Mammo-Bench: A Large-Scale Benchmark Dataset of Mammography Images

Gaurav Bhole, Suba S, Nita Parekh
CCNSB, IIIT Hyderabad, India

Introduction

Motivation

- Global Impact: 2.3 million new breast cancer cases are diagnosed worldwide annually
- 670,000 deaths* from breast cancer occur globally each year
- Prognosis is very good for early detection

Why Mammography Matters:

- Most widely used screening tool - over 39 million mammograms performed annually in the US alone
- Most cost-effective screening method - average cost $100-250, compared to $1000+ for MRI
- Accessibility - available in most healthcare facilities, including rural and remote areas

Impact of Improving Mammographic Analysis:

- Earlier detection leads to better survival rates and reduced treatment costs
- Reduced false positives means fewer unnecessary biopsies and patient anxiety
- AI-assisted analysis can support radiologists in making more accurate diagnoses

*WHO 2022 statistics

The Challenge in Mammographic Analysis

Quality of Images

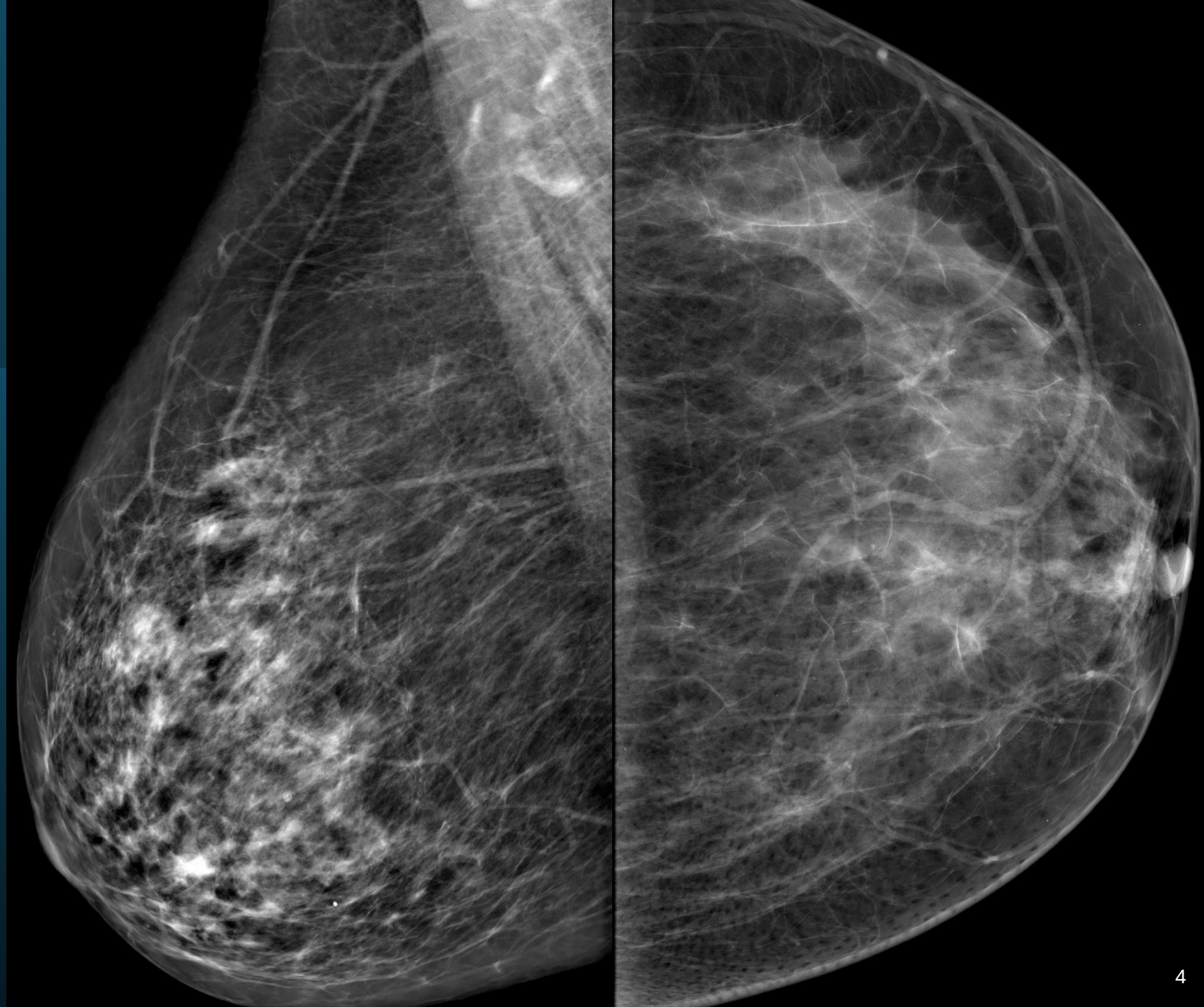
Abnormalities

Breast Density

Architectural distortions and Asymmetries

Human factors affecting diagnosis

Sample Images from Mammo-Bench



Why do we need a Comprehensive Benchmark?

