QUBAHAN JOURNAL OF MEDICAL SCIENCES
VOL. 3, November 2026
https://doi.org/10.48161/qjms.v3a92

Lesion-Location Supervision for Dual-Resolution Mammography Classification: A Patient-Grouped Evaluation on Mini-MIAS

Muhammad Abuarrah[1*], Yogeesh N[2], Afshin Soflaei Shahrbabak[3]

[1] Radiography and Medical Imaging Department, Fatima College of Health Sciences, Abu Dhabi, United Arab Emirates;
[2] Department of Mathematics and Natural Sciences, Gulf University for Science and Technology, Hawalli, Kuwait;
[3] Tehran Medical Sciences Branch, Islamic Azad University, Tehran, Iran.

* **Corresponding author:** Muhammad.Abuarah@actvet.gov.ae.

ABSTRACT: This study evaluated whether lesion-location supervision improves benign-versus-malignant classification and lesion-center coverage in full mammograms. The analysis included 115 Mini-MIAS mammograms from 108 patients, comprising 63 benign and 52 malignant images. Two dual-resolution models combined a downsampled whole-mammogram view with 25 overlapping high-resolution tiles using a shared frozen ResNet-18 encoder. One model used lesion centers to guide tile weighting during training. Both were predicted from pixels alone on the same held-out patient folds. In the reference run, pooled area under the receiver operating characteristic curve (ROC-AUC) was 0.493 with location supervision and 0.491 with image labels alone (paired difference 0.002, 95% patient-bootstrap confidence interval −0.006 to 0.011). Across five training runs on fixed partitions, supervised ROC-AUC ranged from 0.480 to 0.511. Among 111 mammograms with usable coordinates, top-tile lesion-center coverage increased from 6.3% to 23.4% (gain 17.1 percentage points, 95% confidence interval 9.1 to 25.5). Uniform random and fixed central tile selection yielded 6.5% and 39.6% coverage, respectively. On the coordinate-eligible subset, reference ROC-AUC was 0.488 for the annotation-centered crop classifier and 0.508 for the supervised full-image model. Location supervision improved tile coverage, while full-image malignancy discrimination stayed near chance under the evaluated protocol.

Keywords: mammography, lesion localization, spatial supervision, patient-grouped validation, Mini-MIAS.

I. INTRODUCTION

Mammograms contain information at different spatial scales. The full image preserves the distribution of breast tissue and the position of a suspicious finding within it, while smaller regions retain details such as lesion margins, local texture, and calcification patterns. Resizing an entire mammogram to fit a neural network can make these details less visible. Processing smaller regions at higher resolution helps preserve local detail, but each region captures only part of the surrounding tissue. Global–local models address this trade-off by combining features from the full mammogram with features from local patches. Shen et al., for example, used a global saliency map to select patches for breast cancer classification [1]. Their approach provides a basis for examining whether explicit lesion-location annotations can help a model assign greater weight to relevant image regions during training.

Training with lesion locations may affect two distinct aspects of model performance. It may help the model focus on an abnormality, and it may improve the model's ability to distinguish benign from malignant findings. These outcomes need to be measured separately because a tile can cover an annotated lesion while the features extracted from it provide limited information about malignancy. Conversely, a correct classification provides little information about whether the model assigned substantial weight to the annotated region. Evaluating classification alongside lesion coverage allows assessment of whether a change in tile weighting is accompanied by a change in discrimination.

The Mini-MIAS dataset supports this evaluation because it contains digitized film mammograms with annotations describing the type and severity of abnormal findings and, where available, their locations. The present study focuses on mammograms with annotated abnormalities and evaluates discrimination between the dataset's benign and malignant labels. The small sample and uncertainty about how the severity labels were established limit the interpretation of the results [2]. The dataset also contains paired mammograms from the same patient, which makes patient-level grouping important when constructing training, validation, and test sets. Kag et al. showed that changing the order of splitting and augmentation substantially altered reported accuracy in a whole-image Mini-MIAS experiment [3]. In the present study, images from the same patient remain in the same group, and augmented views are used only during training.

This study proposes a dual-resolution model that combines a downsampled whole-mammogram view with overlapping high-resolution tiles. The whole-image branch estimates the relative importance of the tiles, and the classifier combines the weighted tile features with the whole-image representation. During training, available lesion-center annotations guide the tile weights through an additional location loss. During evaluation, the model estimates tile weights from the mammogram alone. A matched comparison uses the same architecture, image inputs, and patient-grouped folds with the location loss disabled. This comparison isolates the contribution of location supervision under the evaluated training protocol. The study makes three contributions. First, it implements a dual-resolution model that uses lesion annotations during training while requiring only the mammogram for prediction. Second, it evaluates the contribution of lesion-location supervision through a matched comparison on identical patient-grouped folds. Third, it reports benign-versus-malignant discrimination alongside the proportion of mammograms in which the highest-weighted tile covers an annotated lesion center. This joint evaluation shows whether improvements in lesion coverage are accompanied by improvements in classification on held-out patients.

II. RELATED WORK

Shen et al. developed a global–local model that combines information from the full mammogram with features extracted from selected high-resolution patches. A global saliency map guides patch selection, allowing the classifier to retain breast-level context while examining smaller regions in greater detail. This approach laid the groundwork for studying how regional weighting influences whole-mammogram classification [1]. In a later study, Shen et al. used a whole-image network to identify regions for higher-resolution processing and evaluated localization learned from image-level labels [4]. Ribli et al. evaluated a lesion detector that also classified findings on mammograms [5]. Lotter et al. studied annotation-efficient cancer detection in mammography and digital breast tomosynthesis [6]. These studies illustrate different uses of spatial supervision and regional evidence.

Chen et al. proposed BRAIxDet to detect malignant lesions in screening mammograms with incomplete spatial annotations. They first pretrained a multi-view classifier using image-level labels and then adapted it for lesion detection using the available spatial annotations and semi-supervised learning. Their work shows how mammograms with different annotation levels can contribute to a common training pipeline. The primary endpoint was malignant lesion detection [7]. Zhu et al. proposed local extremum mapping for mammogram classification and weakly supervised lesion localization. Their method generates regional score maps, identifies informative regions from local extrema, and aggregates regional scores to classify the mammogram. Experiments on the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) [8] and INbreast [9] showed that regional localization can be learned from image-level labels. This approach is relevant to the matched comparison in the present study, in which tile weights are learned through the classification objective [10]. Mahmood et al. developed a convolutional neural network (CNN) to classify suspicious breast regions using Mini-MIAS and a separate private dataset. They reported a test accuracy of 93% and an area under the receiver operating characteristic curve (ROC-AUC) of 0.95 on Mini-MIAS. The corresponding values on the private dataset were 97% and 0.99. Because the pipeline includes suspicious-region extraction, interpretation of its classification performance depends in part on how those regions are identified [11]. Avcı and Karakaya examined how segmentation affects downstream mammogram

classification on Mini-MIAS. They evaluated region-based, edge-based, and deep learning segmentation methods. Region-based approaches, including k-means clustering and Otsu thresholding, performed better in their normal-versus-abnormal classification experiments. Their findings illustrate how defining an image region can affect the features available to a classifier [12].

