Beyond the Tumor Signature: Investigating Global Mammographic Context in Machine-Learning–Assisted Breast Imaging

Tanish Jain

Rock Ridge HS, VA, USA

DOI: 10.29322/IJSRP.16.08.2026.p17616
https://dx.doi.org/10.29322/IJSRP.16.08.2026.p17616

Paper Received Date: 26th July 2026
Paper Acceptance Date: 24th August 2026
Paper Publication Date: 31st August 2026

Abstract

Artificial intelligence (AI) is increasingly studied for mammographic image analysis, but model performance depends on preprocessing, dataset design, and evaluation choices. This study tested whether focusing a traditional machine-learning pipeline on an annotated lesion improves classification of benign versus malignant abnormalities. A Kaggle-distributed copy of the mini-MIAS mammography dataset was used, with dataset identity and annotation definitions checked against the MIAS documentation [1]. Abnormal mammograms with valid lesion coordinates were analyzed using two strategies: the whole mammogram and a lesion-centered region of interest (ROI). Both used identical Histogram of Oriented Gradients (HOG) features [2] and class-weighted logistic regression. Performance was evaluated with 10 repeats of 5-fold stratified, patient-grouped cross-validation. Whole-mammogram processing produced mean ROC-AUC 0.571 (95% CI 0.540–0.602), compared with 0.532 (95% CI 0.496–0.567) for ROI processing. The paired ROI-minus-whole difference was −0.039 (95% CI −0.080–0.001; paired t-test p=0.055). Sensitivity was 0.537 for whole images and 0.498 for ROI images. Lesion-focused cropping therefore did not improve this HOG-plus-logistic-regression classifier. Both approaches performed only modestly above chance, reinforcing recommendations that medical-AI studies report clinically relevant metrics, uncertainty, and transparent validation rather than relying on accuracy alone [7,10,11].

1. Introduction

Mammography is central to breast-cancer screening, and computer-aided detection and diagnosis (CAD) has been investigated for decades. Earlier CAD systems typically used explicit image preprocessing, lesion detection or segmentation, hand-engineered features, and statistical or conventional machine-learning classifiers. More recent systems increasingly use deep learning to learn image representations directly from large collections of mammograms [3].

The growth of AI in mammography has produced promising results but also methodological concerns. A 2021 systematic review of AI for breast-screening mammography identified 12 studies involving 131,822 screened women but found no prospective studies measuring AI test accuracy in actual screening practice at that time [7]. A later systematic review and meta-analysis of standalone AI included 16 studies and more than 1.1 million examinations; reported AUCs were high in several settings, but performance varied by study type and imaging modality [8]. These results demonstrate the potential of modern AI, but they also show why performance from different datasets and study designs cannot be compared as if all systems were tested under identical conditions.

This project intentionally studies a narrower question. Instead of attempting to reproduce a large deep-learning screening system on a small educational dataset, it uses a controlled traditional pipeline to isolate one experimental variable: whether the classifier receives the whole mammogram or a crop centered on the known lesion. The aim is therefore methodological rather than clinical: to determine whether lesion localization alone improves the information available to a HOG-plus-logistic-regression classifier.

2. Literature Review and Related Work

2.1 From Traditional CAD to Deep Learning

Traditional breast-imaging CAD relied heavily on human-designed image descriptors. A review of CAD and AI in breast cancer describes conventional systems as pipelines in which developers translate visual characteristics into mathematical features, while noting that such hand-engineered representations may fail to capture complex differences across patients [3]. HOG is one example of a hand-engineered descriptor. Dalal and Triggs introduced HOG as a representation of local gradient orientations, showing that local gradients, spatial binning, and contrast normalization can provide useful shape information [2]. Although HOG was not created for mammography, its interpretable edge- and structure-based representation makes it suitable for a simple controlled experiment.