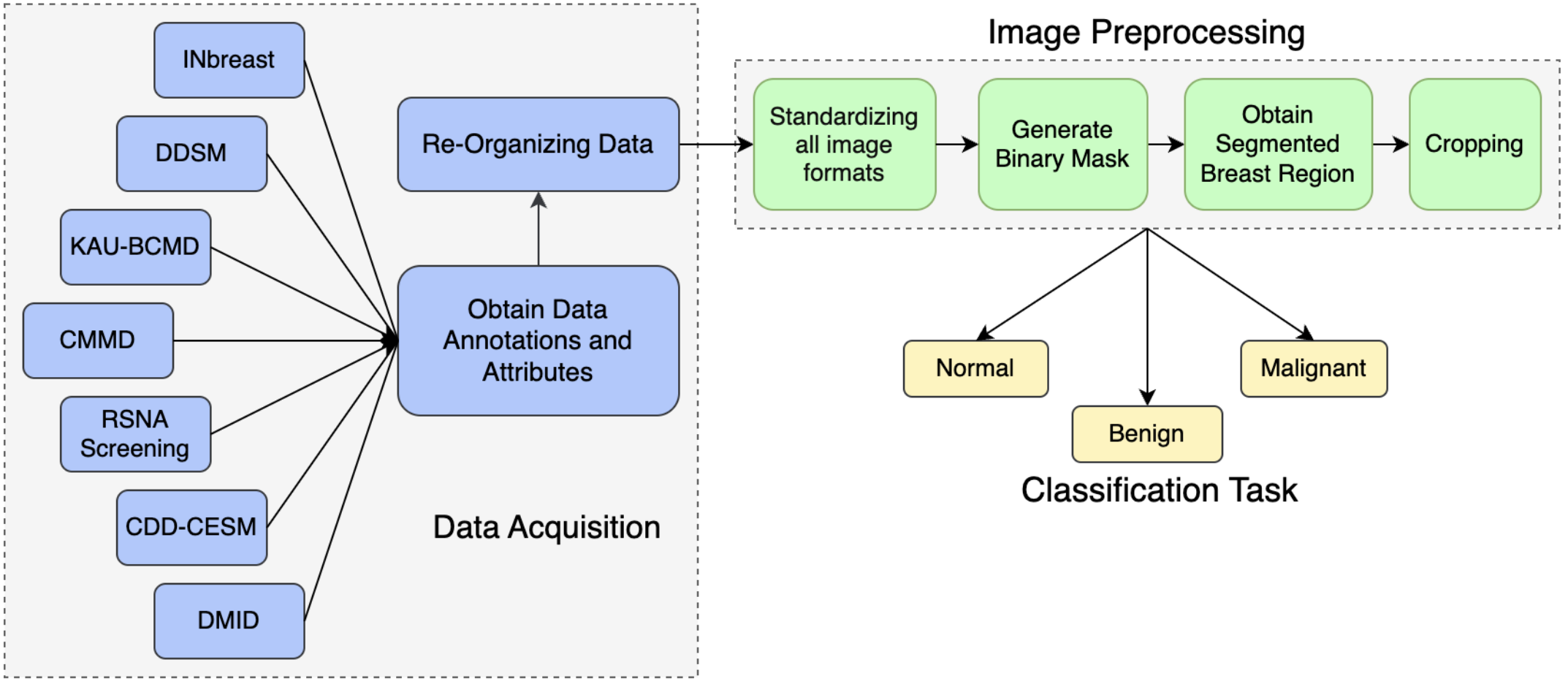
Current Challenges:

- Limited dataset sizes
- Inconsistent image quality
- Incomplete annotations
- Screening dataset biases (e.g. RSNA)

Objectives:

- Development of a large-scale unified benchmark dataset with clinical annotations and standardized images across sources
- Establish baseline performance metrics

