LUMINA: A Multi-Vendor Mammography Benchmark with Energy Harmonization Protocol

Hongyi Pan[1†], Gorkem Durak[1], Halil Ertugrul Aktas[1], Andrea M. Bejar[1], Baver Tutun[2], Emre Uysal[2], Ezgi Bulbul[2], Mehmet Fatih Dogan[2], Berrin Erok[2], Berna Akkus Yildirim[2], Sukru Mehmet Erturk[3], Ulas Bagci[1†]

[1]Department of Radiology, Northwestern University, Chicago, IL, USA
[2]Department of Radiation Oncology, University of Health Sciences
Prof. Dr. Cemil Tascioglu City Hospital, Istanbul, Turkey
[3]Department of Radiology, Istanbul University, Istanbul, Turkey
[†]{hongyi.pan, ulas.bagci}@northwestern.edu

Abstract

*Publicly available full-field digital mammography (FFDM) datasets remain limited in size, clinical annotations, and vendor diversity, hindering the development of robust models. We introduce **LUMINA**, a curated, multi-vendor FFDM dataset that explicitly encodes acquisition energy and vendor metadata to capture clinically relevant appearance variations often overlooked in existing benchmarks. This dataset contains 1824 images from 468 patients (960 benign, 864 malignant), with pathology-confirmed labels, BI-RADS assessments, and breast-density annotations. LUMINA spans six acquisition systems and includes both high- and low-energy imaging styles, enabling systematic analysis of vendor- and energy-induced domain shifts. To address these variations, we propose a foreground-only pixel-space alignment method ("energy harmonization") that maps images to a low-energy reference while preserving lesion morphology. We benchmark CNN and transformer models on three clinically relevant tasks: diagnosis (benign vs. malignant), BI-RADS classification, and density estimation. Two-view models consistently outperform single-view models. EfficientNet-B0 achieves an AUC of 93.54% for diagnosis, while Swin-T achieves the best macro-AUC of 89.43% for density prediction. Harmonization improves performance across architectures and produces more localized Grad-CAM responses. Overall, LUMINA provides (1) a vendor-diverse benchmark and (2) a model-agnostic harmonization framework for reliable and deployable mammography AI.*

1. Introduction

Breast cancer is the most prevalent type of cancer and one of the leading causes of cancer-related deaths among women [18, 34]. Recent advances in deep learning have demonstrated the potential to improve diagnostic accuracy and detect subtle lesions that can be overlooked by human readers [17, 30]. However, training robust and generalizable models requires large-scale, high-quality datasets. In the mammography domain, publicly available datasets such as CBIS-DDSM [14] and INbreast [20] are limited in size and heterogeneity, restricting the development of artificial intelligence (AI) systems that can perform reliably across diverse clinical settings. This underscores the need for comprehensive mammography datasets that capture high-resolution images across multiple views, patient populations, and imaging systems. To address this gap, we introduce the **LUMINA dataset** along with a foreground-only pixel-space CDF alignment method that reduces vendor/energy appearance drift and consistently improves accuracy and AUC across three tasks.

Our contribution can be summarized as follows:

- **LUMINA dataset.** Unlike prior resources that are single-vendor or film-based (e.g., CBIS-DDSM, INbreast), LUMINA is explicitly multi-vendor and energy-annotated FFDM at 12–14-bit depth, with pathology-confirmed outcomes, per-breast BI-RADS, and density labels. This combination enables controlled stress-tests for vendor/energy shifts and evaluation of single- vs. two-view modeling at clinically used resolutions. Fig. 1 presents representative benign and malignant mammograms.
- **Foreground-only histogram harmonization.** A simple,

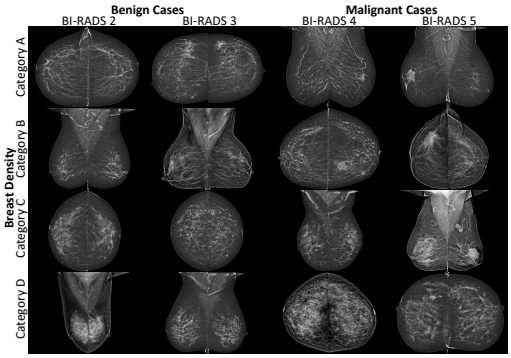


Figure 1. **Representative benign and malignant mammograms.**

vendor-agnostic histogram matching that excludes background pixels and preserves lesion contrast.

- **Three-task benchmark with multi-view modeling.** Unified evaluation on (i) pathology, (ii) BI-RADS risk grouping (binary and 3-class), and (iii) density; two-view models outperformed single-view for diagnosis, with EfficientNet-B0 reaching 93.61% AUC; EfficientNet-B0 also obtains the highest AUC (93.97%) for BI-RADS, and Swin-T yields the best AUC (89.10%) for density.
- **Harmonization improves models and attention.** Foreground histogram matching consistently raises AUC/ACC across settings (Fig. 6 in Sec. 6.3) and yields more focal Grad-CAMs around suspicious regions (Fig. 7 in Sec. 6.3), indicating better localization of lesion-related signal.

2. Related Works

Mammography datasets: The Mammographic Image Analysis Society (MIAS) database [31], the Digital Database for Screening Mammography (DDSM) [4], and its curated extension CBIS-DDSM [14] have been widely used in early computer-aided detection system development, providing digitized film mammograms with associated lesion annotations and pathology verification. However, these datasets contain relatively low-resolution screen-film mammography (SFM) scans, limiting their applicability to current clinical practice. To address these limitations, several institutions have released modern full-field digital mammography (FFDM) datasets such as INbreast [20], VinDR-Mammo [23], RSNA Screening Mammography [6], Chinese Mammography Database (CMMD) [5], CDD-CESM [5], KAU-BCMD [1]. Despite recent progress, many existing mammography datasets face limitations, including small sample sizes, reliance on film-based scans, or incomplete integration of radiological and pathological labels. Compared with these resources in Table 1, LUMINA provides high-resolution multi-view FFDM images with pathology-confirmed outcomes, expert-provided BI-RADS risk categories, and full breast density annotations. These allow LUMINA to fill critical gaps in existing datasets and establish a valuable benchmark for developing and evaluating AI algorithms in breast cancer imaging.

Mammography classification: CNN-based studies adapted architectures such as AlexNet [21], VGG [19, 25], and ResNet [7, 24] for mammography classification, showing clear improvements over handcrafted features [2, 9]. More specialized designs, such as multi-view CNNs that jointly model cranio-caudal (CC) and mediolateral oblique (MLO) views, have further enhanced diagnostic performance [17, 30]. Recently, transformer-based architectures have been explored, leveraging self-attention mechanisms to capture long-range dependencies across mammographic views and regions [3].

