Intelligent Multimedia Processing & Communication Systems Journal

IMPCSJ

J IMPCS (2025) 22: 19-28

DOI 10.71856/IMPCS.2025.1228092

Research Paper

Comparative Evaluation of YOLO Architectures for Automated Breast Cancer Detection in Digital Mammography

Zahra Samadi[1], Mohammad Norouzi[2*], Seyyed Omid Shahdi[3], Mona Haddad Zahmatkesh[4]

1. PhD. Student, Department of Biomedical Engineering, Qa.c., Isalamic Azad university, Qazvin, Iran.
2. Assistant Professor, Department of Computer Engineering and Information Technology, Qa.c., Islamic Azad University, Qazvin, Iran. **Corresponding Author*
3. Assistant Professor, Department of Electrical Engineering, Qa.c., Isalamic Azad university, Qazvin , Iran.
4. Assistant Professor, Department of Pharmaceutical Biotechnology-Nuclear Pharmacy, School of Pharmacy, Guilan University of Medical Sciences, Rasht, Iran

Article Info

Article history:

Received: 21 Oct 2025

Accepted: 30 Nov 2025

Keywords:

Breast Cancer Detection,
Computer-Aided Diagnosis (CAD),
Digital Mammography,
Object Detection,
YOLO.

ABSTRACT

Early detection of breast cancer is critical for improving patient survival; however, accurate interpretation of digital mammography remains challenging due to dense breast tissue, overlapping anatomical structures, and low-contrast lesions. Recent advances in deep learning particularly object detection frameworks from the YOLO family; have shown promise for automated lesion detection. Nevertheless, systematic and controlled comparisons of contemporary YOLO architectures in mammography remain limited.

This study presents a rigorous comparative evaluation of three lightweight YOLO variants YOLOv5n, YOLOv8n and YOLOv11n for automated breast lesion detection in digital mammography. Experiments were conducted on the VinDr-Mammo dataset comprising over 20,000 expert-annotated mammograms. To ensure fairness and reproducibility, all models were trained under identical conditions using a clinically validated preprocessing pipeline, including Contrast_imited_adaptive histogram equalization (CLAHE), bilateral filtering and safety preserving data augmentation. A patient wise five-fold cross validation strategy was employed.

Model performance was assessed using lesion level metrics including mean average precision, precision, recall, and F1-score, alongside image level receiver operating characteristic and precision–recall analyses.

While all models showed good performance, YOLOv11n outperformed all the other models, attaining an $mAP_{0.5}$ of 68.28% and $mAP_{0.5:0.95}$ of 40.82%, which is 5.4% and 9.2% better than YOLOv8n and YOLOv5n, respectively. YOLOv11n also displayed superior performance metrics, achieving better sensitivity (0.69) and precision (0.73) especially on small and slightly contrast lesions, all while achieving real-time performance at 92 fps and lower GPU memory usage. The results represent the best available performance and document an increase in accuracy and efficiency for the clinically actionable AI-based CAD systems.

Copyright © Author(s).

Publisher: Islamic Azad University – Zanjan Branch

I. Introduction

Despite advancements in diagnostic imaging techniques, breast cancer continues to challenge healthcare organizations globally due to its prevalent complicated diagnostic imaging[1]. Digital mammography is the main technique recommended to diagnose breast cancer early and, most importantly, diagnose lesions before they are symptomatic [2]. However, the precise identification of subtle, small, and low-contrast abnormalities is still problematic due to overlapping glandular structural abnormalities, varying levels of image noise, and wide variation in breast density. There are inconsistent abnormalities that are also interpreted as missing, adding to the false/ positive- and false negative-rate in varying levels of breast density [3]. Since the prognosis of the patient is worsened, the treatment gap also raises the cost of the interventions; the necessity to increase the precise measurement of patient outcomes is the enhanced diagnostic imaging of the interpretive mammography [4].

CAD systems provide invaluable assistance for radiologists through objective and automated detection of anomalous areas in medical images. The rapid advancements of artificial intelligence especially deep learning have revolutionized automated medical imaging through real-time feature extraction and classification of complex images [5]. Within this framework, deep learning object detection models, within the YOLO family of algorithms, have outstanding efficacy and real-time localization for different image modalities [6]. Their single-shot detection and high-speed inference cut down processing time and improve the accuracy of real-time CAD systems.

While there has been progress, there are still issues that need to be addressed. Previous research has studied the image classification and image segmentation tasks that are not localized to specific tumor borders. Also, there have been few direct comparisons of the state of the art YOLO architectures. The most recent models, YOLOV8 and YOLOV11, include more advanced backbones, better feature fusion architectures, and more powerful detection heads yielding better accuracy and performance compared to models within the same family [7–9]. However, there has yet to be a thorough assessment of the models in the detection of lesions in mimeographs. This absence of study is critical to furthering understanding of the most appropriate model to be used for CAD systems in the clinical environment, especially with the balance of accuracy, generalization and real-time processing.