Construction of Mammo-Bench



Description of Mammo-Bench

Dataset Features	DDSM	INbreast	KAU-BCMD	CMMD	CDD-CESM	RSNA	DMID	Mammo-Bench
Origin	USA	Portugal	Saudi Arabia	China	Egypt	USA/Australia	India	Diverse
Year	2001	2012	2021	2021	2022	2022	2023	2025
No. of Cases	2,620	115	1,416	1,775	326	20,000	NA	**26,500**
No. of Images	10,400	410	2,206	5,202	1,003	54,705	510	**74,436**
Img. Format	JPG	DICOM	JPG	DICOM	JPG	DICOM	DICOM	JPG
View & Lat.	✔	✔	✔	✔	✔	✔	✔	✔
N/B/M labels	✔	✔	✔	✔	✔	✔	✔	✔
BI-RADS	✘	✔	✔	✘	✔	✔	✔	✔
Breast Density	✔	✔	✔	✘	✔	✔	✔	✔
Abnormality	✘	✘	✘	✔	✘	✘	✘	✔
Mol. Subtype	✘	✘	✘	✔	✘	✘	✘	✔
ROI Mask	✔	✔	✘	✘	✘	✘	✔	✔
Age	✔	✘	✔	✔	✔	✔	✘	✔
Asymmetry	✘	✘	✘	✘	✘	✘	✔	✔

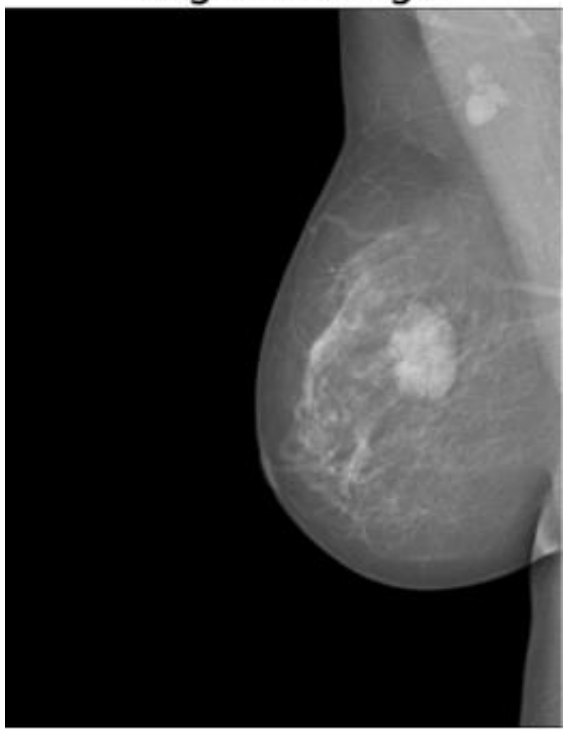
A dataset with diverse clinical annotations

Labels	No. of Images	Class	Images in the Class
Normal/Benign/ Malignant	46,017	Normal (N)	29,264
		Benign (B)	8,334
		Suspicious Malignant (SM)	235
		Malignant (M)	8,184
Density	43,911	ACR A (Fatty)	5,372
		ACR B (Fatty+Scattered Areas of Fibroglandular Density)	18,299
		ACR C (Heterogeneously Dense)	16,418
		ACR D (Extremely Dense)	3,822
BI-RADS Score	30,383	0 (Additional Diagnosis Required)	8,250
		1 (Normal Findings)	18,325
		2 (Benign)	2,670
		3 (Probably Benign)	455
		4 (Suspicious Malignant)	358
		5 (>95% chance Malignant)	313
		6 (Biopsy Proven Malignant)	12
Abnormality	5,712	Mass	3,344
		Calcification	747
		Both	1,411
Molecular Subtype	2,956	Luminal A	600
		Luminal B	1,482
		HER2-enriched	532
		Triple Negative	342