Deep learning changed this design by learning hierarchical features directly from images. For example, Shen et al. developed an end-to-end mammography system that could use both lesion annotations and image-level cancer labels; on an independent CBIS-DDSM test set, their best ensemble reached per-image AUC 0.91 [5]. McKinney et al. later evaluated a deep-learning breast-screening system using large US and UK datasets and reported performance competitive with human readers, while also examining workflow implications [6]. These systems are not direct benchmarks for the present study: they use different datasets, training sizes, architectures, labels, and validation procedures. Their relevance is that they illustrate how modern learned representations differ fundamentally from the small, hand-engineered HOG pipeline tested here.

2.2 Whole-Mammogram Context versus Lesion-Focused Analysis

ROI-based analysis is intuitively attractive because mammographic lesions may occupy only a small fraction of the image. Cropping can suppress unrelated background and magnify the suspicious area. However, whole-image analysis can preserve contextual information such as surrounding tissue, lesion scale, breast density, and relationships between an abnormality and the broader breast structure. Modern mammography systems use both strategies, including whole-image classifiers, lesion detectors, and combined multi-stage systems [3,5].

Recent work shows that the effect of ROI information is not universal. A 2026 study of ROI-stratified deep learning across multiple public mammography datasets reported that ROI-related stratification could improve classification, particularly as data volume increased [12]. That work used modern architectures such as ResNet, EfficientNet, Swin Transformer, ConvNeXt, and MobileNet, making it methodologically different from the present HOG experiment. This contrast strengthens the current research question: an ROI can be useful in some learning frameworks, but it does not follow that simply cropping around a lesion will improve every classifier.

2.3 Evaluation, Errors, and Clinical Relevance

Medical classification cannot be judged by accuracy alone. Sensitivity measures the fraction of malignant cases detected, whereas specificity measures the fraction of non-malignant cases correctly rejected. ROC-AUC summarizes ranking discrimination across thresholds. This distinction matters because screening systems face a trade-off between false positives and false negatives. A 2024 systematic review of AI errors in screening mammography found that false-positive and false-negative behavior changed with positivity thresholds, algorithm versions, and study quality [9].

The importance of transparent evaluation is also reflected in current reporting guidance. TRIPOD+AI, published in 2024, provides updated reporting recommendations for prediction models developed or evaluated with regression or machine-learning methods [10]. PROBAST+AI, published in 2025, emphasizes quality, risk of bias, applicability, data sources, predictors, outcomes, and analysis when assessing healthcare prediction models [11]. These principles motivated this project's use of grouped validation, multiple clinically interpretable metrics, confidence intervals, and explicit limitations.

2.4 Research Gap and Study Objective

The literature establishes three points: conventional CAD depends strongly on feature engineering [3]; modern deep-learning mammography systems can achieve substantially stronger discrimination on much larger

datasets [5–8]; and ROI information can help under some modeling conditions but does not guarantee

improvement [12]. For a small public dataset such as mini-MIAS, there is therefore value in a controlled experiment that changes only the spatial preprocessing while holding the feature extractor, classifier, and validation procedure constant. The present study asks whether lesion-centered ROI preprocessing improves benign-versus-malignant discrimination over whole-mammogram preprocessing when both are represented with HOG and classified with logistic regression.

3. Research Question and Hypothesis

Research question: Does lesion-focused ROI preprocessing improve benign-versus-malignant classification compared with whole-mammogram preprocessing when HOG features, logistic regression, and the evaluation procedure are held constant?

Hypothesis: Because MIAS provides the location of the annotated abnormality, cropping around that lesion was expected to reduce irrelevant background information and improve discrimination between benign and malignant abnormalities.

4. Materials and Methods

4.1 Dataset

The experiment used a Kaggle-distributed copy of the mini-MIAS dataset. Dataset identity was checked using MIAS reference numbers and metadata fields. The mini-MIAS documentation describes images standardized to 1024 × 1024 pixels and truth data containing background-tissue type, abnormality class, severity, lesion-center coordinates, and an approximate lesion radius [1]. The local metadata contained 330 annotation rows corresponding to 322 unique mammograms after repeated image references were collapsed: 207 normal, 63 benign, and 52 malignant. For the ROI comparison, normal mammograms were excluded because they do not have lesion coordinates; abnormal mammograms with valid X, Y, and RADIUS annotations were eligible.

