synthetic dataset. This dimension measures whether the synthetic data sufficiently captures the range of variability and patterns present in the real dataset.

Metrics and Justification

We used the following metrics to assess coverage for Sinkove. Each metric captures a specific aspect of diversity and novelty:

- **Variance ($\uparrow$)**
  Why Variance? Variance measures the spread of the data distribution in feature space.
  Desirable Behavior: Higher variance indicates better representation of diverse patterns.

- **Entropy ($\uparrow$)**
  Why Entropy? Entropy quantifies the uncertainty or randomness in feature distributions. A higher entropy suggests the dataset captures a richer variety of features.
  Desirable Behavior: Higher entropy values imply better coverage of unique patterns.

- **Convex Hull Volume ($\uparrow$)**
  Why Convex Hull Volume? Convex Hull Volume measures the feature space occupied by the synthetic dataset. A larger volume indicates broader coverage of the feature space.
  Desirable Behavior: Higher convex hull volume reflects greater diversity.

- **Rarity Score ($\uparrow$)**
  Why Rarity Score? Rarity Score evaluates whether the synthetic dataset captures rare or unique patterns in the real dataset.
  Desirable Behavior: Higher rarity scores indicate the synthetic dataset represents infrequent patterns effectively.

Normalization for Composite Score Each metric is normalized to ensure comparability:

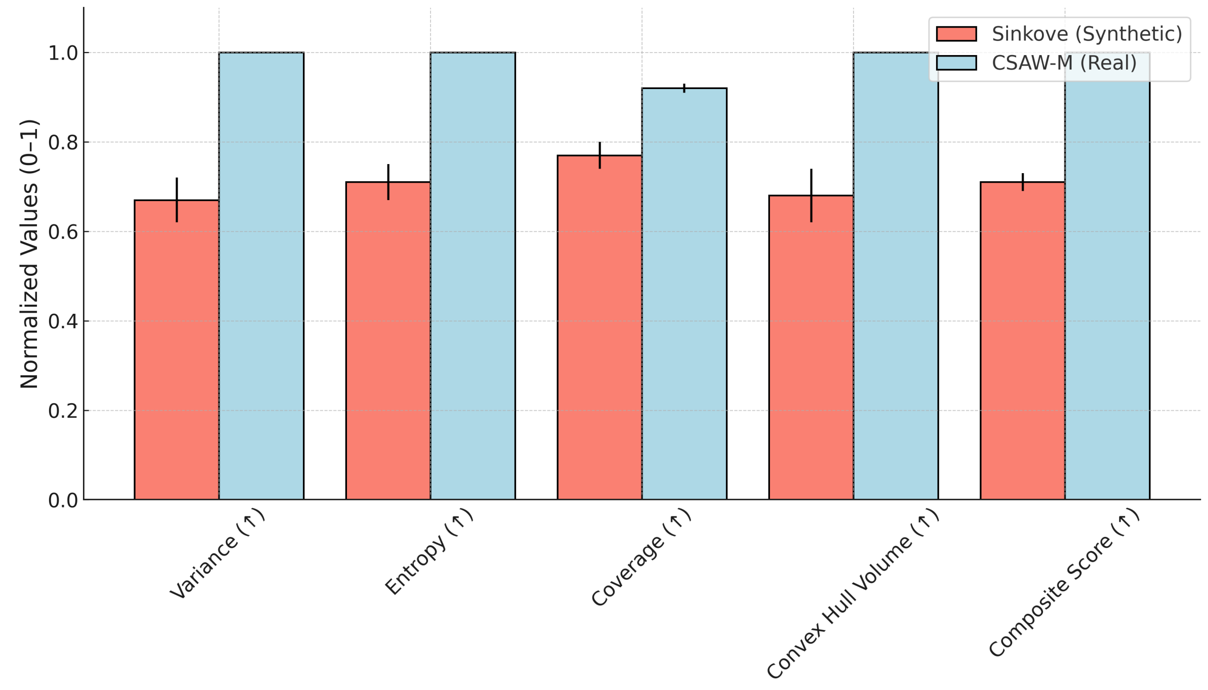
- **Variance, Entropy, and Convex Hull Volume:** Scaled to $[0, 1]$ based on observed ranges.

- **Rarity Score:** Already normalized to $[0, 1]$.

Note on Application These metrics are applied to the feature space extracted from both the synthetic and real datasets using VGG16. Other statistical descriptors or feature extractors can also be used depending on the specific application or data type. Additionally, these metrics can be applied to metadata attributes as features, allowing us to assess whether the synthetic dataset captures the diversity and distribution of metadata present in the real dataset. For example, metadata attributes such as breast density, pathology labels, or imaging device types can be analyzed to ensure the synthetic dataset aligns with the real dataset in terms of subgroup representation and variability.