Medical image harmonization: Harmonizing medical images across different scanners, acquisition protocols, and vendors is critical to enabling robust and generalizable AI systems in healthcare. Statistical approaches such as ComBat [22] employ an empirical Bayes framework to correct batch-effects for MRI images by adjusting location and scale parameters of extracted imaging features, and have been shown to significantly reduce inter-site variability while preserving biological signal. Deep learning approaches, such as HarmoFL [12], further address feature drift by normalizing frequency domain amplitudes and guiding model convergence in federated learning setups across heterogeneous clients.

How we differ: Prior mammography studies typically address *one* task (diagnosis or BI-RADS or density) within a constrained imaging domain, often single-vendor FFDM or film-based datasets. In contrast, **LUMINA** couples (i) *multi-vendor* FFDM with energy metadata, (ii) *unified* evaluation across *three* clinically relevant tasks, and (iii) a *foreground-only, pixel-space* CDF alignment that reduces vendor/energy intensity drift while preserving lesion morphology. Unlike ComBat-style feature harmonizers or federated normalization methods, our approach is *model-agnostic*, runs as a lightweight pre-processing step, and consistently improves AUC and attention localization across backbones (Fig. 6 in Sec. 6.3). We further report view-ablations showing that *two-view* models outperform four-view configurations while maintaining parameter efficiency (Table 4 in Sec. 6.2).

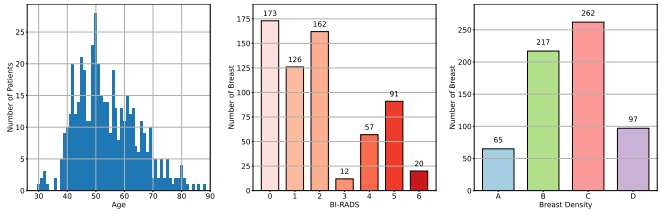
3. LUMINA Dataset Introduction

The LUMINA dataset is a curated collection of FFDM developed in collaboration with the University of Health Sciences, Prof. Dr. Cemil Tascioglu City Hospital, Istanbul, Turkey. It consists of 1,824 images from 468 patients, including 960 images from 250 benign patients and 864 images from 218 malignant patients, with malignancy confirmed by pathology. Patient ages range from 30 to 88 years, as shown in Fig. 2a. Most cases include four-view mammograms (CC and MLO views for both breasts), while

Table 1. **Comparison of publicly available mammography datasets.**

Dataset	Origin	Patients	Images	Pathology Labels	BI-RADS Assessment	Breast Density	Acquisition Technique
MIAS [31]	UK	161	322	Yes	No	No	SFM
DDSM [4]	USA	2,620	10,480	Yes	Yes	Yes	SFM
CBIS-DDSM [14]	USA	2,620	10,239	Yes	Yes	Yes	SFM
INbreast [20]	Portugal	115	410	No	Yes	Yes	FFDM
VinDR-Mammo [23]	Vietnam	5,000	20,000	No	Yes	Yes	FFDM
RSNA [6]	USA	1,970	9,594	Yes	Limited*	Yes	FFDM
CMMD [5]	China	1,775	3,728	Yes	No	No	FFDM
CDD-CESM [13]	Egypt	326	1,003	Yes	Yes	Yes	CESM
KAU-BCMD [1]	Saudi Arabia	442	1,774	No	Yes	No	FFDM
LUMINA (Ours)	**Turkey**	**468**	**1,824**	**Yes**	**Yes**	**Yes**	**FFDM**

*RSNA provides simplified BI-RADS labels (0: follow-up, 1: negative, 2: normal) instead of the full BI-RADS scores from 0 to 6.



(a) Age distribution. (b) BI-RADs distribution. (c) Density distribution.

Figure 2. **Age, BI-RADS, and breast distribution.**

a subset contains only two views due to incomplete acquisition or prior surgical history. Images are accompanied by expert-provided annotations, including BI-RADS risk assessments (categories 0–6) and breast density classifications (A–D), alongside the pathology-confirmed outcome, as summarized in Fig. 2. While the dataset contains 1,824 images in total, only 1,282 images have fully interpretable BI-RADS and density annotations. This discrepancy arises because malignancy typically appears unilaterally. Only the breast exhibiting malignancy (as confirmed by pathology) is included in the malignant class, while the contralateral breast is discarded.

The mammography images were collected from six mammography imaging systems: IMS, Metaltronica, FUJIFILM Corporation, Siemens, Carestream Health, and GE Medical Systems. Details of these vendors are summarized in Table 2. Mammograms are stored in DICOM format with native resolutions ranging from 2364×2964 to 4800×6000 pixels and a bit depth of 12–14 bits per pixel, preserving fine diagnostic details. Most images use the MONOCHROME2 format (higher pixel values correspond to brighter regions), while FUJIFILM systems use MONOCHROME1, where the intensity mapping is inverted. These MONOCHROME1 images were converted to MONOCHROME2 to ensure consistent visualization across the dataset.

Because mammograms have significant domain shifts as mentioned before, we applied a foreground-only histogram harmonization (Section 4) to align intensity distributions across vendors toward a unified low-energy style while preserving lesion integrity. Representative samples from LUMINA are shown in Fig. 1.

Table 2. **Vendor distribution.**

Vendor	# Patients	# Images	Energy
IMS	341	1326	High
Metaltronica	91	354	Low
FUJIFILM Corporation	29	116	Low
SIEMENS	4	16	Low
Carestream Health	1	4	Low
GE MEDICAL SYSTEMS	2	8	High
Total	468	1824	

Table 3. **Breast cancer diagnosis task data distribution.** Numbers outside and inside brackets are cases and images, respectively.

View	Benign	Malignant	Total
Single	592 (592)	344 (344)	936 (936)
Two	296 (592)	172 (344)	468 (936)

3.1. Evaluation Protocol

We evaluate the LUMINA dataset using three clinically relevant classification tasks: breast cancer diagnosis, BI-RADS risk assessment, and breast density prediction. All tasks are performed on standard two-view mammograms (CC and MLO) unless otherwise noted.

Breast cancer diagnosis: This is the primary task in this study, as it is clinically important for early detection and risk assessment. Models are trained using labels confirmed by pathology to distinguish benign from malignant cases. Images with BI-RADS 0 are excluded, as BI-RADS 0 indicates incomplete assessments requiring additional imaging. We evaluate single-view and two-view configurations. Only images corresponding to malignant findings are included in