Teoh et al. combined ResNet-50 and GoogLeNet to identify microcalcifications in mammograms drawn from Mini-MIAS, INbreast, and CBIS-DDSM. Their ensemble achieved 99.07% accuracy in the combined experiment. This result concerns microcalcification identification across three datasets. The endpoint and data composition differ from those of a benign-versus-malignant classification experiment restricted to Mini-MIAS [13]. Zebari et al. reviewed computer-aided approaches to breast cancer detection from mammograms, covering image preprocessing, segmentation, feature extraction, and classification. Their review provides a basis for comparing pipelines according to how they obtain the regions used for classification. For the present study, the relevant distinction is whether lesion locations are available when the model processes a held-out mammogram [14]. Zebari et al. evaluated an improved multifractal feature method for benign-versus-malignant classification across several mammography datasets. In one cross-dataset experiment, the method was trained on the Breast Cancer Digital Repository and achieved 100% reported accuracy on the prepared Mini-MIAS test samples. The training source and sample preparation define the scope of that result and limit its direct comparability with patient-grouped evaluation within Mini-MIAS [15]. Zebari et al. applied transfer learning with a pretrained CNN to breast cancer detection from mammograms. A secondary source reports 95.71% accuracy on Mini-MIAS for this approach. The work is relevant to the present study's use of pretrained convolutional features. The accessible sources leave the patient-grouping procedure unclear, so the reported accuracy is best interpreted in the context of the original experiment [16]. Zebari et al. used deep features to segment suspicious regions in mammograms. Their work addresses spatial delineation of suspicious tissue and is relevant to using spatial information in mammography models. The present study evaluates a different localization outcome: whether the highest-weighted tile contains an annotated lesion center [17]. Kag et al. examined the effect of data splitting and augmentation on reported Mini-MIAS performance. In their whole-image experiment, accuracy decreased from 97.6% to 51.3% after they corrected the order of splitting and augmentation. In a separate experiment, a model trained on CBIS-DDSM achieved a ROC-AUC of 0.734 when evaluated on Mini-MIAS lesion patches. The two experiments illustrate why the split procedure and the availability of lesion crops at evaluation should be documented when comparing published results [3].

Qian et al. combined an evolving convolutional network with a hybrid extreme learning machine and chimp optimization for breast cancer diagnosis [18]. Cao et al. studied an AlexNet–extreme learning machine pipeline with chimp optimization and Nelder–Mead refinement [19]. These studies motivate interest in how feature representations and classification heads affect breast cancer models. The present study instead holds the feature extractor and classification architecture fixed to estimate the contribution of lesion-center supervision.

In other medical imaging tasks, Zebari et al. evaluated DenseNet121 with augmentation for brain tumor classification on magnetic resonance images [20]. Zebari et al. studied a support vector machine for COVID-19 classification on chest radiographs [21], while Salih et al. combined local binary pattern features with convolutional features for that task [22]. These papers illustrate alternative approaches to augmentation and feature construction in medical imaging. Their outcomes concern different modalities and prediction tasks from the Mini-MIAS experiment.

Prior studies show that mammography models can learn to highlight suspicious regions from image-level labels and can use spatial annotations to support lesion detection. The studies reviewed here differ in their prediction targets, input regions, and evaluation protocols. They therefore leave open a specific question about the added value of lesion-center supervision for benign-versus-malignant classification from full Mini-MIAS mammograms. The present study evaluates two variants of the same dual-resolution model on identical patient-grouped folds. Both variants use a downsampled whole-mammogram view and overlapping high-resolution tiles. One variant receives an additional training loss based on annotated lesion centers. At evaluation, both variants estimate tile weights from the mammogram alone. Reporting top-tile lesion coverage alongside ROC-AUC allows the study to assess whether a change in localization is accompanied by a change in benign-versus-malignant discrimination.

III. MATERIALS AND METHODS

This study evaluated whether lesion-location supervision during training improves benign-versus-malignant classification and lesion localization on full mammograms. Two variants of the same dual-resolution model were trained and evaluated on identical patient-grouped folds. The proposed variant used lesion-center annotations to supervise tile weighting during training. Both variants predicted from mammogram pixels on held-out patients.