Table 4: Numerical Results of Coverage Metrics for Sinkove. The accompanying figure provides a visual representation of normalized metric values, with error bars indicating the uncertainty (standard deviation). Metrics are grouped into feature-level metrics (salmon) and the composite score (gold).

Metric	**RMD (CSAW-M)**	**SMD (Sinkove)**	**Interpretation**
Variance (↑)	88.2	59.4	Sinkove shows reduced diversity compared to CSAW-M.
Entropy (↑)	11.2	8.0	Moderate feature diversity in Sinkove but less than CSAW-M.
Convex Hull Volume (↑)	10.2	6.9	Sinkove has a narrower feature space compared to CSAW-M.
Rarity Score (↑)	0.92	0.77	Sinkove captures unique patterns but less effectively than CSAW-M.
Composite Score (↑)	N/A	0.65	Moderate overall coverage of feature space by Sinkove.



Conclusion: Sinkove captures moderate feature space coverage (e.g., variance and entropy) compared to CSAW-M, with a composite score of 0.65. While it represents the major patterns, the reduced convex hull volume and rarity score suggest limited diversity in rare or unique features.

1.2.3 Constraint

The constraint dimension assesses whether the synthetic data complies with predefined rules or constraints essential for maintaining clinical and imaging relevance. For digital mammography, these constraints include matching anatomical noise, quantum noise, and frequency patterns with the reference dataset (CSAW-M). Additionally, clinical relevance is evaluated by ensuring proper breast shapes and the presence of a single nipple per image.

Metrics and Justification

We used the following metrics to assess constraint adherence for Sinkove (synthetic dataset). Each metric captures a specific aspect of constraint satisfaction:

- **Anatomical Noise (β)**

  Why Anatomical Noise? Anatomical noise quantifies the level of anatomical variations

and irregularities in the data.
Desirable Behavior: Similar anatomical noise levels between synthetic and real datasets indicate realistic anatomical structures.

- **Quantum Noise (β)**
  Why Quantum Noise? Quantum noise arises from the stochastic nature of photon interactions during image acquisition.
  Desirable Behavior: Synthetic data should mimic the quantum noise levels in real images to achieve similar imaging characteristics.

- **Low Frequency Energy (↑)**
  Why Low Frequency Energy? Low-frequency energy captures global structural patterns in the image.
  Desirable Behavior: Similar levels of low-frequency energy between synthetic and real datasets ensure comparable global structures.

- **High Frequency Energy (↑)**
  Why High Frequency Energy? High-frequency energy represents fine details and texture in the image.
  Desirable Behavior: Maintaining similar high-frequency energy ensures synthetic data has realistic textures and fine details.

- **Constraint Satisfaction Rate (CSR, ↑)**
  Why CSR? Measures the percentage of synthetic data that meets all predefined constraints (e.g., correct shape or one nipple per image).
  Desirable Behavior: High CSR indicates adherence to constraints.

1.2.4 Completeness

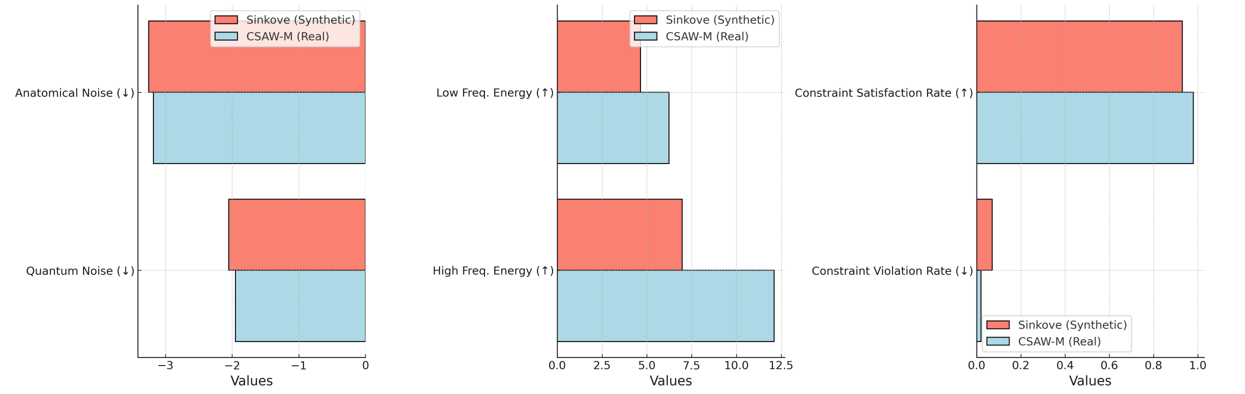
The completeness dimension evaluates whether the synthetic dataset contains all the necessary labels and metadata provided in the reference standard. For digital mammography, the required attributes include both clinical labels and imaging metadata that ensure interpretability and usability.