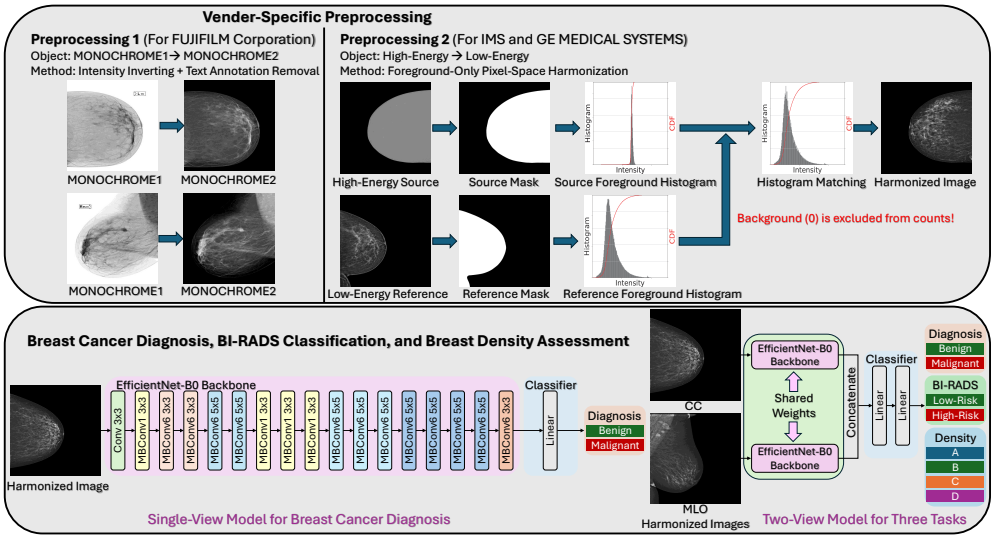


Figure 3. **LUMINA pipeline.** The EfficientNet-B0 backbone can be replaced by other backbones (ResNet-50, DenseNet-121, and Swin-T). Two-view *shared-backbone* reached accuracy comparable to independent-backbone with 48% less parameters (Table 7) and outperformed four-view variants (Table 4).

the single- and two-view malignancy subsets, and other images from the malignant patients are discarded. As Table 3 presents, 936 images are used for single-view evaluation and 468 cases (936 images) for two-view evaluation.

BI-RADS classification: This task focuses on predicting BI-RADS risk categories to mimic radiologists' assessments. Only the standard two-view images are used, as BI-RADS annotations are assigned per breast without pathological confirmation. We evaluate two schemes: (1) binary classification, distinguishing low-risk (BI-RADS 1–3) from high-risk (BI-RADS 4–6), and (2) three-class classification, grouping cases into low-risk (BI-RADS 1–2), intermediate-risk (BI-RADS 3–4), and high-risk (BI-RADS 5–6). Images with BI-RADS 0 are excluded as in the previous task. As a result, 468 cases (936 images) are used for evaluation.

Breast density assessment: Accurate breast density prediction is essential because higher-density breasts (C–D) can obscure lesions and reduce mammography sensitivity. Models are trained to predict the four density categories (A–D) using two-view images for each breast. Since BI-RADS 0 does not affect density determination, these images are included. The evaluation includes 641 cases (1,282 images), enabling development of models that reliably predict breast density across multiple vendors and support risk stratification and AI-assisted interpretation.

4. Pixel-Space Harmonization

Modern FFDM systems typically acquire images using either low-energy or high-energy X-ray settings. Low-energy mammograms are optimized for soft-tissue contrast, enhancing the visibility of subtle lesions, while high-energy mammograms emphasize denser structures such as calcification. However, these acquisition differences, combined with vendor-specific processing, introduce significant intensity and contrast variations across devices. Such inconsistencies can degrade the generalization performance of deep learning models, as models trained on one vendor's imaging style may perform poorly on another. Therefore, harmonizing intensity distributions across systems is critical for developing vendor-agnostic and clinically reliable AI models.

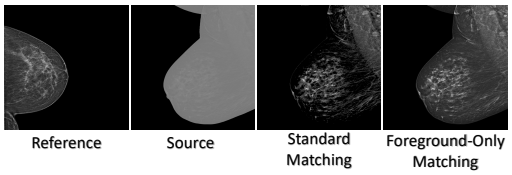


Figure 4. **Background influence.** Standard histogram matching is degraded by large black background regions, whereas foreground-only histogram matching remains unaffected.

To address these discrepancies, the LUMINA dataset was harmonized using a histogram-matching-based approach. This method aligns the intensity distribution of each image to a reference histogram, thereby improving consistency across multi-vendor datasets while preserving fine diagnostic details. Histogram matching [26, 28, 29, 33] transforms the intensity distribution of a source image to match that of a reference image. However, unlike nature images, mammograms contain large black background regions. Including these areas in the histogram can degrade the quality of the matching, as Fig. 4 shows. Therefore, these regions are excluded before applying histogram matching. Let $M_s = \{(x, y) \mid \mathbf{I}_s(x, y) > 0\}$ and $M_r = \{(x, y) \mid \mathbf{I}_r(x, y) > 0\}$ denote the sets of foreground (breast) pixels in the source image $\mathbf{I}_s$ and reference image $\mathbf{I}_r$, respectively. We compute histograms only over these foreground pixels

to avoid bias from black background regions:

$$H_s(k) = \#\{(x,y) \in M_s \mid \mathbf{I}_s(x,y) = k\}, \quad (1)$$

$$H_r(k) = \#\{(x,y) \in M_r \mid \mathbf{I}_r(x,y) = k\}, \quad (2)$$

and the total foreground pixel counts:

$$N_s = \sum_k H_s(k) = |M_s|, \quad (3)$$

$$N_r = \sum_k H_r(k) = |M_r|. \quad (4)$$

We then form the normalized cumulative distribution functions (CDFs) over the foreground intensities:

$$C_s(p) = \sum_{k=1}^{p} H_s(k), \quad \bar{C}_s(p) = \frac{C_s(p)}{N_s}, \quad (5)$$

$$C_r(q) = \sum_{k=1}^{q} H_r(k), \quad \bar{C}_r(q) = \frac{C_r(q)}{N_r}, \quad (6)$$

where the summation starts from intensity 1, as intensity 0 represents the background and is excluded. The mapping function $\mathcal{T}(\cdot)$ is defined by matching these CDFs:

$$\mathcal{T}(p) = \arg\min_{q \in \{1,\ldots,2^b-1\}} \left|\bar{C}_s(p) - \bar{C}_r(q)\right|. \quad (7)$$

Finally, the harmonized image $\mathbf{I}_o$ is obtained by applying $\mathcal{T}$ to foreground pixels and leaving background pixels as zero:

$$\mathbf{I}_o(x,y) = \begin{cases} \mathcal{T}\big(\mathbf{I}_s(x,y)\big), & (x,y) \in M_s, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

Fig. 3 illustrates an example: the source image exhibits high contrast, and its CDF is adjusted to follow that of the reference image, resulting in improved intensity consistency.

Practical notes: We derive the reference histogram from a representative subset of low-energy FFDM images to stabilize soft-tissue contrast. Foreground masking is computed by thresholding at intensity > 0 after MONOCHROME1$\rightarrow$2 conversion to avoid background bias (Fig. 3). Histogram computation uses 12-bit bins to preserve detail. This pre-processing is model-agnostic and applied identically across tasks and backbones. The harmonization algorithm is summarized in Algorithm 1.