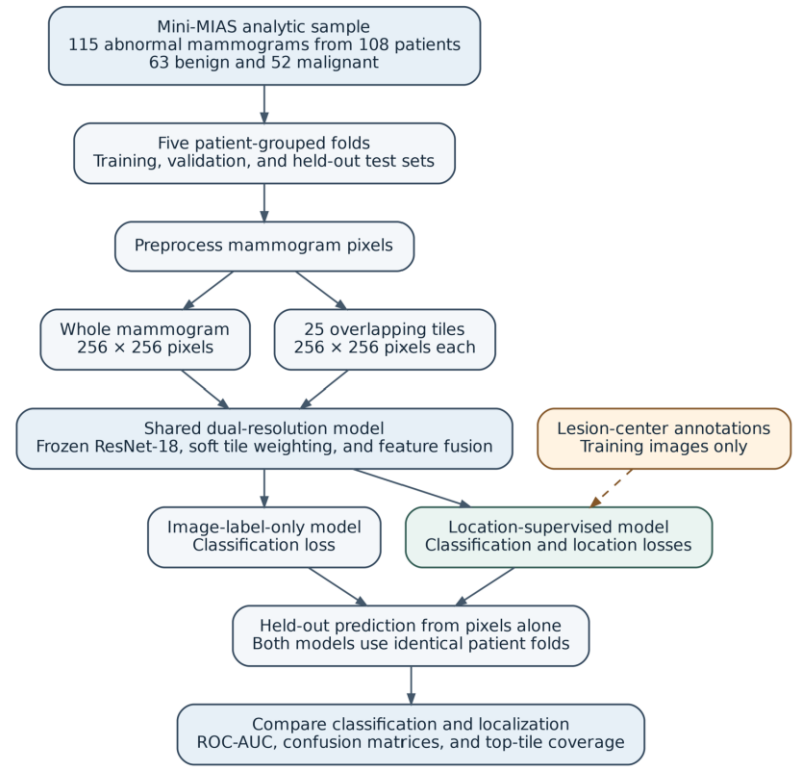


Figure 1. Methodology workflow

3.1. Mini-MIAS Dataset and Sample Selection

This study used the reduced-resolution Mammographic Image Analysis Society dataset (Mini-MIAS), distributed through Kaggle and originally documented by Suckling et al. [23]. The dataset comprised 322 grayscale mammograms from 161 patients, each with a resolution of $1{,}024 \times 1{,}024$ pixels. The dataset documentation identifies each consecutive odd–even image pair as the right and left mammograms of the same patient. A common patient identifier was assigned to each pair so that both images remained in the same partition whenever they were eligible for analysis. The annotation records specified the type and severity of each abnormality and, where available, its center coordinates.

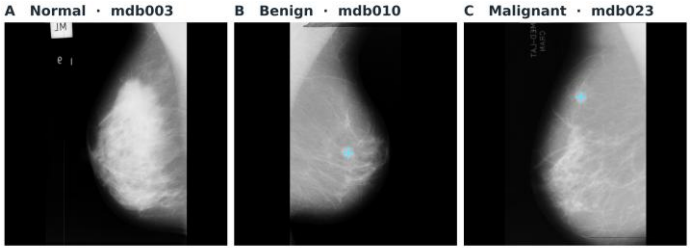


Figure 2. Mini_MIAS Mammogram Samples

The individual mammogram was the unit of prediction, and the patient was the unit of data splitting. A mammogram was labeled malignant if it contained at least one finding annotated as malignant. It was labeled benign if all annotated abnormalities were benign. Under this rule, one mammogram with both benign and malignant findings was assigned to the malignant class.

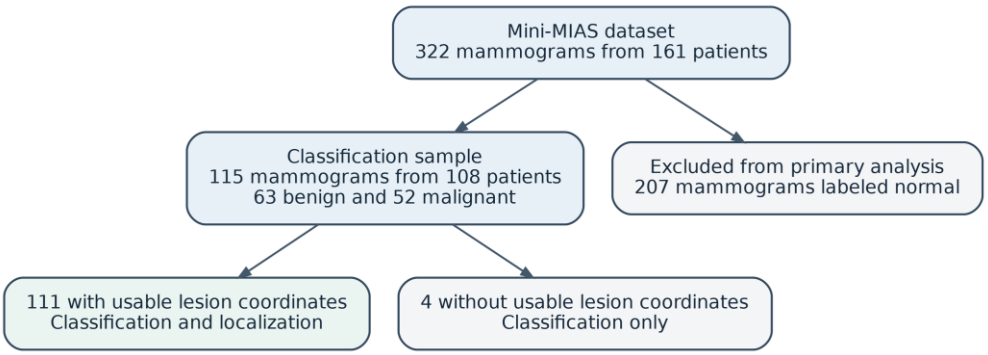


Figure 3. Mini-MIAS sample selection

The 207 mammograms labeled normal were excluded because the primary task was to distinguish benign from malignant findings among mammograms with annotated abnormalities. The resulting classification sample comprised 115 mammograms from 108 patients, including 63 benign and 52 malignant images. All 115 mammograms contributed to the classification evaluation. The localization analysis used the 111 mammograms with usable lesion-center coordinates. The remaining four contributed only to the classification analysis. Figure 3 summarizes the sample selection and the eligibility criteria for the two evaluations.

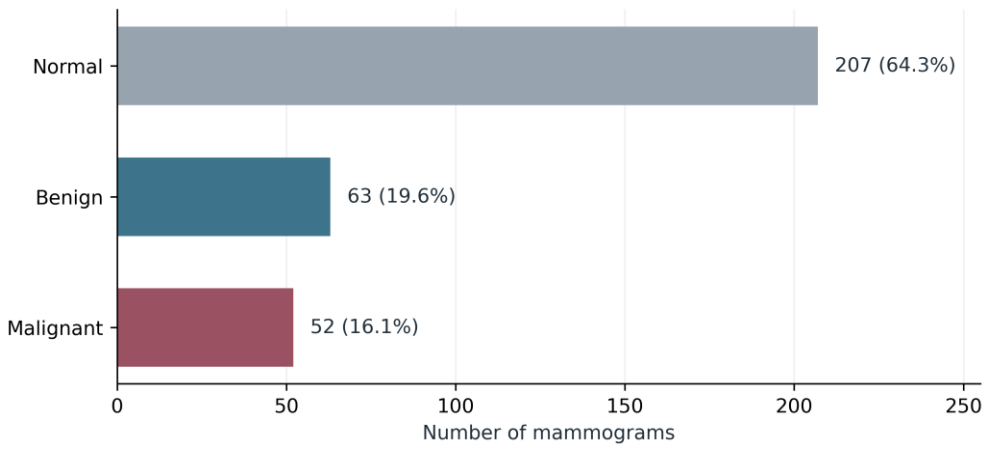


Figure 4. Mini_MIAS Class Distribution

3.2. Image Preprocessing and Tile Grid

Grayscale pixels were scaled to [0, 1], and the outer 16 pixels on each image edge were set to zero. The global branch received 256 × 256 images bilinearly downsampled. The local branch received 25 tiles of 256 × 256 pixels extracted at a stride of 192 pixels on both axes, yielding 64 pixels of overlap. Tile origins were 0, 192, 384, 576, and 768 pixels on each axis. Each grayscale view was repeated across three channels and normalized with the ImageNet channel means and standard deviations. Lesion coordinates were converted from the dataset's bottom-origin convention to top-origin image coordinates and mirrored for horizontally flipped training views.

3.3. Proposed Method

The proposed dual-resolution model combines the anatomical context of the whole mammogram with fine-scale information from high-resolution tiles. A shared ImageNet-pretrained ResNet-18 encoder [24], truncated after the third residual stage, processes both inputs. The encoder weights remain fixed throughout training. The whole-mammogram branch retains a spatial feature map, while spatial average pooling produces one feature vector for each tile.