When a mammogram had multiple annotation rows, the pipeline retained one image-level sample to avoid treating the same image as independent observations. Malignant severity took precedence over benign severity; within the same severity, the larger-radius lesion was retained.

4.2 Image Preprocessing

Two strategies were compared. For whole-mammogram processing, each grayscale mammogram was intensity-normalized, resized to 256 × 256 pixels, and contrast-enhanced with adaptive histogram equalization. For ROI processing, a square crop was centered on the MIAS lesion coordinate. The crop half-width was 1.5 times the annotated radius, with a minimum half-width of 32 pixels; the ROI was then resized to 256 × 256 and contrast-enhanced identically. The MIAS coordinate convention required conversion of the vertical coordinate before array-based cropping.

4.3 HOG Features and Logistic Regression

HOG features were extracted with 9 orientations, 16 × 16-pixel cells, 2 × 2-cell blocks, and L2-Hys block normalization, following the general HOG principle of locally normalized gradient-orientation histograms [2]. Feature vectors were standardized and classified with logistic regression using balanced class weights. The same model settings were used for both strategies so that preprocessing remained the main experimental variable.

4.4 Cross-Validation and Leakage Control

Performance was estimated using 10 repeats of 5-fold StratifiedGroupKFold cross-validation, producing 50 paired folds per preprocessing strategy. Related MIAS mammogram pairs were assigned to the same group so they could not be split between training and testing. Identical fold assignments were used for whole-image and ROI features, enabling paired comparisons. This design follows the broader medical-AI principle that related observations should not leak across model-development and evaluation partitions [10,11].

4.5 Metrics and Statistical Analysis

Metrics were accuracy, balanced accuracy, sensitivity, specificity, precision, F1 score, and ROC-AUC. Mean scores and 95% confidence intervals were calculated across repeated cross-validation folds. Paired ROI-minus-whole differences were assessed with a paired t-test and Wilcoxon signed-rank test, and Cohen's dz was calculated as a paired effect-size measure. Because repeated cross-validation reuses observations across repeats, fold-level inferential statistics are treated as exploratory rather than as 50 independent experiments.

5. Results

Whole-mammogram preprocessing had a higher mean score than ROI preprocessing for all seven metrics. None of the paired comparisons crossed the conventional $p<0.05$ threshold. ROC-AUC was closest: paired t-test $p=0.055$ and Wilcoxon $p=0.059$.

Metric	Whole mean (95% CI)	ROI mean (95% CI)	ROI − Whole (95% CI)	Paired t p	Cohen dz
Accuracy	0.558 (0.529–0.587)	0.522 (0.491–0.552)	-0.037 (-0.077–0.003)	0.068	-0.263
Balanced Accuracy	0.557 (0.527–0.586)	0.516 (0.483–0.550)	-0.040 (-0.082–0.002)	0.062	-0.271
Sensitivity	0.537 (0.488–0.586)	0.498 (0.451–0.546)	-0.038 (-0.100–0.023)	0.216	-0.177
Specificity	0.576 (0.540–0.613)	0.534 (0.489–0.580)	-0.042 (-0.101–0.017)	0.156	-0.204
Precision	0.499 (0.453–0.546)	0.459 (0.404–0.515)	-0.040 (-0.082–0.002)	0.062	-0.270
F1 Score	0.504 (0.462–0.547)	0.468 (0.421–0.515)	-0.036 (-0.083–0.011)	0.130	-0.218
ROC-AUC	0.571 (0.540–0.602)	0.532 (0.496–0.567)	-0.039 (-0.080–0.001)	0.055	-0.278

Table 1. Repeated grouped cross-validation results. Differences are ROI minus whole-mammogram performance.