Algorithm 1 Foreground-only CDF Matching for Vendor/Energy Harmonization

Require: Source image $\mathbf{I}_s$, reference image $\mathbf{I}_r$ (low-energy style), bit depth b

Convert MONOCHROME1 to MONOCHROME2 if needed; remove textual burn-ins (see Fig. 3).

Foreground masks: $M_s \leftarrow \{(x,y) \mid \mathbf{I}_s(x,y) > 0\}$, $M_r \leftarrow \{(x,y) \mid \mathbf{I}_r(x,y) > 0\}$

Compute histograms over foreground: $H_s, H_r \in \mathbb{N}^{2^b}$; set $H_s(0) = H_r(0) = 0$

Compute normalized CDFs $\bar{C}_s, \bar{C}_r$ (Eqs. (5)–(6))

for $p = 1$ **to** $2^b - 1$ **do**

  $\mathcal{T}(p) \leftarrow \arg\min_{q \in \{1,\ldots,2^b-1\}} |\bar{C}_s(p) - \bar{C}_r(q)|$ {monotone mapping}

end for

Initialize $\mathbf{I}_o \leftarrow \mathbf{0}$; **for** $(x,y) \in M_s$: $\mathbf{I}_o(x,y) \leftarrow \mathcal{T}(\mathbf{I}_s(x,y))$

return $\mathbf{I}_o$

5. Deep Learning Approaches for Multi-View Mammography Classification

We adopt EfficientNet-B0 [32] as the representative backbone, though alternative architectures such as ResNet-50 [10], DenseNet-121 [11], and Swin-T [15] can also be utilized. All backbones are pretrained on the ImageNet-1K dataset [8] to provide robust initialization. In the single-view setting, the model is fine-tuned through conventional transfer learning. In the two-view setting, the CC and MLO images of the same breast are independently processed through a shared-weight backbone. The resulting feature embeddings are concatenated and passed through a series of fully connected layers for classification. This shared-weight configuration enables efficient parameter utilization and promotes consistent feature learning across views while leveraging complementary information from both projections for improved diagnostic accuracy. The BI-RADS classification and breast density prediction models follow the same two-view architecture, differing only in the output layer of the classification head.

6. Experimental Results

6.1. Experimental Setup

Baseline fairness: All backbones used the same data splits, augmentations, and training schedule to ensure comparability. We applied horizontal flip and resizing, replicate grayscale to three channels, with identical optimization settings across tasks. CNN backbones (ResNet-50, DenseNet-121, EfficientNet-B0) were trained by AdamW optimizer [16] with an initial learning rate of 1×10^{-3}, while Swin-T started with 1×10^{-5}. All models were trained for 100 epochs with a weight decay of 1×10^{-5}, the learning rate decayed by a factor of 0.1 every 30 epochs, and the selection of the model by the best validation AUC under 5-fold cross validation. This protocol avoids method-specific tricks and isolates architectural effects. PyTorch and CUDA determinism flags are enabled for all reported runs. All experiments were implemented in PyTorch on a server with 8 NVIDIA A6000 GPUs. The LUMINA dataset and source code will be released upon paper acceptance.

Evaluation: We reported accuracy (ACC), area under the ROC curve (AUC), F1-score, sensitivity (recall), precision,

Table 4. **Mammogram cancer diagnosis benchmark.** The best and second-best results are highlighted in green and yellow, respectively.

View	Model	Input	Params	Flops	ACC(%)	AUC(%)	Precision(%)	Recall(%)	F1(%)	Specificity(%)
Single	ResNet-50	224^2	23.51M	4.13G	79.49±5.99	87.42±4.21	**81.29±10.90**	61.97±18.17	67.52±13.09	**89.68±9.63**
	DenseNet-121	224^2	6.95M	2.90G	79.49±3.37	87.14±3.08	80.45±9.45	62.55±17.56	67.79±9.57	89.38±7.52
	EfficientNet-B0	224^2	4.01M	0.41G	**83.12±4.13**	**90.66±4.21**	79.78±6.31	**73.89±15.77**	**75.50±7.95**	88.51±5.00
	Swin-T	224^2	27.52M	3.13G	81.84±5.92	88.62±5.18	78.27±8.63	71.83±12.10	74.16±8.36	87.66±7.45
	ResNet-50	512^2	23.51M	21.58G	77.68±7.18	86.69±4.37	82.79±14.09	59.36±28.73	61.41±22.78	88.38±12.29
	DenseNet-121	512^2	6.95M	15.14G	79.38±5.55	90.52±3.61	81.32±16.39	68.09±21.96	69.22±11.98	85.99±14.42
	EfficientNet-B0	512^2	4.01M	2.13G	**86.43±3.85**	**92.13±4.21**	**86.61±5.58**	**75.03±9.46**	**80.00±6.35**	**93.08±3.31**
	Swin-T	512^2	27.52M	16.47G	84.94±3.57	91.34±4.22	85.33±4.48	71.55±10.46	77.38±6.35	92.73±2.55
Two	ResNet-50	224^2	24.03M	8.26G	76.70±7.80	88.71±2.82	87.99±14.28	48.87±27.42	55.53±24.63	92.89±9.12
	DenseNet-121	224^2	7.22M	5.80G	84.40±1.64	89.86±1.21	**88.67±5.88**	66.81±7.49	75.68±3.62	**94.59±3.63**
	EfficientNet-B0	224^2	4.34M	0.82G	**86.11±2.47**	**92.99±3.04**	86.83±2.93	**73.26±5.63**	**79.39±4.18**	93.58±1.26
	Swin-T	224^2	27.72M	6.26G	81.22±9.77	91.99±3.40	63.41±32.06	66.57±33.56	64.92±32.73	89.88±5.93
	ResNet-50	512^2	24.03M	43.16G	72.62±14.72	87.56±5.14	70.66±16.28	71.45±27.02	64.29±17.57	73.23±30.35
	DenseNet-121	512^2	7.22M	30.29G	72.46±11.81	88.33±5.46	64.93±13.99	**79.01±13.83**	68.88±5.96	68.56±25.60
	EfficientNet-B0	512^2	4.34M	4.27G	**85.04±5.92**	**93.54±3.88**	**91.34±11.37**	69.87±22.03	75.75±12.85	**93.91±9.75**
	Swin-T	512^2	27.72M	32.94G	83.96±4.25	92.42±4.27	80.92±10.00	76.15±9.13	**77.67±5.36**	88.50±6.91