A trainable 1×1 convolution projects the whole-mammogram feature map onto a single-channel location map. The model samples this map at the centers of the 25 tiles and applies a softmax to obtain the tile weights $a_1, \ldots, a_{25}$. It computes a weighted sum of the tile feature vectors and concatenates it with the spatially averaged whole-mammogram representation. A linear classifier maps the combined representation to a malignancy score, which is then transformed by a sigmoid into a value between zero and one. During training, available lesion-center annotations provide a soft target distribution over the tiles. For tile j with center $\mathbf{t}_j$, the target weight q_j is determined by its distance to the nearest lesion center in the set of annotated centers $\mathcal{C}$:

$$q_j = \frac{\exp\left[-\frac{\min_{c\in\mathcal{C}} \| \mathrm{t}_j - \mathrm{c} \|_2^2}{2\sigma^2}\right]}{\sum_{k=1}^{25} \exp\left[-\frac{\min_{c\in\mathcal{C}} \| \mathrm{t}_k - \mathrm{c} \|_2^2}{2\sigma^2}\right]}, \qquad \sigma = 96 \text{ pixels.}$$

This formulation assigns greater target weight to tiles near an annotated lesion center and distributes supervision across neighboring tiles. The training objective combines class-weighted binary cross-entropy for malignancy classification with a location loss that penalizes disagreement between the target and predicted tile weights:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda \mathcal{L}_{\mathrm{loc}}, \qquad \mathcal{L}_{\mathrm{loc}} = -\sum_{j=1}^{25} q_j \log(a_j), \qquad \lambda = 0.5.$$

The positive-class weight in $\mathcal{L}_{\mathrm{cls}}$ was calculated within each fold as the ratio of benign to malignant mammograms in the training subset. The location loss was averaged over training images with usable lesion-center coordinates. Images without usable coordinates still contributed to the classification loss. During validation and testing, the model

derived both tile weights and malignancy scores from mammogram pixels alone. Figure 5 shows the model architecture and the role of lesion annotations during training.

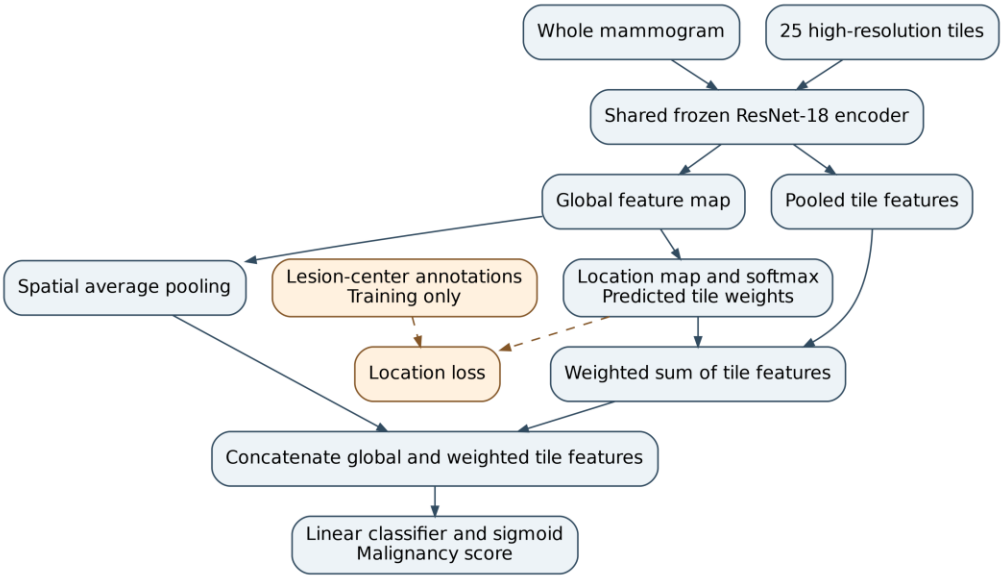


Figure 5. Dual-resolution mammogram model

3.4. Matched Comparison Model

The image-label-only comparison model shared the proposed model's frozen encoder, whole-mammogram input, tile grid, soft tile-weighting mechanism, and classification head. The location-loss weight was set to $\lambda = 0$, so the image-level classification loss alone guided the learning of tile weights. Both variants used identical patient partitions, preprocessing, augmentation, fold-specific initialization seeds, and training budgets. This matched design allowed us to evaluate the contribution of lesion-location supervision while holding the architecture and training protocol fixed.

The evaluation used five stratified, patient-grouped outer folds. In each iteration, one fold was held out for testing, and the remaining patients were divided into training and validation subsets using a grouped inner split. Within each iteration, all eligible mammograms from a patient were assigned to the same subset. Patient identifiers were checked to confirm that the training, validation, and test subsets were disjoint and that each subset contained both benign and malignant images. Each eligible mammogram contributed one out-of-fold test prediction per model.

Each model variant was trained for 15 epochs per outer fold using the AdamW optimizer [25], a learning rate of $3 \times 10^{-4}$and a batch size of four. Only the location and classification heads were optimized. The frozen encoder was used to compute features for the original and horizontally flipped images once, after which the features were reused across folds. Training batches could sample either orientation. Validation and test predictions used the original orientation. For each model and fold, the checkpoint with the highest validation area under the receiver operating characteristic curve (ROC-AUC) was selected. The classification threshold was determined on the same validation subset by maximizing Youden's index [26], defined as sensitivity plus specificity minus one. The selected checkpoint and threshold were then applied to the corresponding held-out test fold. The training objective, validation log loss, and validation ROC-AUC were recorded at each epoch.

3.5. Classification and Localization

The primary classification metric was ROC-AUC, calculated from the malignancy scores pooled across the five held-out folds. Average precision summarized the precision–recall relationship. Accuracy, balanced accuracy,

sensitivity, specificity, and confusion matrices were calculated using the classification thresholds selected on the corresponding validation subsets. Malignancy was treated as the positive class, and balanced accuracy was calculated as the arithmetic mean of sensitivity and specificity.

The paired difference in ROC-AUC was calculated by subtracting the image-label-only model's ROC-AUC from that of the location-supervised model. A 95% percentile confidence interval was estimated using 2,000 bootstrap resamples at the patient level. Each sampled patient contributed all eligible mammograms and the paired out-of-fold predictions from both models. The interval quantifies sampling variation conditional on the fitted models and their out-of-fold predictions.

Lesion localization was assessed using top-tile lesion coverage, defined as the proportion of held-out mammograms in which the highest-weighted tile contained at least one annotated lesion center. This analysis included the 111 mammograms with usable lesion-center coordinates. After prediction, the coordinates were used to determine whether the selected tile covered a lesion center. The resulting metric measures lesion-center coverage at the tile level.