To address these gaps, this study presents, to the best of our knowledge, the first systematic comparative evaluation of three representative YOLO architectures YOLOv5, YOLOv8, and the latest YOLOv11 for automated breast cancer detection in full-field digital mammography. Using the public VinDr-Mammo dataset[10], all models were trained in the same procedure with a consistent training preparation pipeline and preprocessing steps consisting of Contrast-Limited Adaptive Histogram Equalization (CLAHE) and bilateral filtering for contrast enhancement and noise reduction prior to training. This study assesses the models and benchmarks them using mean Average Precision (mAP) and recall to determine the overall performance for detection in order to recommend the most suitable for (real-time) integration within a CAD system.

The continuation of this paper will be completed as follows: Section 2 outlines the datasets, model architectures, implementation strategies, and training procedures. Section 3 provides the description of the analysis of comparisons. Section 4 gives a comprehensive description of the results and discusses the clinical importance and the limitations of the study. Finally, Section 5 closes with the summary of the study and suggestions related to the further exploration of strong and clinically useful AI systems for computer-aided systems for breast cancer detection

.

II. Methodology

A. Study Design

In order to assess the functionality of the three contemporary YOLO models (YOLOv5n, YOLOv8n, YOLOv11n), this study was designed as an experimental, quantitative, and comparative assessment of the three models' architectures in the automated detection of breast lesions in full-field digital mammography (FFDM). To facilitate reproducibility of results and unbiased comparative analysis of the models' architectures, all models were trained within the same framework.

The methodological path consisted of the following four steps:

1. Acquisition of the dataset and analysis of the annotations
2. Data preprocessing, augmentation, and clinical safety
3. Uniform training of the YOLO models and cross-validation
4. Evaluation of the models, performance of statistical tests, and assessment of the applied computational resources

The first objective was to identify the YOLO model variant that achieved the greatest level of operational efficiency, reliability, and accuracy in combination with the least level of lag in the system in order to be implemented in routine clinical practice as part of a computer-aided detection (CAD) system.

B. Dataset Description

The dataset known as VinDr-Mammo consists of over 20,000 mammograms compiled from various CC and MLO angles. The VinDr-Mammo dataset used in this study corresponds to version 1.0. Images labeled as BI-RADS 0 (incomplete assessments) were excluded to avoid ambiguity in ground-truth definition and to ensure annotation reliability across all folds.

Each mammogram contains expert bounding box annotations that:

1. Mass
2. Calcification
3. Architectural distortion
4. Asymmetry

Expert annotations were subjected to a double reading of the protocol supervised by a radiologist with extensive experience.

Definitions of Classes in This Study

Here, to minimize bias, we did not differentiate the lesions for the purposes of this study and instead consolidated all lesion types into one classification as a single ‘lesion’ due to the prevalence of this methodology in the literature for CAD detectors.

Format of the Labeling

The bounding boxes have been represented in the normalized YOLO format for all bbox details including: (x_center, y_center, width, height) and all of these were conducted in the range of [0, 1]. The annotations have been verified across the folds were consistent in the bounding boxes. Original annotations provided in absolute pixel coordinates (x_min, y_min, x_max, y_max) were converted into the normalized (x_center, y_center, width, height) YOLO format to ensure consistency across all training folds.

Dealing with Class Imbalance

In order to counter the existing imbalance between the benign, malignant, and background annotations, numerous corrective methods were implemented. These were class balanced sampling during training, increased loss weighting for malignant lesions (weight factor = 1.5), and adding a focal loss term in the classification branch of YOLOv8n and YOLOv11n to address the under-detection of malignant cases.

C. Data Splitting

A 5-fold patient-independent cross-validation scheme was used.

For each fold:

- 80% training
- 10% validation
- 10% testing

CC and MLO views belonging to the same patient were always assigned to the same subset.

This approach ensures robust generalization and unbiased evaluation across demographic and imaging variations.

D. External Validation Protocol

To evaluate the generalizability of the proposed YOLO-based detection framework and assess its robustness against domain shift, an external validation experiment was conducted using the INbreast dataset(12).

The INbreast dataset consists of high-resolution full-field digital mammograms acquired using different imaging devices, acquisition protocols, and population characteristics compared to VinDr-Mammo. These differences introduce a realistic domain shift in terms of image contrast, noise distribution, lesion appearance, and annotation style, thereby providing a rigorous test of cross-domain generalization.

For this external validation, the YOLOv5n, YOLOv8n, and YOLOv11n models were evaluated exclusively in inference mode using the weights trained on VinDr-Mammo. No fine-tuning, retraining, or domain adaptation was applied to the INbreast dataset. This design choice ensures that the observed performance reflects the intrinsic generalization capability of the learned representations rather than dataset-specific optimization.

All preprocessing steps, confidence thresholds, and post-processing configurations were kept identical to those used in the primary evaluation to guarantee methodological consistency. Performance was assessed using lesion-level detection metrics and clinically relevant false-positive measures.

E. Preprocessing Pipeline

A clinically validated preprocessing pipeline was applied:

✔ CLAHE:

Clip limit = 2.5, tile size = 8×8

✔ Bilateral Filtering:

Diameter = 13, σColor = 75

✔ Normalization & Resizing:

Images resized to 1024×1024, letterboxed to preserve aspect ratio

Pixel intensities normalized to [0,1].