Table 5. **BI-RADS classification benchmark.**

Model	Input	Two-Class (BI-RADS 1/2/3 VS 4/5/6)			Three-Class (BI-RADS 1/2 VS 3/4 VS 5/6)		
		ACC(%)	AUC(%)	F1(%)	ACC(%)	AUC(%)	F1(%)
ResNet-50	224^2	71.98±12.34	88.23±3.52	62.45±10.33	69.02±5.14	79.58±3.00	**55.36±1.97**
DenseNet-121	224^2	77.55±3.67	87.89±3.52	58.98±14.70	68.15±3.12	80.90±3.74	48.63±3.49
EfficientNet-B0	224^2	**83.56±4.76**	**92.80±4.13**	**74.18±9.32**	**71.78±5.03**	**83.27±4.91**	55.15±4.65
Swin-T	224^2	79.29±9.01	90.80±4.61	60.82±31.02	70.51±4.45	81.69±5.55	49.55±6.85
ResNet-50	512^2	75.22±4.46	89.08±5.42	64.78±9.96	66.66±3.10	78.92±4.91	50.71±3.73
DenseNet-121	512^2	74.56±10.73	88.76±3.95	59.09±24.68	64.97±5.95	80.54±4.29	52.50±3.48
EfficientNet-B0	512^2	**85.48±3.04**	**92.75±4.10**	**76.83±5.45**	**71.13±5.22**	**82.85±4.41**	**55.73±3.22**
Swin-T	512^2	83.31±6.21	91.44±4.41	76.29±7.12	70.07±5.02	81.68±5.88	47.98±5.66

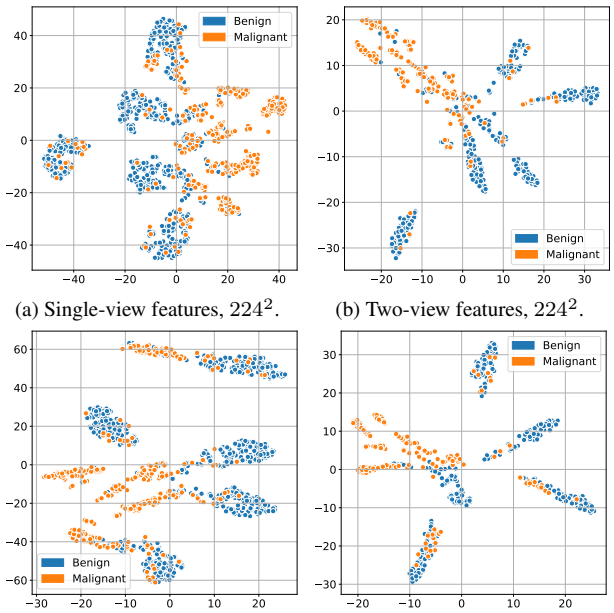
Table 6. **Density classification benchmark.**

Model	Input	ACC(%)	AUC(%)	F1(%)
ResNet-50	224^2	64.74±3.33	85.66±1.62	56.94±6.43
DenseNet-121	224^2	66.29±4.20	86.39±2.92	56.70±2.97
EfficientNet-B0	224^2	59.43±5.23	84.72±3.54	48.25±5.68
Swin-T	224^2	**70.35±4.60**	**89.43±2.22**	**65.28±6.48**
ResNet-50	512^2	65.37±4.07	87.06±2.10	58.29±7.57
DenseNet-121	512^2	66.31±5.31	86.60±3.76	61.42±7.62
EfficientNet-B0	512^2	67.54±3.89	86.83±2.80	**63.10±4.90**
Swin-T	512^2	**69.41±7.23**	**89.14±3.36**	61.50±12.60

and specificity for breast cancer diagnosis, and ACC, AUC, and F1 for the three-class BI-RADS and density classification (macro-AUC and macro-F1 for multi-class). AUC is considered the primary metric from a clinical perspective. Results were reported as mean±std over 5 folds.

6.2. Benchmark Results

Breast cancer diagnosis: Table 4 shows that EfficientNet-B0 consistently outperformed ResNet-50 and DenseNet-121 in terms of AUC, indicating its strong capability for mammographic image analysis. For instance, single-view EfficientNet-B0 achieves 90.66% AUC at 224^2 and 92.13% AUC at 512^2, outperforming both ResNet-50 (87.42% / 86.69%) and DenseNet-121 (87.14% / 90.52%). Higher input resolutions generally improve performance, highlighting the importance of preserving fine mammography details. Nevertheless, 224^2 inputs still provide competitive results, offering computational efficiency. For example, single-view EfficientNet-B0 has 4.01M parameters and 0.41G FLOPs at 224^2, compared to 4.01M parameters and 2.13G FLOPs at 512^2. Two-view models outperform single-view models across all backbones, highlighting the value of combining CC and MLO views. Notably, two-view EfficientNet-B0 with 512^2 inputs achieves the highest overall performance, reaching an AUC of 93.54%, rep-



(a) Single-view features, 224^2. (b) Two-view features, 224^2.

(c) Single-view features, 512^2. (d) Two-view features, 512^2.

Figure 5. **t-SNE visualization for diagnosis task.** Higher resolution increases benign/malignant separation in the latent space.

resenting the most effective configuration in our benchmark. Fig. 5 illustrates the t-SNE visualization, which shows how EfficientNet-B0 learned representations separate benign and malignant cases more clearly at higher input resolution.

BI-RADS classification: The BI-RADS classification benchmarks are summarized in Table 5. Across all experimental settings, EfficientNet-B0 achieved the highest AUCs (92.80% in the two-class classification and 83.27% in the three-class classification), followed by Swin-T, which consistently ranked second (91.44% and 81.68%).

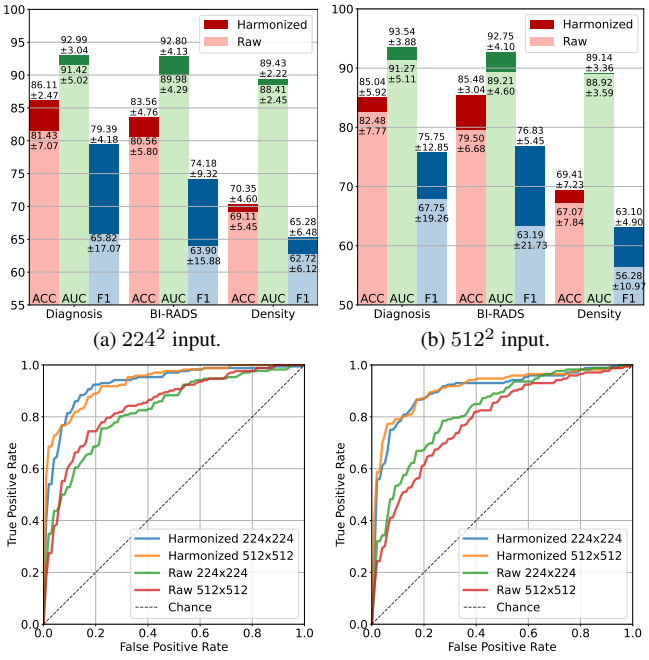
Density classification: As presented in Table 6, Swin-T always achieved the highest AUC across both input resolutions, demonstrating its effectiveness for multi-class density prediction. With 512^2 inputs, Swin-T achieved a Macro-AUC of 89.14%, outperforming ResNet-50 (87.06%), DenseNet-121 (86.39%), and EfficientNet-B0 (84.72%). While higher resolutions generally improved the performance of CNN-based models, Swin-T obtained a slightly higher AUC (89.43%) at 224^2 resolution. We attribute this to the fact that the publicly available pretrained weights (from PyTorch Torchvision) were trained with 224^2 inputs, and Transformer-based attention mechanisms are particularly sensitive to input scale.