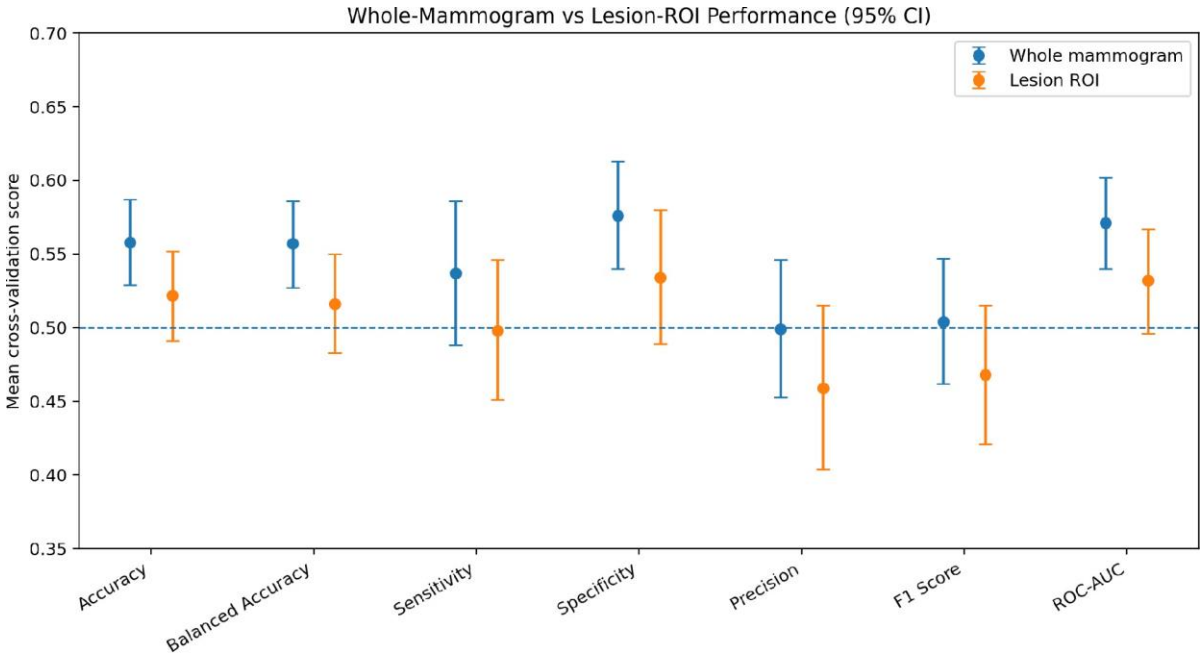

Figure 1. Mean performance with 95% confidence intervals for whole-mammogram and lesion-ROI preprocessing.