3.6. Sensitivity Analysis

The crop experiment retained the original training, validation, and test assignments and restricted each partition to mammograms with usable centers. A 256 × 256 native-resolution crop was extracted around each distinct annotated center, with zero padding at image boundaries. Crops used the same pixel normalization and frozen third-stage ResNet-18 encoder as the tile branch. Spatially pooled features were averaged across crops within each mammogram, then passed to a linear image-level classifier. A secondary probe used the fourth residual stage of the same frozen encoder to examine sensitivity to feature depth. Training used the same loss weighting, optimizer, epoch budget, flip probability, and validation checkpoint rule as the full-image models. The crop experiment used annotation coordinates at evaluation and therefore represents classification with a supplied lesion location.

Uniform random coverage was calculated exactly for each mammogram as the fraction of the 25 tiles containing at least one annotated center. The fixed baseline always selected the central tile, with top-left pixel (384, 384). Coverage differences used 2,000 paired patient-level percentile bootstrap replicates, retaining all eligible mammograms from each sampled patient. A paired interval for the crop ROC-AUC difference versus the supervised full-image model used the same resamples and compared scores on the common coordinate-eligible subset.

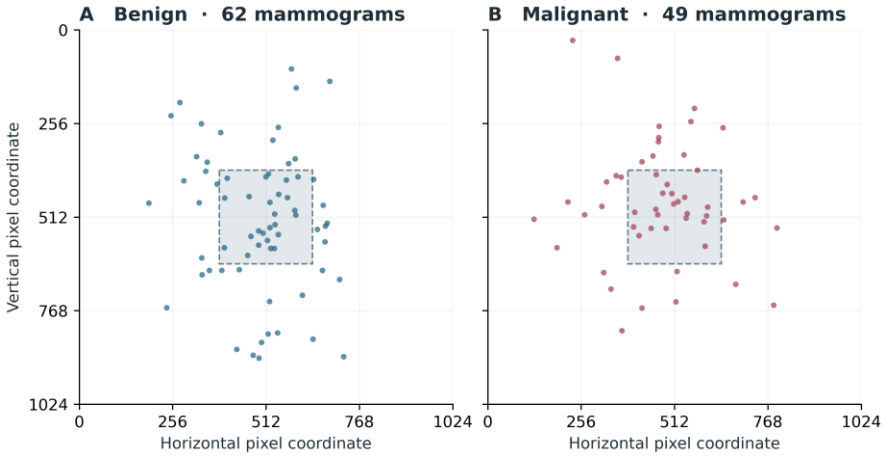


Figure 6. Mini_MIAS Lesion Center Distribution

The reference training run used fold-specific seeds of 42 plus the outer fold number. Four additional runs used base seeds of 142, 242, 342, and 442, each plus the fold number, with the original patient assignments held fixed. The paired

full-image variants shared the seed within each fold and run. Pooled and within-fold ROC-AUC were calculated from held-out scores. The unweighted mean of the five within-fold ROC-AUC values was used to assess discrimination without ranking predictions across independently fitted folds.

IV. RESULTS

4.1. Sample and Fold Composition

The evaluation included 115 mammograms from 108 patients. Of these, 63 were labeled benign (54.8%) and 52 malignant (45.2%). Both model variants used identical patient-grouped partitions. Across the five outer folds, the training, validation, and held-out test subsets contained 68–70, 23–24, and 22–24 mammograms, respectively. Each test fold included 12–13 benign and 10–11 malignant mammograms. Each mammogram contributed one held-out prediction per model.

4.2. Training and Validation Performance

Each model was trained for 15 epochs per outer fold, yielding 150 epoch records across both variants. The selected checkpoints ranged from epoch 1 to epoch 14 (Table 1). The image-label-only model achieved higher validation area under the receiver operating characteristic curve (ROC-AUC) in folds 2–4. Both variants achieved the same validation ROC-AUC in folds 1 and 5. Mean validation ROC-AUC across the selected checkpoints was 0.454 for the image-label-only model and 0.438 for the location-supervised model. Both achieved their highest validation ROC-AUC in fold 2.

Table 1. Fold composition and checkpoint selection

Fold	Training images	Validation images	Test (B / M)	Image labels only (epoch / ROC-AUC)	Location supervised (epoch / ROC-AUC)
1	69	23	13 / 10	1 / 0.455	1 / 0.455
2	69	23	13 / 10	14 / 0.561	4 / 0.538
3	68	24	12 / 11	3 / 0.392	2 / 0.350
4	68	23	13 / 11	14 / 0.446	14 / 0.431
5	70	23	12 / 10	1 / 0.415	1 / 0.415

Values summarize the five partitions and the checkpoint selected using each validation set. Mean validation log loss changed from 0.695 at epoch 1 to 0.694 at epoch 15 for both variants, to three decimal places. Figure 4 shows the trajectories of the training objective, validation log loss, and validation ROC-AUC. The location-supervised training objective includes an additional location loss. Validation log loss and ROC-AUC were calculated on the same basis for both variants.

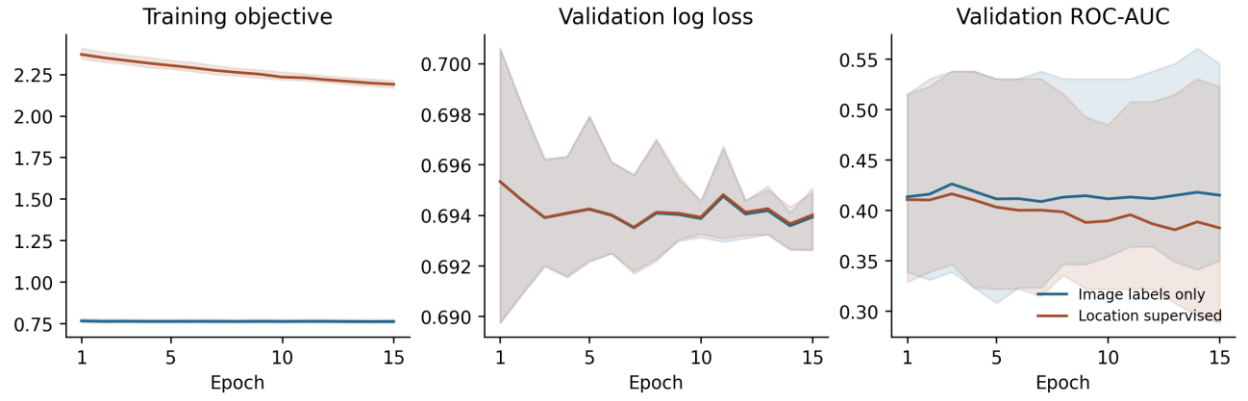


Figure 7. Training and validation trajectories

4.3. Classification Performance

The location-supervised model correctly classified 58 of 115 mammograms (50.4%), compared with 54 of 115 (47.0%) for the image-label-only model. Balanced accuracy was 49.2% and 47.9%, respectively (Table 2). Both accuracy values were below the 54.8% obtained by assigning every mammogram the benign label. At the validation-selected thresholds, location supervision yielded higher specificity and lower sensitivity.