6.3. Ablation Study And Discussion

Shared vs. independent backbone: We compared two-view models using either shared or independent backbone weights. In the independent setting, the two EfficientNet-B0 branches are initialized and trained separately, while

Table 7. **Shared vs. independent two-view EfficientNet-B0 backbones.** Flops are 0.82G for 224^2 and 4.27G under 512^2.

Backbone	Input	Params	ACC(%)	AUC(%)	F1(%)
Independent	224^2	8.34M	84.82±3.00	93.54±2.38	77.41±6.13
	512^2	8.34M	88.46±4.91	92.71±4.63	82.19±8.61
Shared	224^2	4.34M	86.11±2.47	92.99±3.04	79.39±4.18
	512^2	4.34M	85.04±5.92	93.54±3.88	75.75±12.85



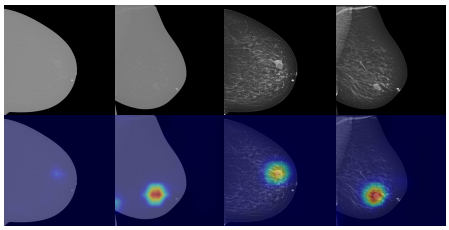
(a) 224^2 input. (b) 512^2 input.

(c) Two-view diagnosis. (d) Two-class BI-RADS.

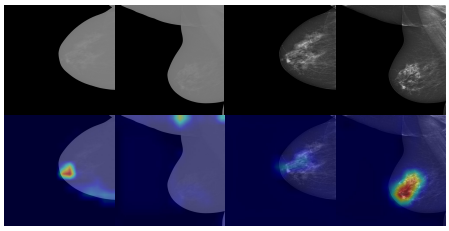
Figure 6. **Harmonization improves ACC, AUC, and F1-Score across tasks.** Bars show mean±std over 5 folds, and curves show the mean ROC. *How to read:* In (a) and (b), darker bars (harmonized) consistently exceed lighter bars (raw). In (c) and (d), areas under the blue and orange (harmonized) curves are respectively larger than areas under the green and red (raw) curves. *Takeaway:* Pixel-space CDF alignment helps regardless of backbone or input size, supporting its use as a robust pre-processing step.

in the shared setting, both views share the same backbone parameters. Table 7 shows that although the computational cost between the two approaches is identical, sharing weights substantially reduces model size from 8.34M to 4.34M parameters without sacrificing performance. The shared-backbone EfficientNet-B0 achieved 92.99% mean AUC at 224^2 and 93.54% at 512^2, comparable to or slightly better than the independent-backbone configuration, which reached 93.54% and 92.71%. These results indicate that shared backbones are sufficient for this task, which offers a favorable trade-off between efficiency and AUC.

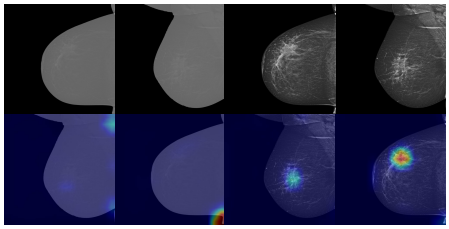
Effect of foreground-only histogram harmonization: We compared the performance of the best model in each task on raw mammograms versus harmonized mammograms.



(a) Malignant case 57L.



(b) Malignant case 83R.



(c) Malignant case 105R.

Figure 7. **Attention becomes more focal after harmonization.** Shown are raw (left two columns) vs. harmonized (right two columns) inputs with Grad-CAM overlays for two malignant cases. *Observation:* Harmonization reduces diffuse activations and concentrates attention on lesion-bearing regions. *Implication:* Harmonization not only improves metrics (Fig. 6) but may also enhance clinical interpretability by focusing on suspicious tissue.

In this section, the conversion from MONOCHROME1 to MONOCHROME2 was still applied. As shown in Fig. 6, histogram matching consistently improved the performance across all settings. These results highlight that aligning intensity distributions across multi-vendor mammograms enhances the model's discrimination between benign and malignant cases, making histogram matching an effective preprocessing step for robust pathology classification.

Furthermore, we visually compared Grad-CAM attention maps [27] generated by the breast cancer diagnosis classifier on raw and histogram-harmonized two-view mammograms (512^2). As shown in Fig. 7, harmonization improves the spatial focus of the model's activations, directing attention toward clinically relevant regions. In Fig. 7a, the model trained on raw images attended primarily to the MLO view, whereas the harmonized model successfully identified the suspicious area in both CC and MLO views. In Fig. 7b and Fig. 7c, the raw-image model's attention was distracted by the chest wall above or below the breast, while harmonization guided the model toward the actual breast tissue, better aligning with lesion-related structures.

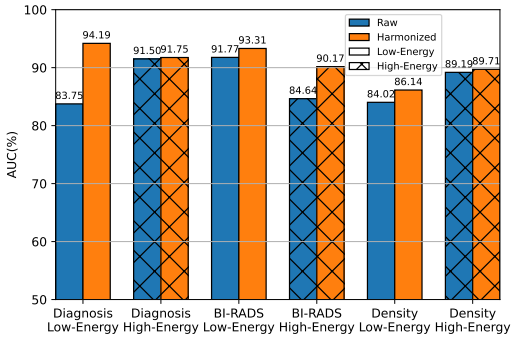


Figure 8. **Energy-specific AUC.** Histogram harmonization improved AUC, particularly for low-energy images.

Energy-specific analysis: To evaluate energy-specific performance, predictions and labels from five patient-specific folds were concatenated to obtain full-dataset predictions. AUC was then computed separately for low- and high-energy subsets using the best models for each task (EfficientNet-B0 for two-view diagnosis and two-class BI-RADS, and Swin-T for density). All models were with 224^2 inputs. As shown in Fig. 8, the histogram harmonization reduces the distribution gap between the two energy types, improving generalization. Notably, when high-energy images dominate the dataset, low-energy images benefit most from harmonization.

7. Conclusion

We present LUMINA, a multi-vendor FFDM dataset with comprehensive pathology, BI-RADS, and density annotations. We also introduced a foreground-only histogram-based harmonization method to mitigate inter-scanner variability. Extensive experiments demonstrate that harmonization consistently improves performance across multiple tasks and architectures. These findings highlight the importance of vendor-level normalization for robust mammography AI. We expect LUMINA to bridge the gap between academic benchmarks and real-world clinical deployment.

8. Data and Code

The LUMINA dataset and the associated source code are publicly available at the following links:

- **OSF (Dataset):** https://osf.io/b63jc/
- **Kaggle (Dataset):** https://www.kaggle.com/datasets/phy710/lumina-mammography-dataset
- **GitHub (Code):** https://github.com/NUBagciLab/LUMINA

9. Acknowledgment

This research was partially funded by NIH: R01-HL171376.