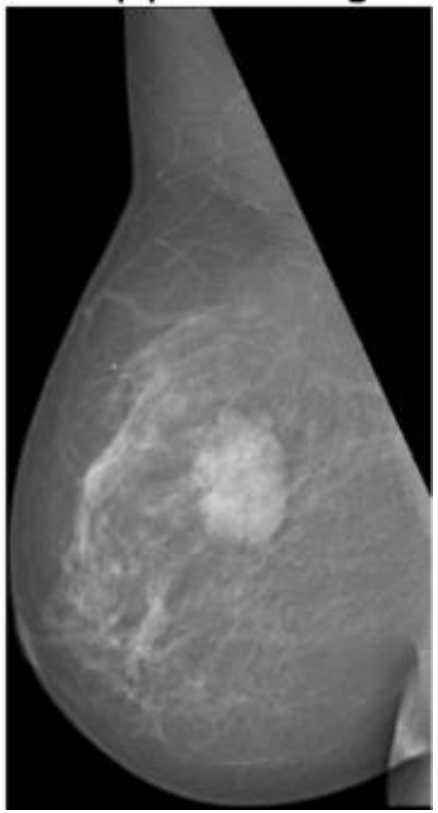
Need for Preprocessing

Example of MLO View

Original Image

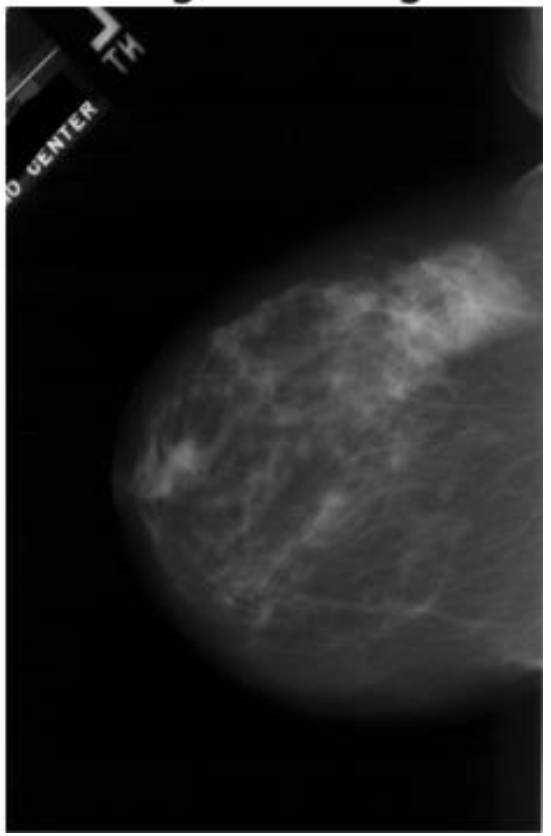


Cropped Image

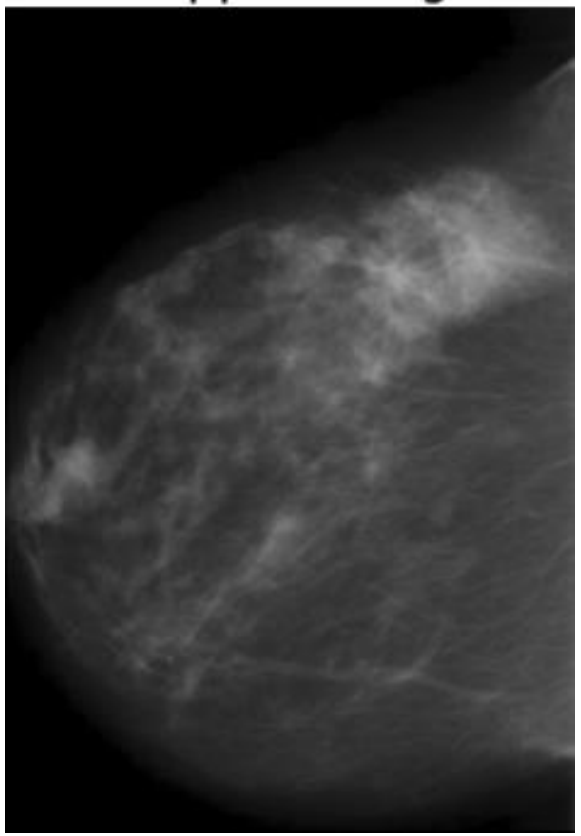


Example of CC View

Original Image

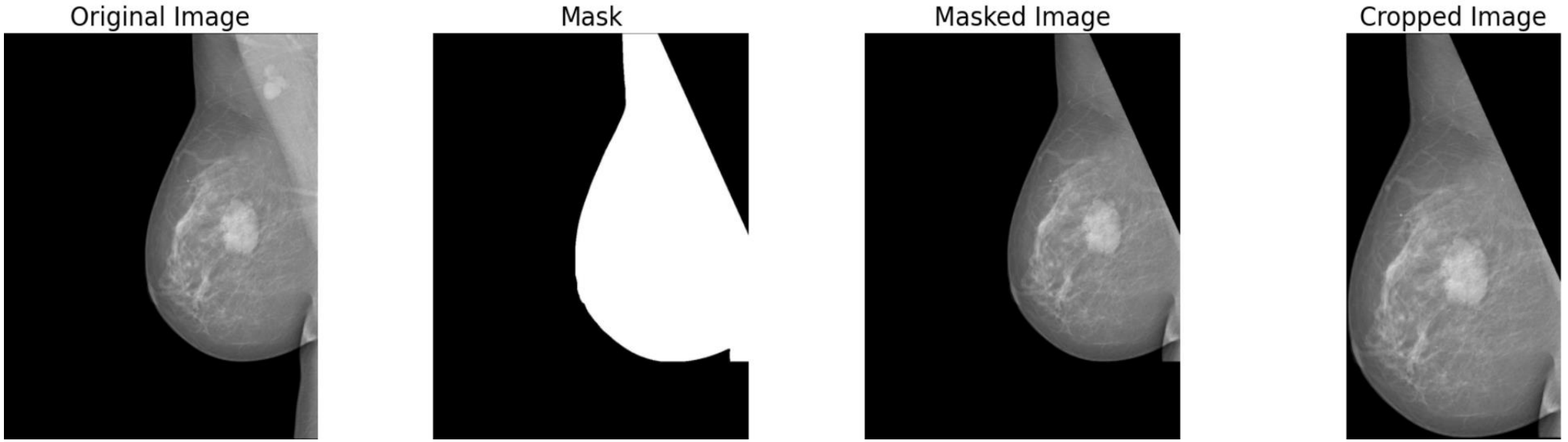


Cropped Image



Need for Preprocessing

Example of MLO view