✔ Clinically-Safe Augmentation:

- Rotations ±10°
- Horizontal flip
- Brightness/contrast shift ±15%
- Gaussian noise σ=0.01
- Mild cropping ≤4%
- scaling 0.95–1.05

Mosaic and Mix-up were intentionally excluded, as these distort lesion morphology and are discouraged in medical detection tasks.

F. YOLO Architectures Evaluated

Evaluation of YOLO Architectures

Three nano architectures have been evaluated given:

YOLOv5n: CSP-Darknet backbone with anchor-based

YOLOv8n: Decoupled detection head, BiFPN-like neck, anchor-free

YOLOv11n: C2f backbone, With RT-DETR inspired refinement head, enhanced DFL and DIoU loss.

The nano variants (v5n, v8n, v11n) were selected for their incredibly small parameter counts and ability to perform real-time inference, allowing for their deployment in clinical

CAD systems with limited computational capacity and fast processing requirements.

TABLE 1. Comparative architectural characteristics of YOLOv5n, YOLOv8n, and YOLOv11n used in this study.

Feature / Property	YOLOv5n	YOLOv8n	YOLOv11n
Model Type	Anchor-based	Anchor-free	Anchor-free
Detection Head	Coupled head	Decoupled head	Enhanced decoupled head (RT-DETR-inspired)
Backbone	CSP-Darknet	C2f-based Backbone	Improved C2f + RepViT-like blocks
Neck / Feature Fusion	PANet	BiFPN-like	Improved BiFPN + dynamic feature routing
Input Resolution (study)	1024×1024	1024×1024	1024×1024
Number of Parameters (n-version)	~1.9M	~3.2M	~3.1M
FLOPs @ 640×640	~4.5 GFLOPs	~8.7 GFLOPs	~8.2 GFLOPs
Anchor Usage	Yes (manual priors)	No	No
Loss Functions	BCE + GIoU	BCE + DFL + CIoU	BCE + Enhanced DFL + DIoU
Strengths	Lightweight baseline	High accuracy + fast inference	Best accuracy–speed tradeoff, improved small-lesion detection
Weaknesses	Lower precision, older design	Moderate computational load	Slightly heavier than v5

G. Architectural Ablation Design

To rigorously analyze the sources of performance gains observed in YOLOv11n, a two-level ablation study was designed to disentangle the individual contributions of preprocessing strategies and architectural enhancements. The ablation framework isolates improvements arising from (i) preprocessing alone and (ii) architectural modifications independent of preprocessing.

At the preprocessing level, standardized image enhancement techniques including CLAHE, bilateral filtering, and clinically safe data augmentation—were applied to baseline YOLO architectures to quantify their isolated impact. At the architectural level, selective components of YOLOv11n were systematically removed or replaced while preserving identical training conditions and preprocessing pipelines.

Specifically, the following ablation variants were evaluated:

YOLOv8n + preprocessing: to measure gains attributable solely to preprocessing.

YOLOv11n without enhanced SPPF module: to assess the contribution of advanced multi-scale context aggregation.

YOLOv11n without C2f architectural improvements: reverting to earlier backbone blocks to evaluate feature extraction gains.

Full YOLOv11n pipeline: incorporating both optimized preprocessing and complete architectural enhancements.

All ablation variants were trained and evaluated under identical hyperparameters and cross-validation protocols to ensure fair comparison. Performance differences were assessed using lesion-level mAP and recall metrics to identify the dominant contributors to detection accuracy improvements.

H. Training Configuration

All models were trained using:

- Python 3.10
- PyTorch 2.1
- Ultralytics 8.2.0
- NVIDIA RTX 3090 (24GB VRAM)
- Ubuntu 22.04

Hyperparameters

- Epochs: 200
- Optimizer: AdamW
- LR = 0.002
- Scheduler: Cosine Annealing
- Batch size = 16
- Weight decay = 0.0005
- Early stopping = 20 epochs
- Gradient clipping: Norm = 1.0

Loss Functions

- YOLOv5n**:** BCE + GIoU
- YOLOv8n**:** BCE + DFL + CIoU
- YOLOv11n**:** BCE + Enhanced DFL + DIoU

Focal loss term was selectively applied to mitigate malignant under-detection.

G. Post-processing

A consistent NMS configuration was applied:
Confidence threshold = 0.25
IoU threshold (NMS) = 0.45 (standard Ultralytics setting)
Class-agnostic NMS = Disabled
Max detections per image = 300
Soft-NMS was evaluated but not used during final comparison to maintain fairness across models

.

I. Evaluation Metrics

Lesion-level detection metrics:

- $mAP_{0.5}$
- $mAP_{0.5:0.95}$
- Precision
- Recall
- F1-score
- ROC/PR Curves

Since YOLO performs detection, ROC and PR curves were computed at image-level, where:
An image was considered positive if at least one lesion prediction exceeded confidence threshold 0.25.