Table 2. Held-out classification performance

Metric	Image labels only	Location supervised
ROC-AUC	0.491	0.493
Average precision	0.482	0.485
Accuracy	47.0%	50.4%
Balanced accuracy	47.9%	49.2%
Sensitivity	57.7%	36.5%
Specificity	38.1%	61.9%

Pooled held-out ROC-AUC was 0.491 for the image-label-only model and 0.493 for the location-supervised model. Both estimates were close to the chance reference of 0.50. The paired difference was +0.002 (95% patient-bootstrap confidence interval −0.006 to +0.011), based on the fixed out-of-fold predictions. The interval included zero. Average precision was 0.482 for the image-label-only model and 0.485 for the location-supervised model. Both values were slightly above the malignant-class prevalence of 0.452, the reference level for precision–recall analysis. Figure 5 shows the pooled curves.

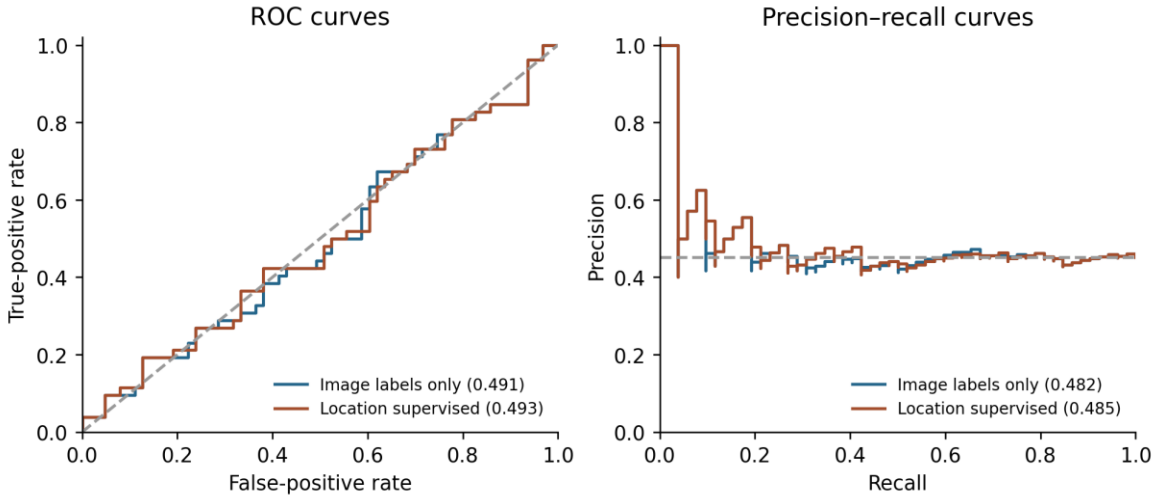


Figure 8. ROC and precision–recall curves

At the validation-selected thresholds, the location-supervised model correctly identified 39 of 63 benign mammograms and 19 of 52 malignant mammograms. The image-label-only model correctly identified 24 benign and 30 malignant mammograms. Location supervision was accompanied by 15 fewer false positives and 11 more false negatives (Figure 9).

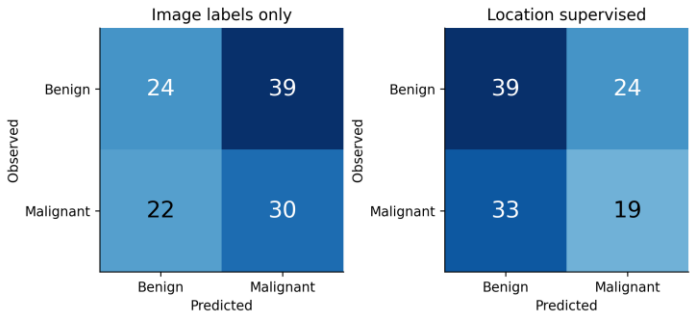


Figure 9. Held-out confusion matrices

4.4. Lesion Localization Performance

Among the 111 mammograms with usable centers, the supervised model selected a tile containing a lesion center in 26 images (23.4%), compared with 7 images (6.3%) for the image-label-only model (Figure 10). The paired gain was 17.1 percentage points (95% patient-bootstrap confidence interval 9.1 to 25.5). The supervised model missed all centers in 85 mammograms. Supervised coverage exceeded image-label-only coverage in every run, ranging from 18.0% to 23.4% versus 4.5% to 10.8%.

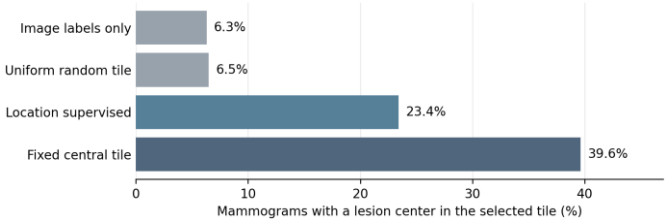


Figure 10. Top-tile lesion-center coverage

Uniform random tile selection yielded 6.5% expected coverage. The fixed central tile covered a lesion center in 44 of 111 images (39.6%). Relative to these baselines, supervised coverage differed by 16.9 and -16.2 percentage points, respectively (95% confidence intervals 9.2 to 25.0 and -26.8 to -6.2).

Table 3. Held-out ROC-AUC by fold in the reference run

Held-out fold	Image labels only	Location supervised
1	0.638	0.623
2	0.508	0.515
3	0.545	0.561
4	0.322	0.301
5	0.400	0.400
Mean across folds	0.483	0.480
Pooled ROC-AUC	0.491	0.493

Within-fold estimates compare predictions from the same fitted model. Pooled estimates rank predictions across all five test folds. Across the five runs, pooled ROC-AUC ranged from 0.470 to 0.505 for image labels alone and 0.480 to 0.511 for location supervision (Table 4). The paired change in pooled ROC-AUC ranged from -0.013 to 0.026. The mean within-fold ROC-AUC ranged from 0.460 to 0.493 and from 0.458 to 0.500 for the two models, respectively.