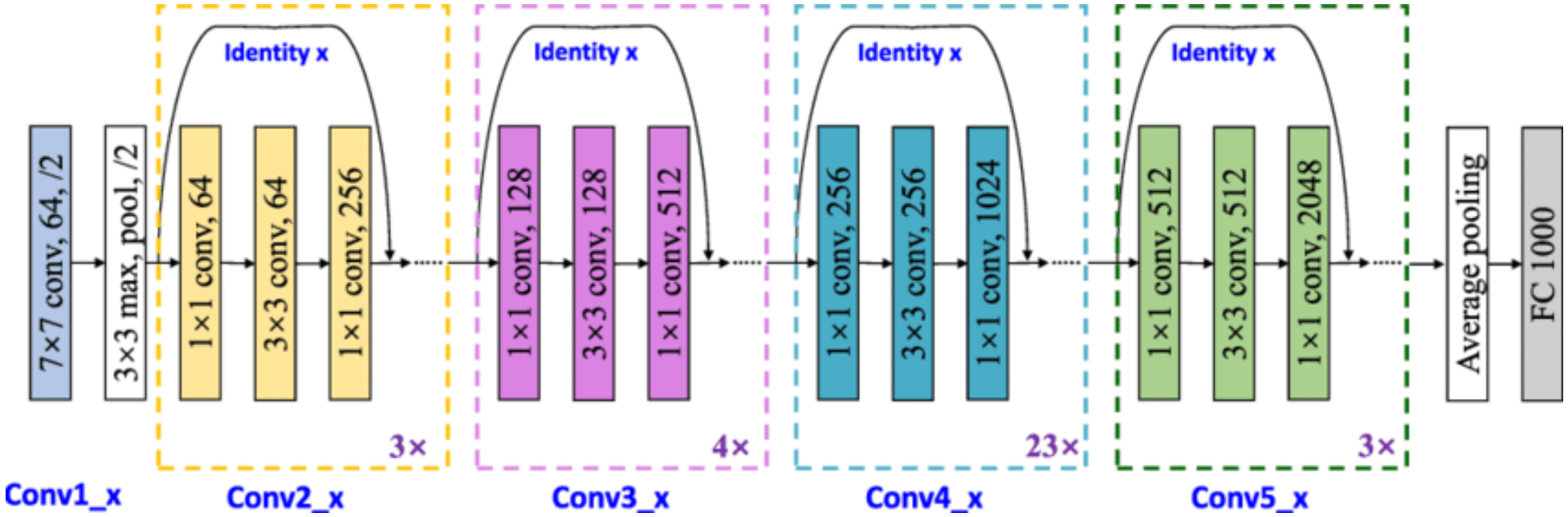


Data

- 3 experiments were performed using Mammo-Bench:
  1. Three-class classification without augmentation
  2. Three-class classification with augmentation on minority classes
  3. Hierarchical binary classification
- Data Split: 80:20 for train-test sets and total images used were 34,721.