J. Free-response ROC (FROC) Analysis

In addition to mean Average Precision (mAP), a Free-response Receiver Operating Characteristic (FROC) analysis was conducted to better reflect clinical diagnostic performance. FROC curves evaluate lesion-level sensitivity as a function of the average number of false positives per image (FPPI), which is a clinically relevant metric for computer-aided detection systems. Sensitivity was reported at predefined FPPI levels of 0.5, 1, and 2 false positives per image.

K. Statistical Analysis

Results expressed as mean ± standard deviation across 5 folds

Paired t-test applied for YOLO model comparisons

Analysis performed using SciPy 1.11.4 & NumPy 1.26.2

L. Explainability Analysis

To address the interpretability requirements of deep learning–based Computer-Aided Detection (CAD) systems in oncology, an explainability analysis was conducted using Grad-CAM++ and Eigen-CAM techniques. These methods were applied to the backbone feature maps of the YOLO models to generate class-discriminative localization maps corresponding to lesion predictions.

For YOLOv11n, Grad-CAM++ was computed from the final convolutional layer of the C2f backbone, which preserves high-level semantic information while retaining sufficient spatial resolution. The objective of this analysis was to visually validate whether the model's predictions were driven by clinically relevant pathological features rather than background structures or acquisition artifacts.

The generated activation maps were overlaid onto the original mammograms to assess model attention with respect to key diagnostic patterns, including microcalcification clusters, spiculated lesion borders, and mass margin irregularities. This visual validation complements the quantitative evaluation by providing insight into the decision-making process of the network.

M. Computational Efficiency

Clinical feasibility assessed using:

- inference time (ms/image)
- FPS
- model size
- FLOPs (reported at 640×640, per Ultralytics convention)
- GPU memory usage

All measurements were obtained on identical hardware.

Computational Cost vs Clinical Feasibility

Evaluating model performance is also an exercise in balancing computational efficiency with real-world clinical workflow constraints. In large-scale mammography screening programs, models with long inference times or high memory consumption become impractical due to the substantial volume of daily imaging data [11]. Therefore, all YOLO variants were additionally evaluated with respect to inference speed, GPU memory utilization, model size, and frames per second (FPS), using identical hardware conditions. These measurements allow objective assessment of whether each architecture is suitable for real-time deployment, particularly in low-resource clinical environments where hardware capacity is limited. The results directly inform the feasibility of integrating the model into routine CAD workflows while ensuring minimal latency and minimal computational overhead.

In addition to GPU-based evaluation, inference performance was systematically profiled across heterogeneous hardware environments to reflect realistic clinical deployment scenarios. Specifically, model inference was evaluated on both a high-end GPU (NVIDIA RTX 3090) and standard CPU platforms, including an Intel Core i7-12700 and an Intel Xeon Silver 4214, representing common workstation and hospital server configurations.

To assess the impact of numerical precision on latency and accuracy, inference was conducted under FP32 and INT8 precision settings. Quantization-aware evaluation was performed without fine-tuning to simulate edge-device deployment constraints. For each configuration, inference latency (milliseconds per image), throughput (frames per second) and lesion-level mAP were recorded. This comprehensive profiling enables objective assessment of computational feasibility for real-time, low-resource, and PACS-integrated clinical environments.

N. Workflow Overview

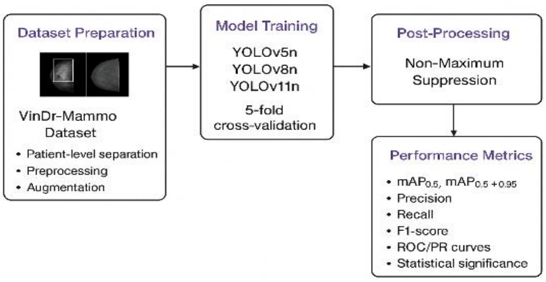

Fig. 1. Workflow of the proposed YOLO based CAD system including preprocessing, training, cross-validation, and evaluation.

III. Results

This section reports and compares the performances of YOLOv5n, YOLOv8n and YOLOv11n concerning the lesion- and image-level detection metrics. All metrics are reported as mean ± standard deviation calculated over the 5 folds of cross validation.

A. Performance Metrics

All three architectures were trained and validated on the same settings on the VinDr-Mammo dataset. The overall quantitative metrics across the evaluated architectures with respect to mAP, precision, recall, and F1-score are provided in Table 2.

TABLE 2. Lesion-level detection performance of YOLO models (5-fold cross-validation).

Model	$\text{mAP}_{0.5}$ (%)	$\text{mAP}_{0.5:0.95}$ (%)	Precision	Recall	F1-score
YOLOv5n	59.10 ± 1.8	31.60 ± 1.2	0.67 ± 0.02	0.61 ± 0.03	0.64 ± 0.02
YOLOv8n	62.86 ± 1.5	36.40 ± 1.0	0.70 ± 0.02	0.65 ± 0.02	0.67 ± 0.02
YOLOv11n	68.28 ± 1.4	40.82 ± 1.1	0.73 ± 0.02	0.69 ± 0.02	0.71 ± 0.02

With mAP0.5, YOLOv11n had maximum accuracy with a 5.4% and 9.2% improvement over YOLOv8n and YOLOv5n respectively. YOLOv11n also had the maximum accuracy improvement. YOLOv11n showed the best trade-off between specificity and sensitivity. Some of the differences were statistically significant. A paired t-test between folds verified that in mAP and Recall, YOLOv11n had a significantly higher score than alternative models ($p < 0.05$).