Table 4. Pooled held-out ROC-AUC across training seeds

Base seed	Image labels only	Location supervised	Crop, stage 3	Crop, stage 4
42	0.491	0.493	0.488	0.514
142	0.505	0.511	0.524	0.431
242	0.470	0.496	0.468	0.514
342	0.485	0.504	0.483	0.529
442	0.493	0.480	0.484	0.416

The full-image columns use 115 mammograms. The annotation-centered crop columns use 111 mammograms with usable centers, assigned to the same patient partitions. The coordinate-eligible subset comprised 62 benign and 49 malignant mammograms from 104 patients. In the reference run, the crop classifier achieved a pooled ROC-AUC of 0.488 with stage-three features and 0.514 with stage-four features. On the same mammograms, the image-label-only and location-supervised full-image models achieved 0.506 and 0.508, respectively. The paired difference between the stage-three crop and location-supervised full-image models was -0.020 (95% patient-bootstrap confidence interval -0.186 to 0.152). Across seeds, crop ROC-AUC ranged from 0.468 to 0.524 with stage-three features and 0.416 to 0.529 with stage-four features.

V. DISCUSSION

Adding lesion-location supervision shifted tile selection toward annotated lesions while leaving full-image benign-versus-malignant discrimination close to chance in the reference run. The supervision targets spatial proximity, and a tile that covers a lesion center can contain substantial surrounding tissue. The localization estimates describe center coverage at the tile level. They do not describe segmentation quality or the ability to detect previously unidentified abnormalities. The supervised model exceeded uniform random selection (6.5%) but fell below the fixed central tile (39.6%). This comparison suggests that lesion centers were spatially concentrated in this dataset and that a simple location prior can outperform learned tile weighting on this coverage measure.

Supplying lesion centers at evaluation yielded no consistent improvement across seeds with either the third- or fourth-stage frozen features. The wide paired interval leaves room for a meaningful difference in either direction. The crop probes test class information accessible to a linear head under the selected crop size and pooling scheme. They leave open whether alternative representations, fine-tuning, or a larger training sample would improve discrimination. Access to lesion centers at evaluation makes this experiment a diagnostic comparison within the study. It does not estimate the performance of a pixel-only system on previously unannotated images.

Between-seed variation reflects changes in initialization, training-example order, and augmentation on the same held-out patients. The paired patient bootstrap quantifies variability across patients conditional on the fitted predictions. It does not capture the variability that would arise from refitting on new cohorts. The small individual folds also make checkpoint selection and threshold estimates sensitive to a few images. The observed sensitivity-specificity trade-off should be interpreted in that context.

The study uses a single collection of digitized film mammograms and excludes normal examinations. Consequently, the results address discrimination among already-abnormal mammograms with Mini-MIAS severity labels. The crop analysis further depends on the availability of usable spatial annotations. No external test set was used, and the frozen ImageNet encoder, fixed grid, and small sample constrain generalizability to other imaging systems or clinical populations. External screening evaluations, including the comparison of three commercial algorithms by Salim et al., address a broader population and a different clinical endpoint [27]. Freeman et al. documented variation in the design and quality of evidence for mammography screening algorithms [28]. Future evaluation can use annotated full-field digital datasets such as VinDr-Mammo [29], with explicit harmonization of label definitions and imaging protocols. Reporting acquisition, partitions, and evaluation procedures in that setting is consistent with the Checklist for Artificial Intelligence in Medical Imaging [30].

VI. CONCLUSION

This study evaluated whether lesion-location supervision during training improves benign-versus-malignant classification on full mammograms. In a patient-grouped evaluation of 115 abnormal Mini-MIAS mammograms from 108 patients, supervision increased the proportion of images in which the highest-weighted tile contained an annotated lesion center from 6.3% to 23.4%. The paired gain was 17.1 percentage points (95% patient-bootstrap confidence interval, 9.1 to 25.5). Supervised coverage exceeded the 6.5% expected under uniform random tile selection and fell below the 39.6% achieved by selecting the fixed central tile. Full-mammogram discrimination remained close to chance. In the reference run, the area under the receiver operating characteristic curve (ROC-AUC) was 0.491 with image labels alone and 0.493 with location supervision. Across five training seeds, supervised ROC-AUC ranged from 0.480 to 0.511. Supplying annotated lesion centers at evaluation also yielded no consistent classification improvement across the two frozen feature stages and five seeds. Under the evaluated protocol, lesion-location supervision improved tile selection relative to image-label-only training, while malignancy discrimination remained limited even when the lesion location was supplied. These findings apply to abnormal digitized film mammograms from a single dataset and to the specific frozen encoder and tile grid evaluated. Future studies could test whether mammography-specific feature learning improves discrimination and whether any gains generalize to independent digital mammography cohorts. Reporting lesion coverage alongside classification performance would help establish which part of the pipeline improves.

REFERENCES

[1] Shen, Y., Wu, N., Phang, J., Park, J., Kim, G., Moy, L., Cho, K., & Geras, K. J. (2019). Globally-aware multiple instance classifier for breast cancer screening. In *Machine Learning in Medical Imaging* (pp. 18–26). Springer.

[2] Logan, J., Kennedy, P. J., & Catchpoole, D. (2023). A review of the machine learning datasets in mammography, their adherence to the FAIR principles and the outlook for the future. *Scientific Data, 10*, Article 595.

[3] Kag, H., Kumar, D., & Diwan, A. (2026). Breast lesion classification: Leakage-free patch-based transfer learning with cross-dataset evaluation on CBIS-DDSM, MIAS, and INbreast. *Jordanian Journal of Computers and Information Technology, 12*(3), 388–411.

[4] Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S. G., Moy, L., Cho, K., & Geras, K. J. (2021). An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical Image Analysis, 68, 101908*. https://doi.org/10.1016/j.media.2020.101908

[5] Ribli, D., Horváth, A., Unger, Z., Pollner, P., & Csabai, I. (2018). Detecting and classifying lesions in mammograms with deep learning. *Scientific Reports, 8, 4165*. https://doi.org/10.1038/s41598-018-22437-z

[6] Lotter, W., Diab, A. R., Haslam, B., Kim, J. G., Grisot, G., Wu, E., Wu, K., Onieva Onieva, J., Boyer, Y., Boxerman, J. L., Wang, M., Bandler, M., Vijayaraghavan, G. R., & Sorensen, A. G. (2021). Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nature Medicine, 27, 244–249*. https://doi.org/10.1038/s41591-020-01174-9

[7] Chen, Y., Liu, Y., Wang, C., Elliott, M., Kwok, C. F., Peña-Solorzano, C., Tian, Y., Liu, F., Frazer, H., McCarthy, D. J., & Carneiro, G. (2024). BRAIxDet: Learning to detect malignant breast lesions with incomplete annotations. *Medical Image Analysis, 96*, 103192.