Train Set	Class
20,634	Normal
6,925	Benign
7,162	Malignant

ResNet101



ResNet101 Architecture

Performance Evaluation: Three-Class Classification

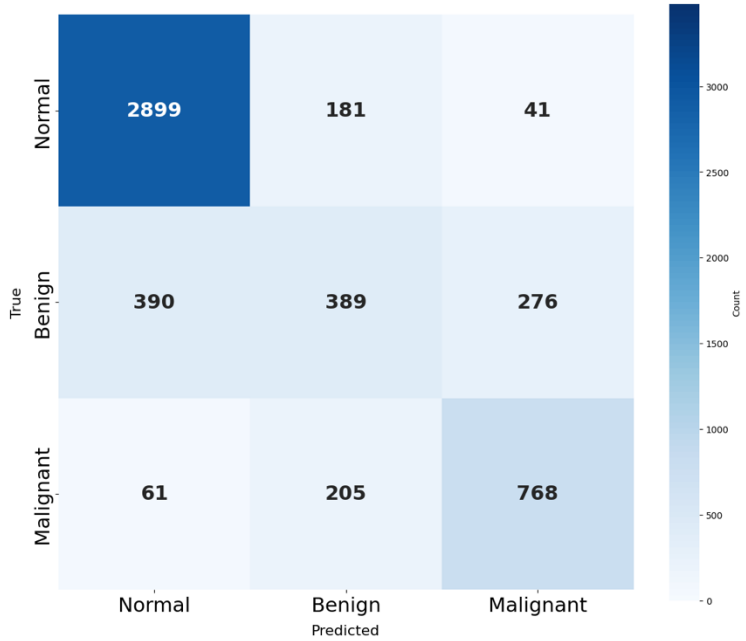
Dataset Used	Class	Precision	Recall	F1-Score	Accuracy
Three-Class Classification	Normal	0.865	0.928	0.895	0.778
	Benign	0.502	0.369	0.425	
	Malignant	0.708	0.743	0.725	
Three-Class Classification*	Normal	0.869	0.929	0.898	0.788
	Benign	0.546	0.382	0.45	
	Malignant	**0.709**	0.777	0.741	
Hierarchical Binary Classification	Normal	**0.876**	**0.954**	**0.913**	0.891
	Abnormal	0.92	0.798	0.854	
	Benign	**0.78**	**0.67**	**0.72**	0.736
	Malignant	0.7	**0.81**	**0.75**	

* with minority class Augmentation

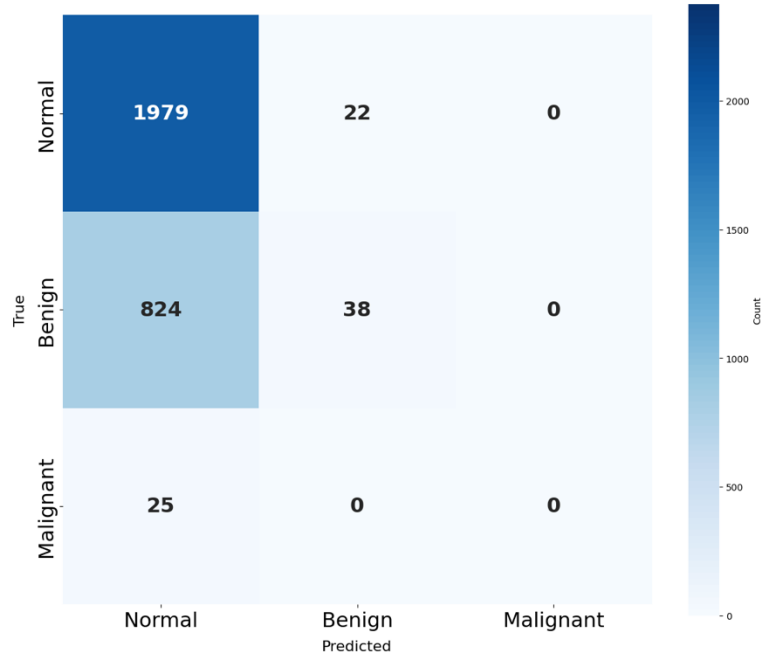
Performance Evaluation: Hierarchical Binary Classification