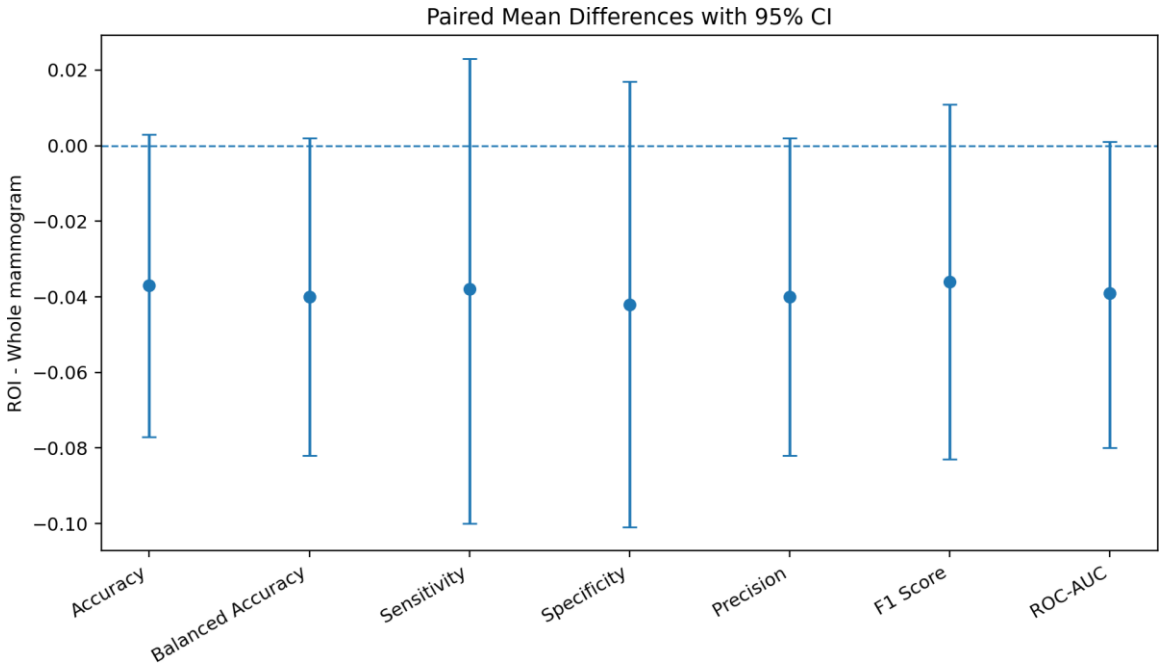

Figure 2. Paired mean differences (ROI − whole mammogram) with 95% confidence intervals. Every interval includes zero.

5.1 Primary Findings

Mean ROC-AUC was 0.571 (95% CI 0.540–0.602) for whole-mammogram features and 0.532 (95% CI 0.496–0.567) for ROI features. The paired ROI-minus-whole difference was −0.039 (95% CI −0.080–0.001), paired t-test p=0.055, Wilcoxon p=0.059, and Cohen's dz=−0.278. Sensitivity was 0.537 versus 0.498, and specificity was 0.576 versus 0.534. Balanced accuracy was 0.557 versus 0.516. Thus, point estimates consistently favored whole-image processing, but uncertainty intervals did not support a conclusive difference.

6. Discussion

The hypothesis was not supported. Lesion-centered cropping did not improve benign-versus-malignant classification with HOG and logistic regression. Instead, whole-mammogram processing produced slightly higher mean performance across all metrics, although the paired differences were not conventionally statistically significant.

This result is plausible in light of prior CAD research. Hand-engineered descriptors encode only selected aspects of image appearance and may not capture subtle differences across heterogeneous clinical images [3]. HOG emphasizes gradients, edges, and local shape [2]; malignant mammographic appearance can involve more complex texture, margins, surrounding tissue, density, and spatial context. Cropping may therefore remove useful contextual information while also rescaling lesions of different physical sizes to a common image size.

The finding should not be generalized to ROI methods as a whole. Recent ROI-stratified deep-learning research has reported benefits from ROI information in substantially different architectures and datasets [12]. The contrast suggests that the usefulness of localization depends on how ROI information is represented and learned, as well as dataset size. The present study tests only the simpler claim that direct lesion cropping improves a HOG-plus-logistic-regression pipeline; it did not.

The weak absolute performance is also important. Whole-image ROC-AUC of 0.571 and ROI ROC-AUC of 0.532 are far below values reported for large modern deep-learning screening studies and meta-analyses [5–8]. That comparison is contextual rather than head-to-head: the datasets, populations, tasks, image technology, model capacity, and validation designs differ substantially. It would be misleading to conclude that deep learning

is '0.3 AUC better' based on these unrelated experiments.

The project's earlier exploratory run further demonstrated why accuracy alone is unsafe in imbalanced medical classification. In that run, normal and benign mammograms were combined as non-malignant, yielding 75.3% accuracy but 0% sensitivity: all 13 malignant test cases were missed. Systematic review evidence likewise shows that mammography-AI evaluation must consider false-positive and false-negative behavior, which changes with operating thresholds and study conditions [9]. This motivated the present emphasis on sensitivity, specificity, balanced accuracy, ROC-AUC, confidence intervals, and grouped validation, consistent with current transparent-reporting principles [10,11].

7. Limitations

• The mini-MIAS dataset is small, and the abnormal-only experiment uses only a subset of the 322 mammograms. Estimates therefore have substantial uncertainty.

• mini-MIAS is an older digitized-film resource; results should not be assumed to transfer to modern full-field digital mammography or tomosynthesis.

• The ROI experiment uses provided lesion annotations and therefore does not evaluate automatic lesion localization or an end-to-end screening system.