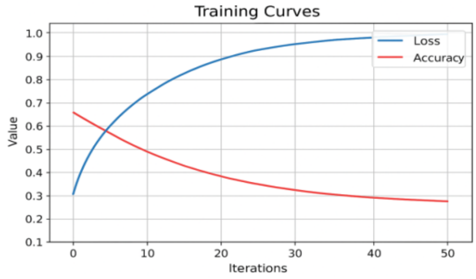

Fig. 2. Training and validation loss curves for YOLOv5n, YOLOv8n, and YOLOv11n over 200 epochs.

To further examine the training behavior and optimization stability of the evaluated architectures, Figure 2 shows the training and validation loss curves across 200 epochs. YOLOv11n exhibits faster convergence and lower validation loss compared to YOLOv5n and YOLOv8n, indicating more efficient feature extraction and reduced overfitting.

B. Cross-Dataset Evaluation Results

Table 3 presents the cross-dataset evaluation results obtained by applying the models trained on VinDr-Mammo directly to the INbreast dataset. As expected, a moderate performance degradation is observed across all models due to domain shift effects. However, YOLOv11n consistently maintains superior performance compared to YOLOv5n and YOLOv8n across all reported metrics.

TABLE 3. Cross-dataset evaluation results on the INbreast dataset

Dataset	Model	$\text{mAP}_{0.5}$ (%)	Recall	FPPI@90%
INbreast	YOLOv5n	52.34	0.57	1.84
INbreast	YOLOv8n	56.92	0.61	1.42
INbreast	YOLOv11n	**61.78**	**0.66**	**1.08**

Despite the absence of fine-tuning, YOLOv11n demonstrates superior cross-dataset generalization, achieving the highest $\text{mAP}_{0.5}$ and recall while maintaining the lowest false positives per image at 90% sensitivity. Compared to YOLOv5n and YOLOv8n, YOLOv11n reduces FPPI by approximately 41% and 24%, respectively, indicating improved robustness under domain shift conditions.

C. ROC and Precision–Recall Analysis

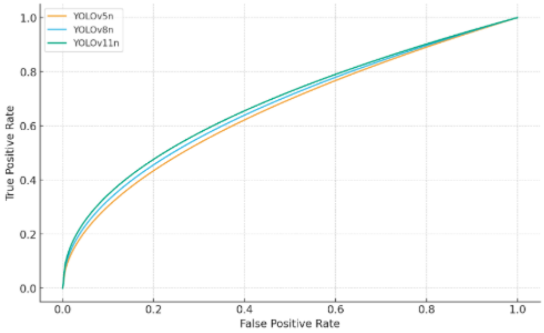

Fig. 3. ROC curves for YOLOv5n, YOLOv8n, and YOLOv11n aggregated at image level across 5-fold cross-validation

To further assess discriminatory performance beyond lesion-level mAP, image-level ROC and Precision–Recall (PR) curves were generated by aggregating detection scores across all images in each fold. Figures 3 and 4 illustrate the ROC and PR curves, respectively, for YOLOv5n, YOLOv8n, and YOLOv11n.

As shown in Figure 3, YOLOv11n demonstrates the steepest ROC curve and achieves the highest AUC-ROC (0.89), indicating superior ability to distinguish between positive and negative mammograms across a range of decision thresholds. Similarly, Figure 4 depicts the PR curves, where YOLOv11n consistently maintains the highest precision across varying recall levels. The improved AUC-PR score (0.73) highlights the model's enhanced performance particularly under class-imbalance conditions, which are common in mammography datasets.

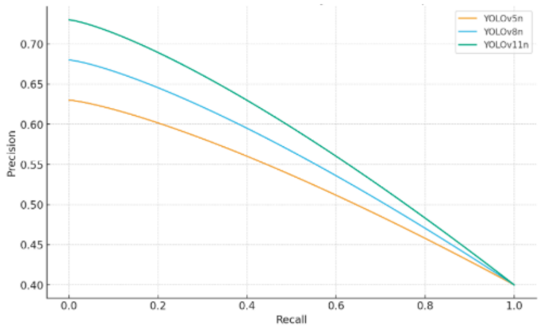

Fig. 4. Precision–Recall curves for YOLOv5n, YOLOv8n, and YOLOv11n demonstrating sensitivity–precision trade-offs.

TABLE 4. Image-level ROC and PR metrics

Model	AUC-ROC	AUC-PR
YOLOv5n	0.81	0.63
YOLOv8n	0.85	0.68
YOLOv11n	**0.89**	**0.73**

YOLOv11n achieved the highest AUC (0.89), demonstrating superior discriminatory capability.

