Deep Learning for Breast Cancer Diagnosis on Mammography

Ruijing Li

01

Previous Work

The Purpose of Research

Develop a deep learning model to assist in mammogram-based breast cancer diagnosis :

- Reduce this burden of radiologists and improve overall efficiency.
- Provide consistent and objective second opinions.
- Identify subtle biomarkers that may be imperceptible to the human eye.

Methodology

Dataset

RSNA Breast Cancer Detection dataset

- Construct balanced small and large subsets of RSNA for training and validation.

INbreast dataset

- External Validation

Model Selection

Resnet 101

- Residual learning
- Deeper layers capture high-level abstract features

Data Preprocessing

Process DICOM images to PNGs (256×256)

Normalize training data into ImageNet standards

Training Strategy

Hyperparameters

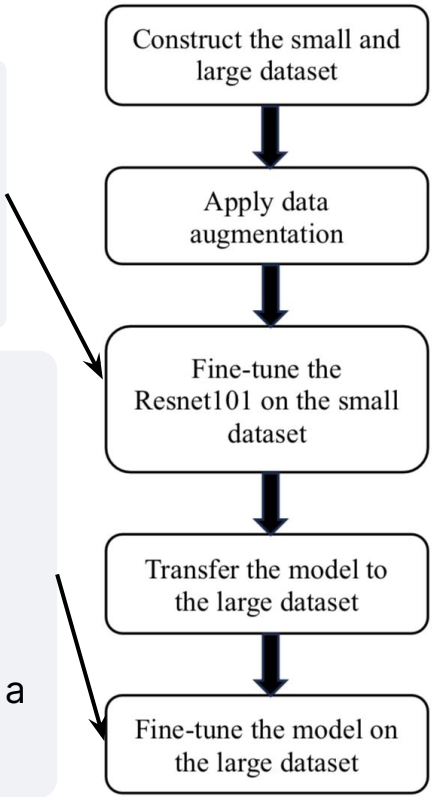
Initialization:
Pretrained on ImageNet
Data augmentation:
Random flipping, rotation
Batch size:
128
Learning rate:
Dynamically adjusting

Phased Training Strategy

- Warm-up the model on small dataset.
- Transfer to large dataset for fine-tuning.

Progressive Unfreezing Strategy

- Freeze convolutional layers and train only the fully connected layers.
- Gradually unfreeze and fine-tune the model with a smaller learning rate.



Results

The Small Dataset (421 images):

	Accuracy	F1 Score	AUC
Initial Training	73.8%	0.23	0.71
Fine-tuned Re-test	85.7%	0.66	0.823

The Large Dataset (2300 images):

	Accuracy	F1 Score	AUC
Initial Transfer	48.8%	0.20	0.495
Fine-tuned	93.8%	0.80	0.86