References

[1] Asmaa S Alsolami, Wafaa Shalash, Wafaa Alsaggaf, Sawsan Ashoor, Haneen Refaat, and Mohammed Elmogy. King abdulaziz university breast cancer mammogram dataset (kau-bcmd). *Data*, 6(11):111, 2021. 2, 3

[2] John Arevalo, Fabio A González, Raúl Ramos-Pollán, Jose L Oliveira, and Miguel Angel Guevara Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer methods and programs in biomedicine*, 127:248–257, 2016. 2

[3] Gelan Ayana, Kokeb Dese, Yisak Dereje, Yonas Kebede, Hika Barki, Dechassa Amdissa, Nahimiya Husen, Fikadu Mulugeta, Bontu Habtamu, and Se-Woon Choe. Vision-transformer-based transfer learning for mammogram classification. *Diagnostics*, 13(2):178, 2023. 2

[4] Kevin Bowyer, David Kopans, William P Kegelmeyer, R Moore, M Sallam, K Chang, and K Woods. The digital database for screening mammography. In *Third international workshop on digital mammography*, page 27, 1996. 2, 3

[5] Hongmin Cai, Jinhua Wang, Tingting Dan, Jiao Li, Zhihao Fan, Weiting Yi, Chunyan Cui, Xinhua Jiang, and Li Li. An online mammography database with biopsy confirmed types. *Scientific Data*, 10(1):123, 2023. 2, 3

[6] C Carr, Felipe Kitamura, G Partridge, J Kalpathy-Cramer, J Mongan, K Andriole, Vazirabad M Lavender, M Riopel, R Ball, S Dane, et al. Rsna screening mammography breast cancer detection. *Kaggle*, 2022. 2, 3

[7] Yuanqin Chen, Qian Zhang, Yaping Wu, Bo Liu, Meiyun Wang, and Yusong Lin. Fine-tuning resnet for breast cancer classification from mammography. In *The international conference on healthcare science and engineering*, pages 83–96. Springer, 2018. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[9] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. Deep learning and structured prediction for the segmentation of mass in mammograms. In *International Conference on Medical image computing and computer-assisted intervention*, pages 605–612. Springer, 2015. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[12] Meirui Jiang, Zirui Wang, and Qi Dou. Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1087–1095, 2022. 2

[13] Rana Khaled, Maha Helal, Omar Alfarghaly, Omnia Mokhtar, Abeer Elkorany, Hebatalla El Kassas, and Aly Fahmy. Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. *Scientific data*, 9(1):122, 2022. 3

[14] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9, 2017. 1, 2, 3

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[17] William Lotter, Abdul Rahman Diab, Bryan Haslam, Jiye G Kim, Giorgia Grisot, Eric Wu, Kevin Wu, Jorge Onieva Onieva, Yun Boyer, Jerrold L Boxerman, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nature medicine*, 27(2):244–249, 2021. 1, 2

[18] Sergiusz Łukasiewicz, Marcin Czeczelewski, Alicja Forma, Jacek Baj, Robert Sitarz, and Andrzej Stanisławek. Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. *Cancers*, 13(17):4287, 2021. 1

[19] Sidratul Montaha, Sami Azam, Abul Kalam Muhammad Rakibul Haque Rafid, Pronab Ghosh, Md Zahid Hasan, Mirjam Jonkman, and Friso De Boer. Breastnet18: a high accuracy fine-tuned vgg16 model evaluated using ablation study for diagnosing breast cancer from enhanced mammography images. *Biology*, 10(12):1347, 2021. 2

[20] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012. 1, 2, 3

[21] Emmanuel Lawrence Omonigho, Micheal David, Achonu Adejo, and Saliyu Aliyu. Breast cancer: tumor detection in mammogram images using modified alexnet deep convolution neural network. In *2020 international conference in mathematics, computer engineering and computer science (ICMCECS)*, pages 1–6. IEEE, 2020. 2

[22] Fanny Orlhac, Jakoba J Eertink, Anne-Segolene Cottereau, Josee M Zijlstra, Catherine Thieblemont, Michel Meignan, Ronald Boellaard, and Irène Buvat. A guide to combat harmonization of imaging biomarkers in multicenter studies. *Journal of Nuclear Medicine*, 63(2):172–179, 2022. 2

[23] Hieu Huy Pham, H 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. 2, 3

[24] T Sathya Priya and T Ramaprabha. Resnet based feature extraction with decision tree classifier for classificaton of mammogram images. *Turkish Journal of Computer and Mathematics Education*, 12(2):1147–1153, 2021. 2

[25] Vinoth Rathinam, R Sasireka, and K Valarmathi. An adaptive fuzzy c-means segmentation and deep learning model for efficient mammogram classification using vgg-net. *Biomedical Signal Processing and Control*, 88:105617, 2024. 2

[26] Jannick P Rolland, V Vo, B Bloss, and Craig K Abbey. Fast algorithms for histogram matching: Application to texture synthesis. *Journal of Electronic Imaging*, 9(1):39–45, 2000. 4

[27] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[28] Dori Shapira, Shai Avidan, and Yacov Hel-Or. Multiple histogram matching. In *2013 IEEE international conference on image processing*, pages 2269–2273. IEEE, 2013. 4

[29] Dinggang Shen. Image registration by local histogram matching. *Pattern Recognition*, 40(4):1161–1172, 2007. 4

[30] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):12495, 2019. 1, 2

[31] John Suckling. The mammographic images analysis society digital mammogram database. In *Exerpta Medica. International Congress Series, 1994*, pages 375–378, 1994. 2, 3

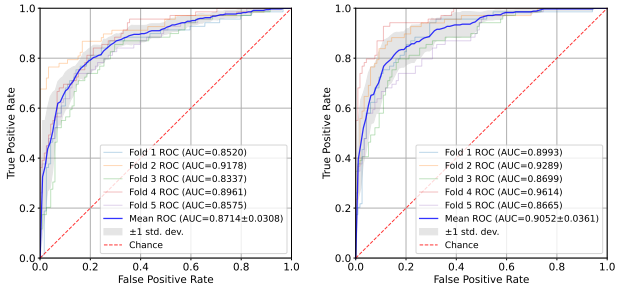
[32] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[33] Liangping Tu and Changqing Dong. Histogram equalization and image feature matching. In *2013 6th International Congress on Image and Signal Processing (CISP)*, pages 443–447. IEEE, 2013. 4

[34] Louise Wilkinson and Toral Gathani. Understanding breast cancer as a global health concern. *The British journal of radiology*, 95(1130):20211033, 2022. 1