Dataset	Precision	Recall	F1-Score	Accuracy	MCC*
CDD-CESM	0.504	0.5	0.491	0.5	0.256
VinDr-Mammo	0.677	0.698	0.592	0.698	0.105
DMID	0.55	0.25	0.234	0.25	0.155
Mammo-Bench	**0.760**	**0.778**	**0.766**	**0.778**	**0.595**

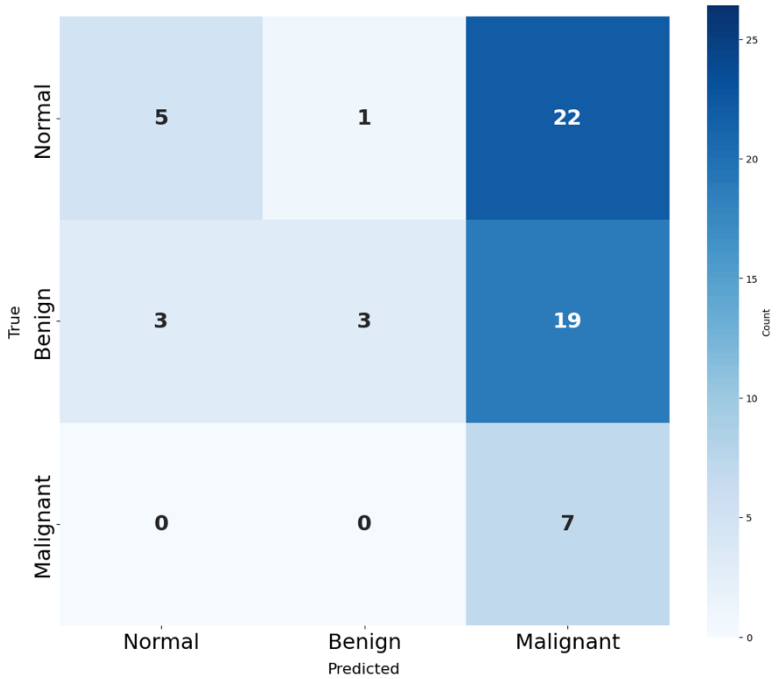
*Mathews Correlation Coefficient



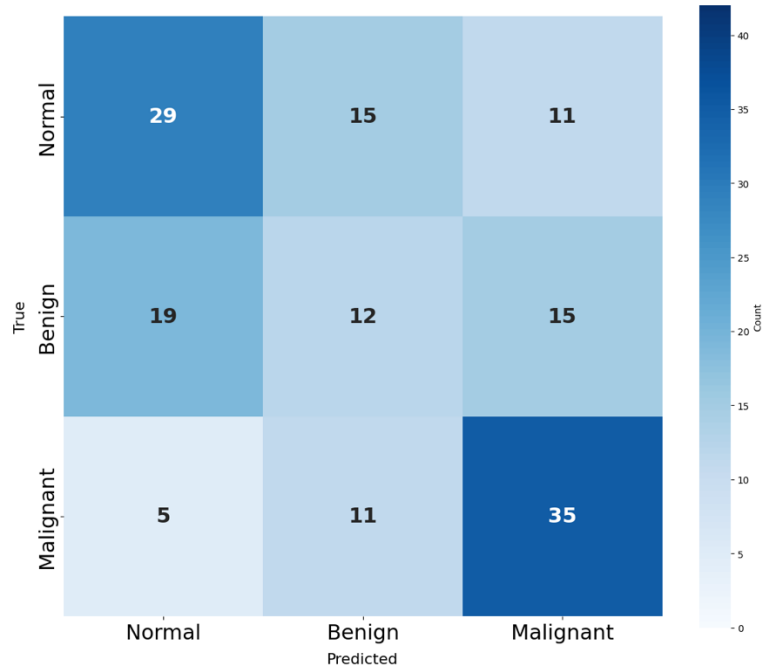
Mammo-Bench



VinDr-Mammo



DMID



CDD-CESM

Conclusion

- Largest open source dataset available with diverse ethnical and geographical distribution
- Improved image quality by preprocessing along with provision of binary mask for segmentation of the breast region
- Resnet101 with Mammo-Bench shows better results than individual datasets and hierarchical binary classification showed best results for all classes with major improvement in benign class
- Limitations: Some data imbalance still persists and all annotations not available for each image
- Future work: Integration of multi-modal data like clinical attributes along with images as inputs

Thank You!!!