• Only one hand-engineered feature representation and one classifier were evaluated. Other features, classifiers, or deep neural networks may behave differently.

• Repeated cross-validation folds are correlated because observations reappear across repeats. The reported fold-based confidence intervals and paired p-values are exploratory, not evidence from 50 independent datasets.

• No external dataset was used for validation. Current prediction-model guidance emphasizes careful distinction between internal and external validation and attention to applicability and risk of bias [10,11].

• This is a computational educational research project with no clinical validation and must not be used for diagnosis or patient care.

8. Conclusion

In this mini-MIAS experiment, lesion-focused preprocessing did not improve benign-versus-malignant classification using HOG features and logistic regression. Whole-mammogram processing achieved mean ROC-AUC 0.571 compared with 0.532 for lesion ROIs, but the paired difference did not reach conventional statistical significance (p=0.055). The main conclusion is not that whole mammograms are universally superior. Rather, simply localizing and enlarging an annotated lesion was insufficient to make this traditional machine-learning pipeline effective. The study demonstrates the value of controlled comparisons, leakage-aware validation, class-sensitive metrics, uncertainty estimates, and honest reporting of negative results in medical-AI research.

9. Reproducibility

The project was implemented in Python using pandas, NumPy, scikit-image, scikit-learn, SciPy, and Matplotlib. The V2 workflow included an initial mammography experiment, the whole-image versus ROI comparison, and final statistical analysis. A fixed base random seed of 42 was used, with different seeds across repeated grouped cross-validation runs. Outputs included fold-level predictions and metrics, summary CSV files, confidence-interval figures, paired-difference figures, and a statistical summary.

References

[1] Suckling J, Parker J, Dance DR, et al. The Mammographic Image Analysis Society Digital Mammogram Database. In: Digital Mammography: Proceedings of the 2nd International Workshop on Digital

Mammography. Excerpta Medica, International Congress Series 1069. 1994:375–378.

[2] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005:886–893. doi:10.1109/CVPR.2005.177.

[3] Giger ML. CAD and AI for breast cancer—recent development and challenges. British Journal of Radiology. 2020;93:20190580. PMCID: PMC7362917.

[4] Li H, Giger ML, Olopade OI, Lan L. Fractal analysis of mammographic parenchymal patterns in breast cancer risk assessment. Academic Radiology. 2007;14(5):513–521. (Background example of quantitative mammographic image features.)

[5] Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Scientific Reports. 2019;9:12495. PMCID: PMC6715802.

[6] McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577:89–94. doi:10.1038/s41586-019-1799-6.

[7] Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ. 2021;374:n1872. doi:10.1136/bmj.n1872.

[8] Yoon JH, Strand F, Baltzer PAT, et al. Standalone AI for Breast Cancer Detection at Screening Digital Mammography and Digital Breast Tomosynthesis: A Systematic Review and Meta-Analysis. Radiology. 2023;307(5):e222639. doi:10.1148/radiol.222639.

[9] Zeng A, Houssami N, Noguchi N, Nickel B, Marinovich ML, et al. Frequency and characteristics of errors by artificial intelligence (AI) in reading screening mammography: a systematic review. Breast Cancer Research and Treatment. 2024;207:1–13. doi:10.1007/s10549-024-07353-3.

[10] Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.

[11] Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505.

[12] Improving Normal/Abnormal and Benign/Malignant Classifications in Mammography with ROI-Stratified Deep Learning. 2026. PubMed PMID: 41749745; PMCID: PMC12938819. Used here as recent related work on ROI-aware mammography classification.

Appendix A. Key Experimental Parameters

Parameter	Setting
Input	Grayscale mini-MIAS mammograms
Target	Benign vs Malignant
Comparison	Whole mammogram vs annotated lesion ROI
Resize	256 × 256
Contrast	Adaptive histogram equalization
HOG	9 orientations; 16×16 cells; 2×2 blocks; L2-Hys
Classifier	Class-weighted Logistic Regression
Validation	10 repeats × 5-fold StratifiedGroupKFold
Base seed	42
Primary interpretation	ROC-AUC with sensitivity, specificity, and balanced accuracy