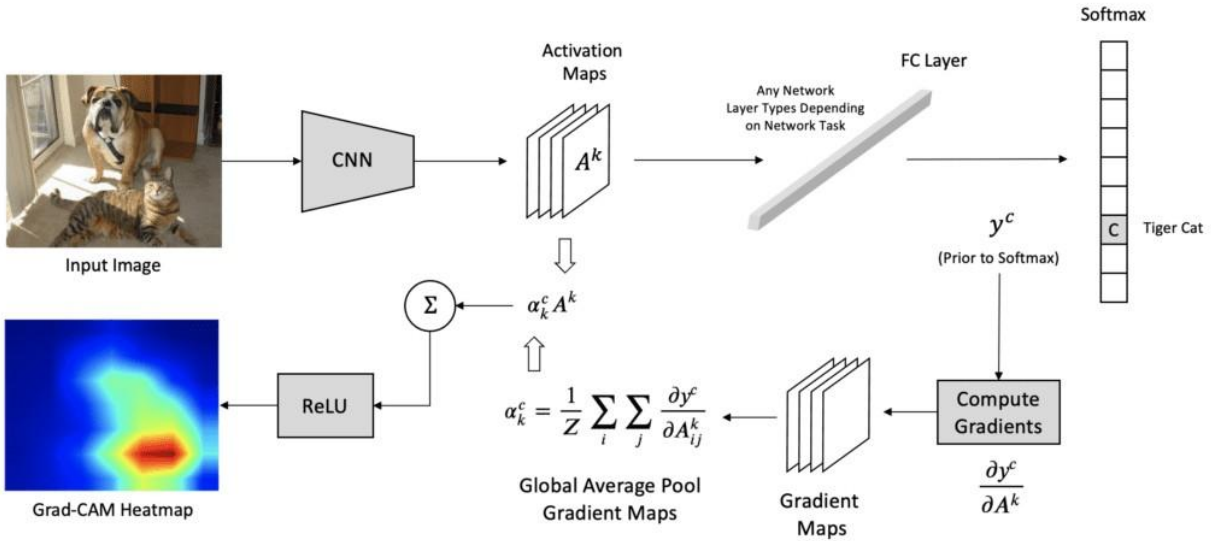
INBreast Dataset:

Accuracy: 82.2%

Interpretability Analysis of Best Model

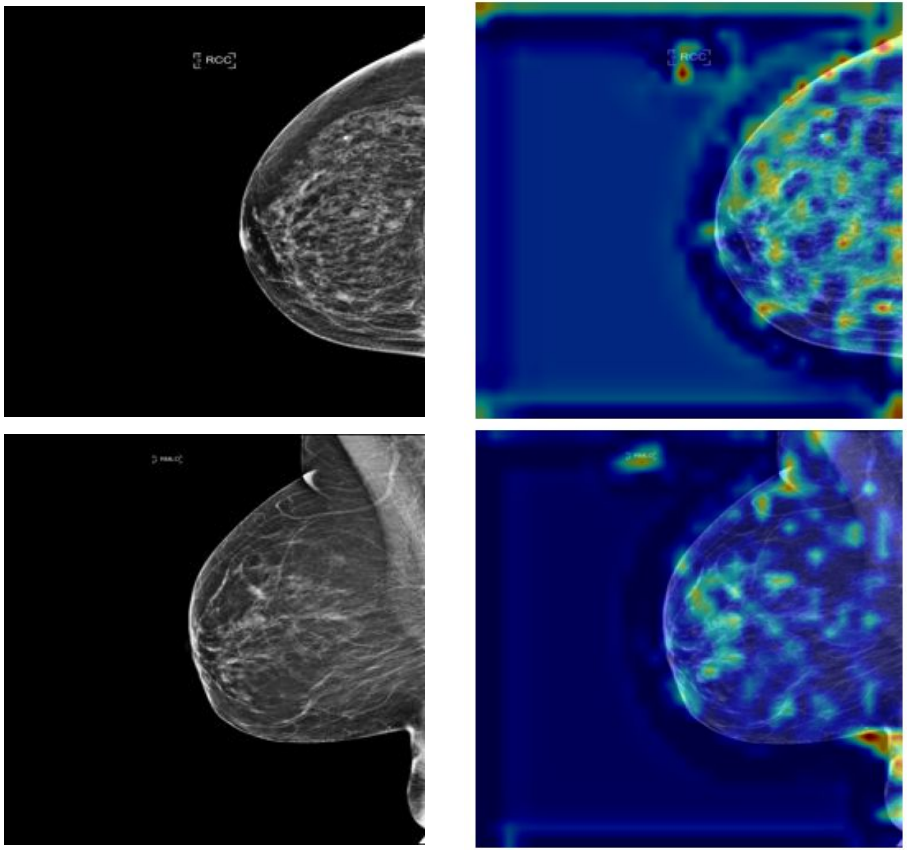
Grad-CAM
(Gradient-weighted Class Activation Mapping)

Grad-CAM is a a class-discriminative localization technique that visualizes the important regions in an image that influence a CNN's decision.