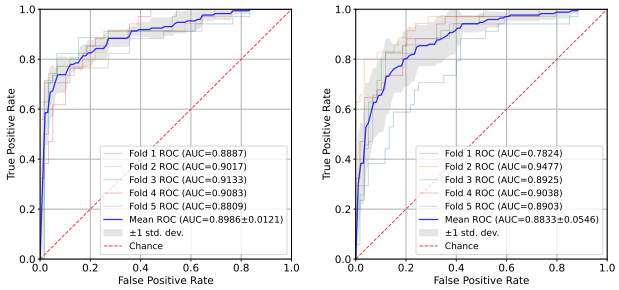
LUMINA: A Multi-Vendor Mammography Benchmark with Energy Harmonization Protocol

Supplementary Material



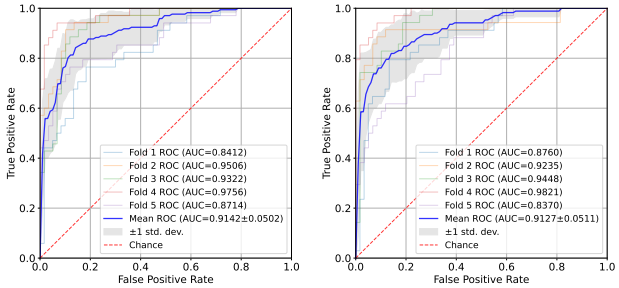
(a) Single-view 224^2 harmonized.

(b) Single-view 512^2 harmonized.



(c) Two-view 224^2 harmonized.

(d) Two-view 512^2 harmonized.

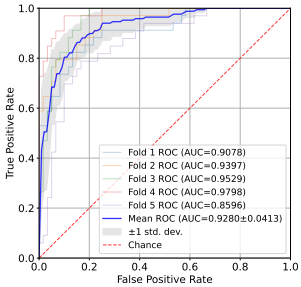


(e) Two-view 224^2 raw input.

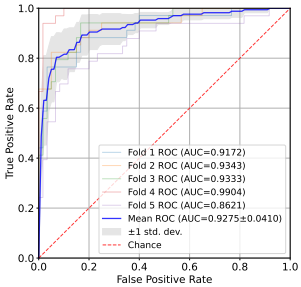
(f) Two-view 512^2 raw input.

Figure 9. **ROC curves of the EfficientNet-B0 model for breast cancer diagnosis.**

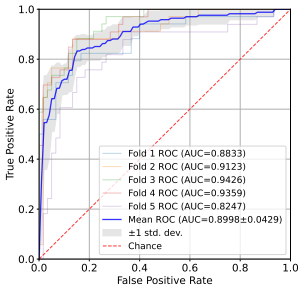
Fig. 9 shows the ROC curves of EfficientNet-B0 for the breast cancer diagnosis task, where we plot the ROC for each fold and the mean ROC. The two-view model consistently outperforms the single-view variant, demonstrating the benefit of combining CC and MLO information. Higher input resolution (512^2) further improves AUC compared to the 224^2 setting. In all configurations, object-histogram-based harmonization yields a noticeable upward shift in the ROC curves, confirming its effectiveness in stabilizing image appearance and enhancing diagnostic performance. Fig. 10 shows the ROC curves of EfficientNet-B0 for the two-class BI-RADS classification task. It also supports that



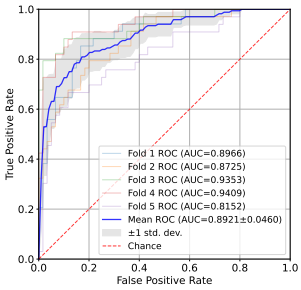
(a) 224^2 harmonized input.



(b) 512^2 harmonized input.



(c) 224^2 raw input.



(d) 512^2 raw input.

Figure 10. **ROC curves of the EfficientNet-B0 model for two-class BI-RADS classification.**

higher resolution and harmonization improves the results.

Table 8 lists the distinguishing characteristics of LUMINA compared to existing mammography datasets and prior harmonization strategies.

Table 9 summarizes the unified hyperparameter settings and pre-processing operations used across all model architectures, respectively. These settings and operations ensured a controlled and fair comparison in our benchmark experiments.

Table 8. **Positioning.** LUMINA provides pathology, BI-RADS, and density on multi-vendor FFDM, plus a simple, effective pixel-space harmonization baseline.

Family	Examples	Labels/Supervision	Architecture/Objective	Limitations vs. LUMINA
Film (SFM) datasets	MIAS [31], DDSM [4], CBIS-DDSM [14]	Pathology (yes), BI-RADS (varies)	Early CNNs / CAD; single-task classification	Film scans, lower resolution; limited modern FFDM relevance.
Digital (FFDM) datasets	INbreast [20], VinDR-Mammo [23], RSNA [6], CMMD [5], KAU-BCMD [1]	Pathology or BI-RADS; density varies	CNNs/ViTs; single- or dual-task	Often single vendor/system; partial labels; limited multi-task scope.
Harmonization (feature space)	ComBat [22]	Batch-effect correction on features	Empirical Bayes (location/scale) on radiomics/latent features	Requires feature extraction; not pixel-space; modality-agnostic assumptions.
Harmonization (federated learning)	HarmoFL [12]	Federated frequency-domain drift normalization	Joint optimization across clients	Heavy infra; task/model coupling; not a simple preproc.
LUMINA (Ours)	Multi-vendor FFDM (6 systems); energy metadata	Pathology + BI-RADS + density	*Foreground-only* pixel-space CDF matching + unified 3-task benchmark; single-/two-/four-view baselines	Vendor/energy diversity; three tasks; consistent harmonization gains; improved attention focus.

Table 9. **Hyperparameter parity across backbones.** Shared schedule and augmentations ensure a fair comparison (see Sec. 6.1).

Backbone	Pre-trained	Learning Rate	Weight Decay	Epochs	Learning Rate scheduler	Batch	Augmentation
ResNet-50	ImageNet-1K	1×10^{-3}	1×10^{-5}	100	$\times0.1$ every 30	32	flip, resize
DenseNet-121	ImageNet-1K	1×10^{-3}	1×10^{-5}	100	$\times0.1$ every 30	32	flip, resize
EfficientNet-B0	ImageNet-1K	1×10^{-3}	1×10^{-5}	100	$\times0.1$ every 30	32	flip, resize
Swin-T	ImageNet-1K	1×10^{-5}	1×10^{-5}	100	$\times0.1$ every 30	32	flip, resize

Table 10. **Shared pre-processing.** Ensures identical inputs across backbones; isolates model differences.

Stage	Operation (applied to all models equally)
DICOM handling	Convert MONOCHROME1 to MONOCHROME2; remove text burn-ins.
Foreground mask	Define $M = \{(x, y) \mid \mathbf{I}(x, y) > 0\}$; exclude background from histograms.
Harmonization	Foreground CDF matching to low-energy reference (Eqs. 1–8); training-free, model-agnostic.
Resize/replicate	Resize to 224^2 or 512^2; replicate grayscale to 3 channels for ImageNet backbones (Sec. 6.1).
Augment	Random horizontal flip (all backbones, all tasks).
Normalization	As in backbone defaults; no per-model special tuning beyond LR noted in Sec. 6.1.