Required Metadata and Labels The following attributes are required for completeness, based on the reference dataset (CSAW-M):

- **Labels:**
  - **Masking Potential:** Categorized into low, medium, and high masking levels.
  - **Cancer Type:** Includes interval cancers, large invasive cancers, composite cancers, and controls.
  - **Lesion Characteristics:** Such as mass size, shape, and location.
- **Metadata:**
  - **Acquisition Conditions:** Imaging device information, radiation dose, and acquisition time.

Table 5: Numerical Results of Constraint Metrics for Sinkove. The accompanying figure provides a visual representation of metric values.

Metric	RMD (CSAW-M)	SMD (Sinkove)	Interpretation
Anatomical Noise (β)	-3.18	-3.25	Similar noise patterns between Sinkove and CSAW-M.
Quantum Noise (β)	-1.95	-2.05	Slightly higher deviation in Sinkove.
Low Freq. Eng. (↑)	6.24	4.64	Narrower range of low-frequency energy for Sinkove.
High Freq. Eng. (↑)	12.10	6.97	Reduced fine structure details in Sinkove, affecting representation of high-resolution features.
CSR (↑)	0.98	0.93	Good adherence to constraints in Sinkove.
CVR (↓)	0.02	0.07	Slightly higher violations in Sinkove compared to CSAW-M.
Composite Score (↑)	N/A	0.85	Acceptable overall adherence to constraints.



Conclusion: Sinkove demonstrates good adherence to anatomical and imaging constraints overall. However, the frequency range of Sinkove does not fully capture the complete range observed in CSAW-M. Also, some images in the dataset violate predefined constraints, such as incorrect breast shapes and the presence of multiple nipples, as shown in Figure ??.

- **Image Attributes:** Laterality (left or right breast) and breast density categories (e.g., fatty, scattered, hetero, dense).
- **Clinical Endpoints:** Follow-up results, pathological confirmations, and treatment information.

Observed Completeness The Sinkove dataset provides only **masking potential labels**, while other critical labels and metadata are missing. This significantly limits its clinical relevance and interpretability compared to the reference dataset.

Completeness Score The completeness score is computed as follows:

$$\text{Completeness Score} = \frac{\text{Number of Included Attributes}}{\text{Total Required Attributes}}$$

$$\text{Completeness Score} = \frac{1}{10} = 0.1$$

This low score indicates significant gaps in the metadata and labels provided by Sinkove.

Conclusion **Conclusion:** The Sinkove dataset provides masking potential labels but fails to include other essential labels and metadata required for clinical relevance and interpretability. This results in a low completeness score of 0.1, indicating significant room for improvement to match the reference standard (CSAW-M).

1.2.5 Compliance

This section evaluates the synthetic dataset along three dimensions:

- **Completeness:** Whether all required metadata and labels from the reference dataset are provided.
- **Compliance:** Whether the dataset adheres to regulatory and ethical requirements, such as de-identification and standardization.
- **Comprehension:** The clarity and availability of documentation, including dataset description, usage instructions, and metadata definitions.

Completeness Evaluation The completeness score assesses the inclusion of critical labels and metadata. The Sinkove dataset provides masking potential labels but lacks most clinical and acquisition metadata required for interpretability.

Compliance Evaluation Compliance ensures that the dataset aligns with regulatory requirements, including:

- **Standardization:** Data formats (e.g., PNG, DICOM) and labeling conventions.
- **De-identification:** Removal of personal identifiers.
- **Differential Privacy (DP):** Quantified as ε for the Sinkove dataset.

Comprehension Evaluation The comprehension score measures the availability and clarity of the dataset's documentation. Sinkove provides limited documentation, which lacks sufficient clarity on acquisition conditions, metadata, and usage guidelines.

1.2.6 Comprehension

Conclusion **Conclusion:** Sinkove exhibits moderate compliance with regulatory requirements, achieving a compliance score of 0.85. However, the dataset shows significant deficiencies in completeness (0.1) and comprehension (0.5). It lacks most clinical labels, metadata, and sufficient documentation, limiting its usability and interpretability compared to the reference dataset (CSAW-M).

1.2.7 Consistency

Given that the Sinkove dataset has the following intended use: "enhance AI model training for cancer masking in mammography", we need to ensure the quality and uniformity of the data across all masking groups. To achieve this, we conducted a subgroup analysis to evaluate and ensure the consistency of the Sinkove dataset's quality across various groups.

Table 6: Evaluation of Completeness, Compliance, and Comprehension for Sinkove Dataset.