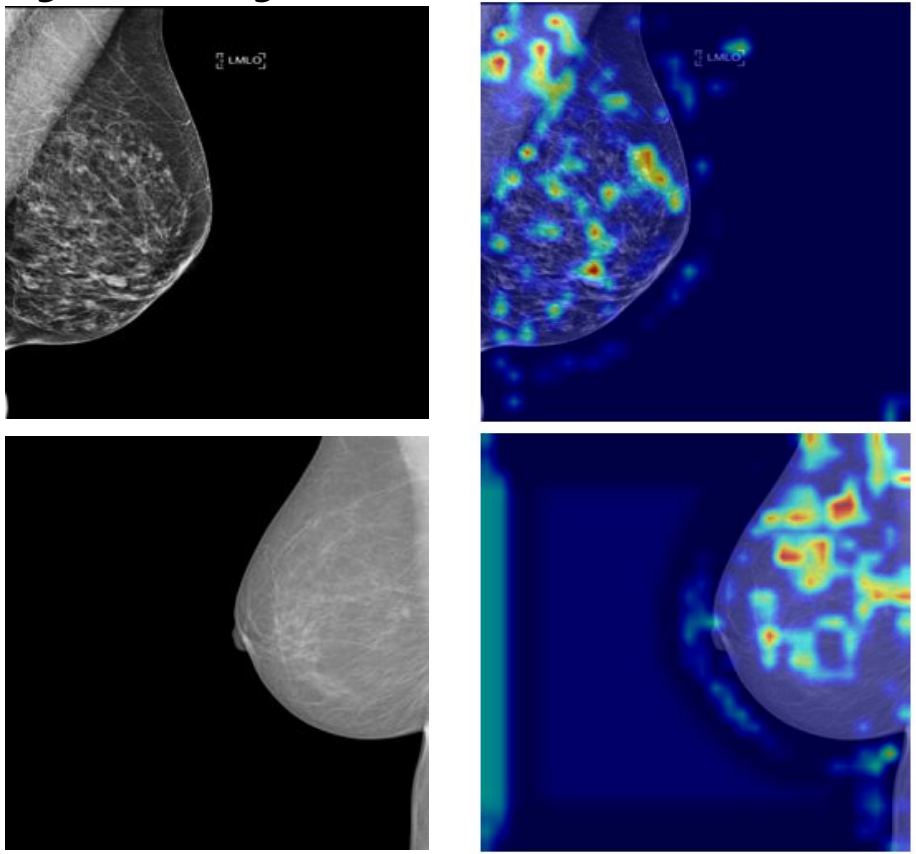
Interpretability Analysis of Best Model

The Benign Cases



Interpretability Analysis of Best Model

The Malignant Cases



Interpretability Analysis of Best Model

Observed Pattern

🔴 **Malignant Predictions**:
Grad-CAM highlighted intense red regions, often focused on masses, nodules, or irregular structures within the breast tissue.
This suggests the model may have learned to focus on features (high density area) associated with malignancy.

🟡 **Benign Predictions**:
The highlighted areas were more yellowish, with weaker intensity and less spatial focus.
The model showed less confidence localization in these cases, maybe reflect the location of suspicious lesions.

Based on the visualization results, it can be inferred that the model may have learned to distinguish between benign and malignant cases using structural features, rather than relying solely on global texture patterns.