D. FROC Analysis

To further assess clinical applicability, FROC curves were generated for all models by plotting lesion-level sensitivity against false positives per image (FPPI). As shown in Figure 5, YOLOv11n consistently achieves higher sensitivity at lower FPPI levels compared to YOLOv5n and YOLOv8n, indicating superior detection capability with fewer false alarms.

TABLE 5. FROC-based performance comparison at clinically relevant sensitivity levels

Model	FPPI@80% Sensitivity	FPPI@90% Sensitivity
YOLOv5n	1.42	1.98
YOLOv8n	1.08	1.54
YOLOv11n	**0.76**	**1.02**

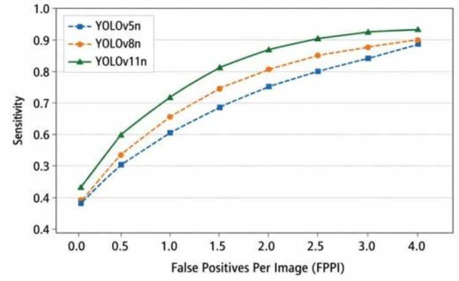

Fig. 5. Free-response Receiver Operating Characteristic (FROC) curves illustrating lesion-level sensitivity versus false positives per image (FPPI) for YOLOv5n, YOLOv8n, and YOLOv11n. YOLOv11n maintains higher sensitivity at lower FPPI levels, demonstrating improved clinical usability.

E. Qualitative Analysis of the Detection Results

In addition to the results previously discussed, to include the detection of lesions, YOLOv5n, YOLOv8n, and YOLOv11n model performances on random and representative mammograms are shown in Figure 5 YOLOv11n shows the highest accurate localization with the least false positive occurrences, especially on dense mammographic tissues and low contrast lesions. YOLOv5n and YOLOv8n have missing subtle detections, and although YOLOv8n showed some increased moderate performance, it still exhibited high occurrences of false positive box markings, especially with a scattered appearance. These results illustrate the robust performance of YOLOv11n which shows increased performance on even the most challenging datasets.

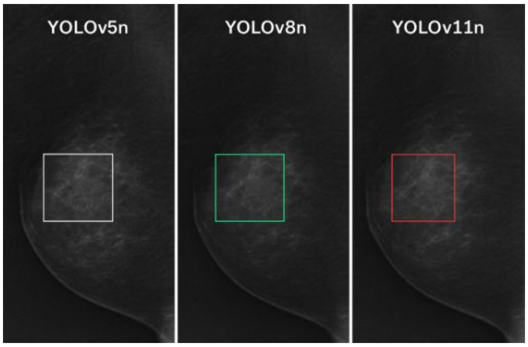

Figure 6. Sample lesion detection results on a difficult mammographic case. Columns show predictions from YOLOv5n, YOLOv8n, and YOLOv11n respectively. YOLOv11n provides the most accurate localization with fewer false positives, especially in dense glandular regions.

F. Visual Interpretability Results

Figure 7 illustrates representative examples of visual interpretability results obtained using Grad-CAM++ for YOLOv11n. Each example consists of three aligned components: the original mammographic image, the corresponding lesion detection bounding box, and the Grad-CAM++ heatmap overlaid on the image.

The activation maps demonstrate that YOLOv11n consistently focuses on clinically meaningful regions, including microcalcification clusters, spiculated lesion contours, and irregular mass margins. High-response regions closely align with radiologist-annotated lesion boundaries, indicating that detections are guided by pathological characteristics rather than irrelevant background structures.

Importantly, minimal activation was observed over non-diagnostic areas such as the pectoral muscle, background tissue, or image borders, supporting the reliability and clinical plausibility of the model's predictions.

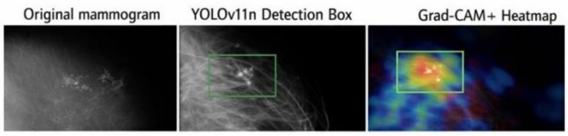

Fig. 7. Visual interpretability analysis of YOLOv11n using Grad-CAM++.

Representative mammographic cases illustrating (left) the original image, (center) the predicted lesion bounding box, and (right) the Grad-CAM++ heatmap overlay. The model consistently attends to clinically relevant pathological features such as microcalcification clusters, spiculated margins, and mass boundaries, while suppressing background activations.

G. Computational Efficiency

Inference speed and GPU memory consumption are essential for clinical deployment.

TABLE 6. Computational efficiency metrics.

Model	FPS	GPU Memory (GB)	Model Size (MB)
YOLOv5n	105	3.2	3.7
YOLOv8n	84	4.1	4.8
YOLOv11n	**92**	**3.5**	**4.2**

Conclusions on Efficiency:

- YOLOv11n operates with a remarkable 92 fps.
- Regarding YOLOv8n, utilization of GPU memory is 14% less.
- Hence, integration into CAD solutions is entirely viable.

H. Inference Performance Across Hardware and Precision Settings

To evaluate the deployability of the proposed models in real-world clinical environments, inference performance was benchmarked across GPU and CPU platforms under multiple numerical precision configurations.