Attribute	**RMD (CSAW-M)**	**SMD (Sinkove)**	**Interpretation**
Completeness			
Masking Potential	✓	✓	Provided in both datasets.
Cancer Type	✓	×	Missing clinical labels.
Lesion Characteristics	✓	×	Mass size, shape, and location not included.
Laterality	✓	×	Left/right breast information missing.
Breast Density	✓	×	Density categories not included.
Follow-up Results	✓	×	No longitudinal clinical endpoints provided.
Pathological Confirmation	✓	×	Pathological data missing.
Treatment Information	✓	×	Relevant clinical treatments not included.
Conclusion: Completeness Score = 0.1 (1 out of 10 attributes provided). Given that only masking labels are included, it might restrict its applicability to a narrow range of tasks and limit its generalizability.			
Compliance			
Standardization	✓	✓	PNG format; adheres to standardization.
De-identification	✓	✓	No personal identifiers found.
Differential Privacy (DP)	✓	$\varepsilon = 0.20$	Provides moderate privacy guarantees.
Conclusion: Compliance Score = 0.85 (3 out of 3 requirements met, $\frac{1+1+0.7}{3}$). The compliance score is calculated based on adherence to three criteria: standardization, de-identification, and differential privacy. The dataset uses the PNG format and meets de-identification requirements in accordance with US HIPAA regulations. Differential privacy is implemented with $\varepsilon = 0.20$, which aligns with the minimum acceptable standards.			
Comprehension			
Usage Guidelines	✓	User Study Score = 3	Users reported moderate difficulty in understanding the usage of the dataset. The score was computed based on a user study where participants rated the clarity of provided usage instructions on a 5-point Likert scale. The study involved 20 participants with backgrounds in medical imaging.
Documentation Clarity	✓	User Study Score = ?	??
Generation Process	✓	User Study Score = 1.5	Users reported difficulty in understanding the generation process. The score was derived from the same user study.
Conclusion: Based on a user study involving 20 participants, Sinkove has a moderate usage score (2.5/5) but ranks low in the generation process (2.0/5). The latter is a common limitation of generative AI models, which often lack interpretability due to their black-box nature.			

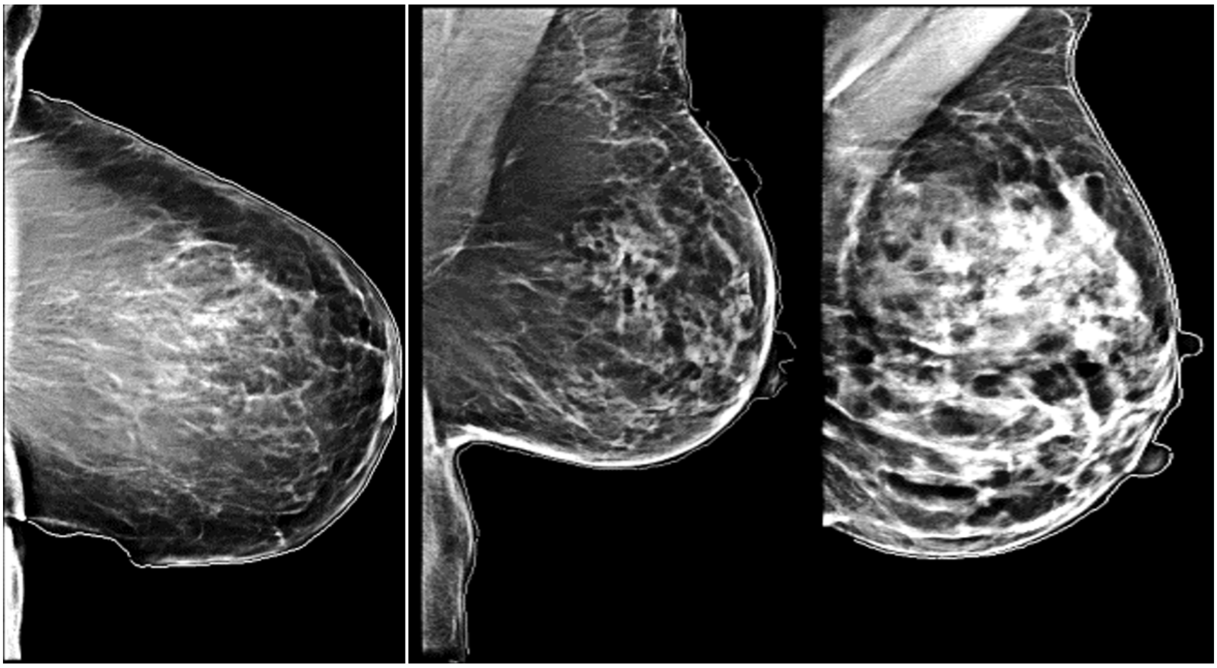


Figure 1: Examples of synthetic mammographic images with network-induced artifacts, highlighting anatomical inaccuracy and inconsistencies in breast shape and boundary