02

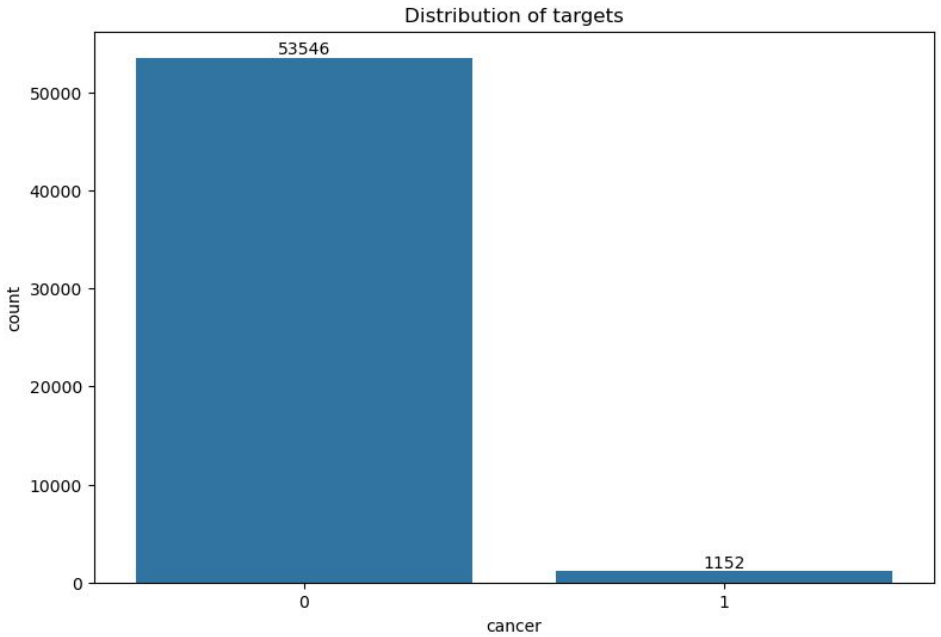
Recent Work

Aim

- Validate our training strategies in a more realistic setting
  - Use the full RSNA dataset with real cancer prevalence
  - Enforce patient-level independence between train and validation
  - Quantitatively evaluate:
    - small-subset warm-up
    - progressive unfreezing

Realistic Data Setting:

Dataset: full RSNA training set
Keep the original cancer prevalence (~2%)



loss weighting :
pos_weight

In binary classification with BCEWithLogitsLoss, pos_weight is a parameter that increases the loss contributed by positive samples.

$$\text{pos\_weight} \approx \frac{\text{number of negative samples}}{\text{number of positive samples}}$$

Avoid Data Leakage

- **Group by (patient_id, laterality)**

  All images from the same breast of one patient = one group

- **Train / validation split at the group level**

  No patient appears in both train and validation

Evaluation Under Extreme Imbalance

With ~2% positives: "All-negative" model → ≈98% accuracy

ROC-AUC : higher AUC means better global discrimination between cancer or not
trade-off between sensitivity & specificity
threshold-free summary of discrimination

PR-AUC : higher PR-AUC means the model can detect more true positive case while keep low false-positive rate
precision vs. recall for the positive class
more informative when positives are rare

Current Baseline Benchmark

Setting	ROC-AUC	PR-AUC
Single-stage training (full RSNA)	0.6720	0.0614

- ROC-AUC of 0.669 indicates that the model has a clear discriminative ability between cancer and non-cancer
- PR-AUC = 0.0529 is clearly above the prevalence (~0.02), meaning the model selects suspicious cases better than random

a reasonable but improvable baseline

Warm-up Subset

Design :

- Construct warm-up subset at group level
- Include all positive groups in training set
- Randomly sample the same number of negative groups

Result: 1:1 positive:negative subset (by groups)

Setting	ROC-AUC	PR-AUC
Random (theoretical)	0.5000	0.0210
Warm-up on 1:1 subset	0.6082	0.0325

03

Future Work

Future Work

Effect of Small-subset Warm-up

A: single-stage training on full RSNA (baseline)
B: warm-up on 1:1 subset → full RSNA training

Same validation split, compare:
ROC-AUC, PR-AUC
convergence speed (epochs vs. AUC)

Effect of Progressive Unfreezing

- Fine-tune all layers at once vs. progressively unfreeze
- Test with and without warm-up initialization

Thanks