Current Landscape of Mammography Datasets

Dataset Categories:

Open Source Datasets:

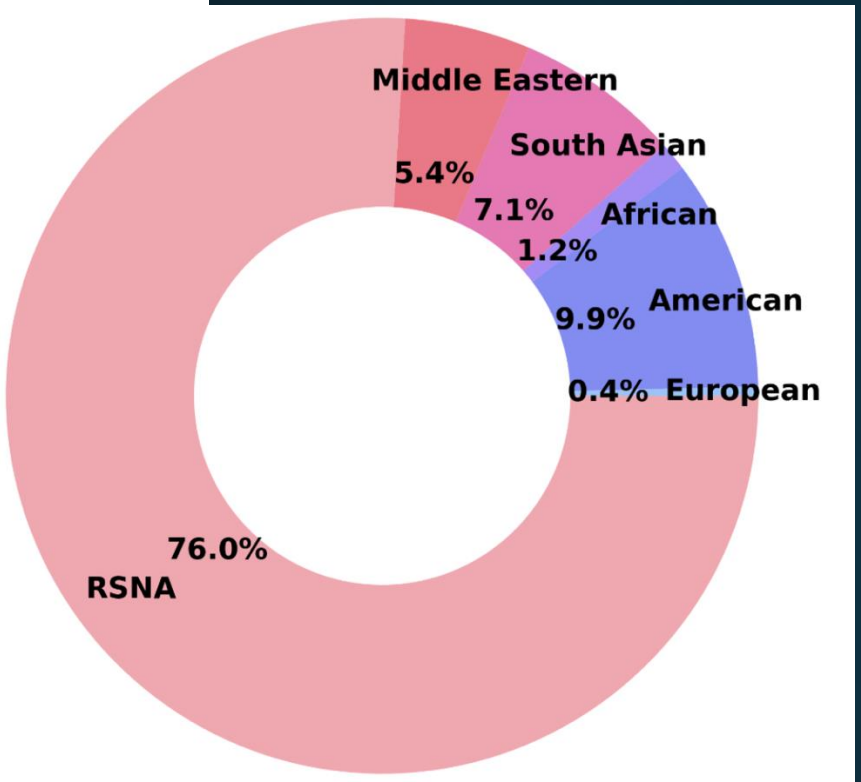
- e.g. DDSM, INbreast, KAU-BCMD, etc.

Restricted Access Datasets:

- e.g. OPTIMAM, VinDr-Mammo, etc.

Geographic & Ethnical Distribution

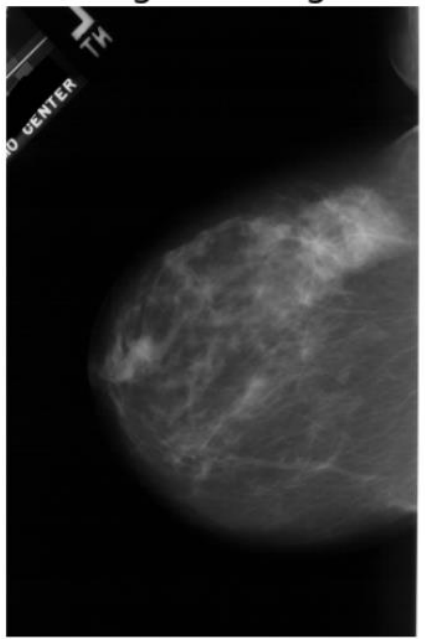
- Incorporates data from 7 countries across multiple continents
- Most diverse representation among existing mammography datasets, with RSNA providing broad US/Australian coverage and regional datasets adding unique populations
- Enables development of more inclusive and generalizable AI models by capturing diverse features across different ethnicities



Need for Preprocessing

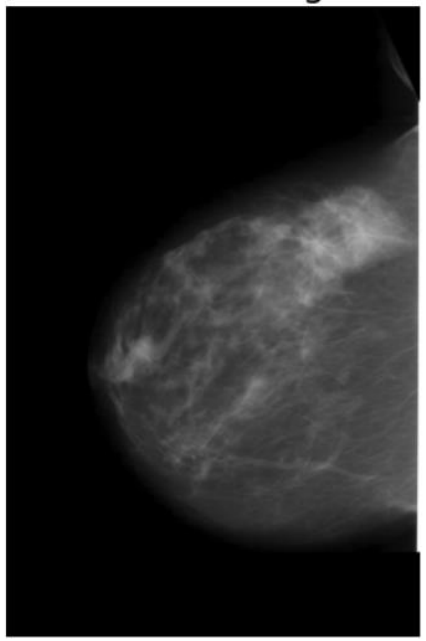
Example of CC View

Original Image



Mask

Masked Image



Cropped Image