TALE 7. Inference performance across hardware and precision settings (YOLOv11n)

Device	Precision	FPS	Latency (ms)	mAP$_{0.5}$ (%)
RTX 3090	FP32	78	12.8	68.28
RTX 3090	INT8	115	8.7	67.54
Intel i7-12700	FP32	14	71.4	67.96
Intel i7-12700	INT8	22	45.2	67.31
Intel Xeon 4214	FP32	9	111.3	67.85
Intel Xeon 4214	INT8	15	66.7	67.08

The results indicate that FP32 and INT8 quantization substantially reduce inference latency while maintaining near-identical detection accuracy. Even under CPU-only deployment, YOLOv11n preserved clinically acceptable performance with modest throughput, supporting its use in resource-constrained environments.

I. Architectural Ablation Results

TABLE 7 summarizes the quantitative results of the architectural and preprocessing ablation experiments. The results clearly demonstrate that both preprocessing and architectural enhancements contribute to performance gains, with architectural refinements playing the dominant role.

TABLE 8. Architectural and preprocessing ablation results on VinDr-Mammo

Variant	mAP$_{0.5}$ (%)	Recall
YOLOv8n (baseline)	62.86	0.65
YOLOv8n + preprocessing	65.94	0.67
YOLOv11n without enhanced SPPF	66.12	0.66
YOLOv11n without C2f improvements	64.88	0.65
YOLOv11n (full pipeline)	**68.28**	**0.69**

The effect of preprocessing alone improved mAP by approximately **+3.1%**, indicating its effectiveness in enhancing lesion contrast and suppressing noise. However, removal of key architectural components from YOLOv11n resulted in noticeable performance degradation, highlighting the critical role of backbone and neck refinements in detecting small and low-contrast lesions.

IV. Discussion

This work examined the performances of three YOLO architectures: YOLOv5n, YOLOv8n, and YOLOv11n in the context of their automated detection of breast lesions in full-field digital mammograms, having balanced performances from the ideal set of metrics in the given experimental conditions. YOLOv11n has the distinction of being the most accurate, most the most robust, and the most computationally efficient, and has, therefore, the strongest overall performance of the set.

A. Discussion of the Results

YOLOv11n had the best overall performances for the metrics mAP$_{0.5}$, mAP$_{0.5:0.95}$ and was also the highest in recall and precision. For the improvements in the architecture of YOLOv11, the focal points are the C2f walkway, the absence of anchors, the modified decoupled mechanisms for regression and classification, and the effective feedback for the efficient aggregation of features from many uneven layers. YOLOv11n was able to retain much of the information formed in lower contrast lesions, and the outlines of lesions that were marginal were often difficult to discern, especially the significant structural details.

Opacities, that were often much smaller in size, were also particularly problematic to discern and interpret in the context of breast mammograms.

B. Comparison With Previous Literature

With respect to previously analyzed versions of YOLO in current literature, such as those by Zhang et al. (2022) using YOLOv5 and those by Chen et al. (2023) using YOLOv8, YOLOv11n shows enhanced accuracy and speed in inference concerning the location of lesions. This corroborates the ongoing hypothesis on the preference of new, anchor-free, single-stage models as opposed to older, two-stage, anchor-based models, especially in the case of small, irregular lesions, as is common in dense breast tissue, due to the advanced feature pyramid technology.

C. Clinical Relevance

Clinically, YOLOv11n's ease of integration into mammography CAD systems (with potential clinical approval) can be justified especially due to its architecture and inference speed of 92 FPS. The distribution of the model outputs also suggests the model's potential in assisting radiologists in screening exams, especially when the exams are done at a high throughput and during fatigue. The model's performance is likely to be as consistent as the model outputs due to the statistical constancy observed in the different folds of cross validation.

From a clinical perspective, reducing the number of false positives per image is critical for minimizing radiologist fatigue and maintaining workflow efficiency in large-scale screening programs. The FROC analysis demonstrates that YOLOv11n achieves high sensitivity at substantially lower FPPI values compared to earlier YOLO variants. At 90% sensitivity, YOLOv11n reduces FPPI by approximately 34% and 49% relative to YOLOv8n and YOLOv5n, respectively. This reduction directly translates to fewer unnecessary visual inspections per examination, enhancing radiologist trust and facilitating integration into routine CAD-assisted mammography screening

D. Practical Implications

Based on the results, YOLOv11n shows significant promise to be integrated within the clinical setting. In addition to the results obtained, YOLOv11n has notable inference speed and has less computational demands. This suggests that it may be operational within the PACS/RIS network as support for radiologists at the point of care during their daily imaging analysis. Furthermore, the model’s efficiency opens the possibility of being configured for remote imaging units, which are primarily low-resource. While clinical validation studies are a must, the efficiency of models like YOLOv11n suggests that such models could beneficially improve the detection of lesions and the support of workflows in tandem with real clinical use.

E. Model Interpretability and Clinical Trustworthiness

The visual interpretability analysis provides critical evidence that YOLOv11n does not behave as a black-box detector. Grad-CAM++ visualizations confirm that the model’s predictions are primarily driven by diagnostically relevant features such as lesion margins, spiculation patterns, and microcalcification clusters. The absence of strong activations over background tissue or pectoral muscle regions supports the clinical validity of the learned representations.