[8] Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data, 4, 170177.* https://doi.org/10.1038/sdata.2017.177

[9] Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). INbreast: Toward a full-field digital mammographic database. *Academic Radiology, 19(2), 236–248.* https://doi.org/10.1016/j.acra.2011.09.014

[10] Zhu, M., Zhang, L., Wang, L., Wang, Z., Wang, Y., & Qian, G. (2025). Local extremum mapping for weak supervision learning on mammogram classification and localization. *Bioengineering, 12*(4), 325.

[11] Mahmood, T., Li, J., Pei, Y., Akhtar, F., Rehman, M. U., & Wasti, S. H. (2022). Breast lesions classifications of mammographic images using a deep convolutional neural network-based approach. *PLOS ONE, 17*(1), e0263126.

[12] Avcı, H., & Karakaya, J. (2024). Segmentation techniques in mammography: Their value in breast cancer diagnosis. *Journal of Clinical Practice and Research, 46*(6), 548–556.

[13] Teoh, J. R., Hasikin, K., Lai, K. W., Wu, X., & Li, C. (2024). Enhancing early breast cancer diagnosis through automated microcalcification detection using an optimized ensemble deep learning framework. *PeerJ Computer Science, 10*, e2082.

[14] Zebari, D. A., Ibrahim, D. A., Zeebaree, D. Q., Haron, H., Salih, M. S., Damaševičius, R., & Mohammed, M. A. (2021). Systematic review of computing approaches for breast cancer detection based on computer-aided diagnosis using mammogram images. *Applied Artificial Intelligence, 35*(15), 2157–2203.

[15] Zebari, D. A., Ibrahim, D. A., Zeebaree, D. Q., Mohammed, M. A., Haron, H., Zebari, N. A., Damaševičius, R., & Maskeliūnas, R. (2021). Breast cancer detection using mammogram images with improved multi-fractal dimension approach and feature fusion. *Applied Sciences, 11*(24), 12122. https://doi.org/10.3390/app112412122

[16] Zebari, D. A., Haron, H., Sulaiman, D. M., Yusoff, Y., & Othman, M. N. M. (2022). CNN-based deep transfer learning approach for detecting breast cancer in mammogram images. In *2022 IEEE 10th Conference on Systems, Process & Control (ICSPC)* (pp. 256–261). IEEE.

[17] Zebari, D. A., Ibrahim, D. A., & Al-Zebari, A. (2022). Suspicious region segmentation using deep features in breast cancer mammogram images. In *2022 International Conference on Computer Science and Software Engineering (CSASE)* (pp. 253–258). IEEE.

[18] Qian, L., Bai, J., Huang, Y., Zeebaree, D. Q., Saffari, A., & Zebari, D. A. (2024). Breast cancer diagnosis using evolving deep convolutional neural network based on hybrid extreme learning machine technique and improved chimp optimization algorithm. *Biomedical Signal Processing and Control, 87, 105492.* https://doi.org/10.1016/j.bspc.2023.105492

[19] Cao, J., Zeebaree, D. Q., Chen, Q., & Zebari, D. A. (2023). Breast cancer diagnosis using hybrid AlexNet-ELM and chimp optimization algorithm evolved by Nelder–Mead simplex approach. *Biomedical Signal Processing and Control, 85, 105053.* https://doi.org/10.1016/j.bspc.2023.105053

[20] Zebari, N. A., Alkurdi, A. A. H., Marqas, R. B., & Salih, M. S. (2023). Enhancing brain tumor classification with data augmentation and DenseNet121. *Academic Journal of Nawroz University, 12(4).* https://journals.nawroz.edu.krd/files/article/view/864

[21] Zebari, D. A., Sulaiman, D. M., Sadiq, S. S., Zebari, N. A., & Salih, M. S. (2022). Automated detection of COVID-19 from X-ray using SVM. *2022 4th International Conference on Advanced Science and Engineering (ICOASE), 130–135. IEEE.* https://doi.org/10.1109/ICOASE56293.2022.10075586

[22] Salih, M. S., Zebari, N. A., & Doski, H. (2025). A hybrid LBP and CNN-based approach for COVID-19 detection using chest X-ray images. *Artificial Intelligence & Robotics Development Journal, 5(4), 396–408.* https://doi.org/10.52098/airdj.20255465

[23] Suckling, J., et al. (1994). The Mammographic Image Analysis Society Digital Mammogram Database. International Congress Series, 1069, 375–378. https://www.andrew.cmu.edu/course/15-440-s12/applications/labs/lab4/mias.html

[24] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.* https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[25] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations.* https://openreview.net/forum?id=Bkg6RiCqY7

[26] Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3(1), 32–35.* https://pubmed.ncbi.nlm.nih.gov/15405679/

[27] Salim, M., Wåhlin, E., Dembrower, K., Azavedo, E., Foukakis, T., Liu, Y., Smith, K., Eklund, M., & Strand, F. (2020). External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncology, 6(10), 1581–1588.* https://doi.org/10.1001/jamaoncol.2020.3321

[28] Freeman, K., Geppert, J., Stinton, C., Todkill, D., Johnson, S., Clarke, A., & Taylor-Phillips, S. (2021). Use of artificial intelligence for image analysis in breast cancer screening programs: Systematic review of test accuracy. *BMJ, 374, n1872.* https://doi.org/10.1136/bmj.n1872

[29] Nguyen, H. T., Nguyen, H. Q., Pham, H. H., Lam, K., Le, L. T., Dao, M., & Vu, V. (2023). VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data, 10, 277.* https://doi.org/10.1038/s41597-023-02100-7

[30] Tejani, A. S., Klontzas, M. E., CLAIM 2024 Update Panel, Gatti, A. A., Mongan, J. T., Moy, L., Park, S. H., & Kahn, C. E., Jr. (2024). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiology: Artificial Intelligence, 6(4), e240300.* https://doi.org/10.1148/ryai.240300