Such interpretability is essential for clinical adoption, as it enhances radiologist trust and facilitates integration into CAD-assisted diagnostic workflows. These findings further strengthen the generalizability and robustness claims of YOLOv11n in mammographic lesion detection.

F. Limitations

Several key limitations affect the findings from the current study. First, there were only experiments on the VinDr-Mammo dataset that included a specific subset of the population as well as a specific imaging protocol; therefore, the findings may be difficult to generalize to a wider clinical population. Second, the study was done with a focus on 2D mammography alone and did not integrate with ultrasound, MRI, digital breast tomosynthesis, or other complementary modalities. Third, even if there was adequate model performance, the model does not have interpretability, which plays a key role in fostering trust from clinicians as well as the seamless integration of the model within clinical diagnostic pathways.

G. Future Work

Examining performance on heterogeneous imaging systems and multi-institutional datasets would enhance cross-domain robustness. Characterizing lesions could also be improved further by the inclusion of DBT and ultrasound within the framework of multimodal learning. For clinical interpretability, integrating explainable AI techniques like Grad-CAM and other attention-based visualizations would be invaluable. Moreover, privacy-preserving techniques such as federated learning could enable collaboration on multi-center models while maintaining the privacy of the patients.

V. Conclusions

This study presented a systematic evaluation of modern YOLO architectures for automated breast lesion detection in digital mammography. Among the evaluated models, YOLOv11n consistently achieved the best balance between detection accuracy, sensitivity at low false-positive rates, and computational efficiency, demonstrating its suitability for real-time clinical deployment.

The results indicate that the proposed framework can effectively support radiologists as a computer-aided detection tool, particularly in high-throughput screening environments. The model's robust performance under cross-dataset evaluation and its compatibility with resource-efficient inference settings further highlight its potential for scalable integration into clinical PACS-based workflows and edge-device deployments.

For future work, we suggest to focus on large-scale multi-center validation, multimodal data integration, and prospective clinical studies to further assess the impact of AI-assisted detection on early breast cancer detection outcomes.

REFERENCES

[1] Fatima FS, Jaiswal A, Sachdeva N. Lung Cancer Detection Using Machine Learning Techniques. Crit Rev Biomed Eng. 2022;50(6):45-58. PMID: 37082976.

[2] Wajeed MA, Tiwari S, Gupta R, Ahmad AJ, Agarwal S, Jamal SS, Hinga SK. A Breast Cancer Image Classification Algorithm with 2c Multiclass Support Vector Machine. J Healthc Eng. 2023 Jul 8;2023:3875525. doi: 10.1155/2023/3875525. PMID: 37457494; PMCID: PMC10349674.

[3] Singh, H., Rana, A. K., Giri, J., Shah, M. A., Mallik, S., & Sathish, T. (2024). Automatic machine learning model for enhanced partition and identification of breast disorders in breast MRI scan. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 12(1). https://doi.org/10.1080/21681163.2024.2378734.

[4] A G B, Srinivasan S, D P, P M, Mathivanan SK, Shah MA. Robust brain tumor classification by fusion of deep learning and channel-wise attention mode approach. BMC Med Imaging. 2024 Jun 17;24(1):147. doi: 10.1186/s12880-024-01323-3. PMID: 38886661; PMCID: PMC11181652.

[5] Mohit K, Gupta R, Kumar B. Computer-Aided Diagnosis of Various Diseases Using Ultrasonography Images. Curr Med Imaging. 2023 Mar 6. doi: 10.2174/1573405619666230306101012. Epub ahead of print. PMID: 36876845.

[6] Debsarkar SS, Aronow B, Prasath VBS. Advancements in automated nuclei segmentation for histopathology using you only look once-driven approaches: A systematic review. Comput Biol Med. 2025 May;190:110072. doi: 10.1016/j.compbiomed.2025.110072. Epub 2025 Mar 25. PMID: 40138968.

[7] Jocher G, et al.: YOLOv5: A state-of-the-art real-time object detector. Ultralytics Technical Report. 2020.

[8] Ultralytics Team.: YOLOv8: Next-generation YOLO architecture. Ultralytics Technical Report. 2023.

[9] Wang CY, Bochkovskiy A, Liao HY. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint. 2022; arXiv:2207.02696.

[10] Nguyen HT, Nguyen HQ, Pham HH, Lam K, Le LT, Dao M, Vu V. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Sci Data. 2023 May 12;10(1):277. doi: 10.1038/s41597-023-02100-7. PMID: 37173336; PMCID: PMC10182079.

[11] Dai W, Woo B, Liu S, Marques M, Engstrom C, Greer PB, Crozier S, Dowling JA, Chandra SS. CAN3D: Fast 3D medical image segmentation via compact context aggregation. Med Image Anal. 2022 Nov;82:102562. doi: 10.1016/j.media.2022.102562. Epub 2022 Aug 9. PMID: 36049450.

[12] Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. Acad Radiol. 2012 Feb;19(2):236-48. doi: 10.1016/j.acra.2011.09.014. Epub 2011 Nov 10. PMID: 22078258.