A YOLO-based Model for Breast Calcification Areas Detection in Screening Mammography

Yun-Ting Chang[a] and Chiao-Min Chen*[a]

[a]Department of Mathematics, National Changhua University of Education, No.1, Jin-De Road, Changhua, Taiwan 500, R.O.C.

ABSTRACT

Breast cancer remains one of the most prevalent and life-threatening diseases among women worldwide. Early diagnosis of breast cancer is pivotal in improving patient outcomes and survival rates. The earliest signs of non-palpable breast cancer are calcifications. This paper proposes a deep learning network for breast calcification areas detection based on YOLO with self-attention mechanism. By using Bi-Level Routing Attention (BRA) mechanisms, the model's performance can be significantly enhanced. Later, the modified Bi-directional Feature Pyramid Network (BiFPN) technique was used. The advanced model architecture is a modification of the YOLOv8 framework. In order to improve the instances detection of breast calcification, we applied several image preprocessing steps. The contrast of each input image was enhanced and standardized, and the images were resized to a fixed resolution. Utilizing k-fold cross-validation, multiple supervised machine learning techniques were compared. The model demonstrated effective performance across various metrics in the task of calcification detection, achieving a precision rate of 99.32%, a recall rate of 85.0% and an F1-score of 91.59% at the IoU threshold of 0.6. Based on these experimental results, the model is shown to reliably detect areas of breast calcification.

Keywords: Mammography, deep learning, objects detection, self-attention, YOLO-based

1. INTRODUCTION

Breast cancer is a significant global health concern that affects millions of women and their families each year. In recent years, there has been a trend toward an earlier age of onset for breast cancer. When breast cancer is detected early and is in the localized stage, the 5-year relative survival rate is higher than 99%.[1] This underscores the critical importance of early detection of breast cancer, especially for the earliest signs of non-palpable breast cancer, which often manifest as calcification.[2] The development of a detection model for breast calcification areas can assist physicians in the early diagnosis of breast cancer, thereby improving clinical decision-making and patient care.

This study proposes a novel multi-scale model upon YOLOv8 embedding a self-attention guidance mechanism for breast calcification detection. In assessing the effectiveness of the proposed framework, we utilized the recognized benchmark dataset CBIS-DDSM,[3] which is tailored to evaluate distinct aspects of the framework's performance in classification and detection, especially in the task of Breast Cancer Detection (BCD). Experiments show that the model can detect breast calcification more efficiently while satisfying the real-time requirements of medical inspection.

As outlined below, this study makes several key contributions to the instance detection of breast calcification based on machine learning and deep learning architectures. The specific work in this study is as follows:
1) Employing hyperparameter tuning to optimize each machine learning framework and enhance performance.
2) Utilizing a primary dataset for framework evaluation.
3) Demonstrating that the proposed model achieved an accuracy of 99.32% and an F1-score of 91.59%, surpassing other frameworks.

Further author information: (Send correspondence to Chiao-Min Chen)
Yun-Ting Chang: E-mail: changyustina@gmail.com
Chiao-Min Chen: E-mail: cmchen@cc.ncue.edu.tw, Telephone: +886-4-7232105 ext. 3223

2. MATERIALS AND METHODS

2.1 Datasets

The 'Curated Breast Imaging Subset of the Digital Database for Screening Mammography' (CBIS-DDSM)[3] is considered a well-established benchmark in digital mammogram-based breast cancer screening, comprising 3,103 mammographic images sourced from 1,566 patients. This dataset includes both mediolateral oblique (MLO) and craniocaudal (CC) views for each breast.

A subset of 1,806 craniocaudal (CC) images focused on calcification areas was used to evaluate the proposed methodology. The distribution of the images is as follows: 1,514 images pertain to single tumors, while 292 images correspond to multi-tumor cases. The dataset was partitioned into three subsets: 1,280 images for training, 320 images for validation, and 206 images for testing.

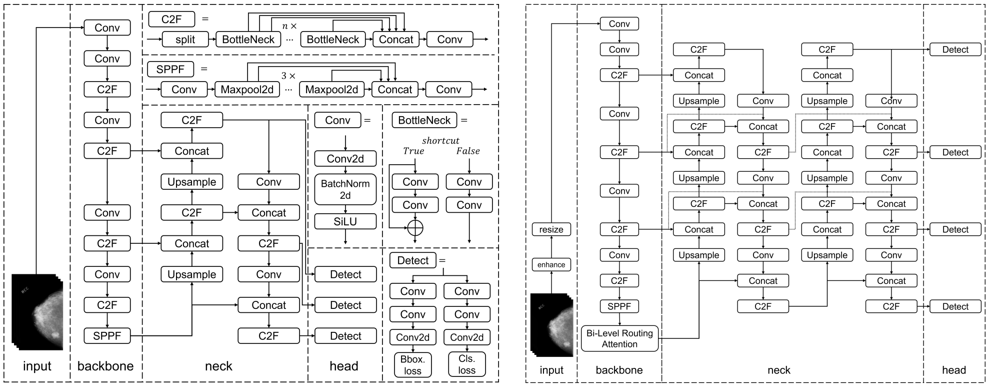
We customized the CBIS-DDSM for our purposes, manually annotating the dataset with LabelImg, an open-source image annotation tool. Data labeling adheres to a uniform standard. During annotation, the bounding box coordinates were created. The YOLO format was used to save annotations as text files (.txt). Each image file corresponds to a label file. Every line in the label file contains five numbers that represent an instance. The five numbers respectively represent the category of the instance, the abscissa of the center point, the ordinate of the center point, width, and height.

2.2 Image Preprocessing

To enhance image quality for analytical purposes, the contrast of each input image was enhanced and subsequently standardized, ensuring consistency across samples. In order to optimize the trade-off between training and inference speed, as well as memory efficiency, all images were resized to a fixed resolution of 640×640 pixels. This approach maintained uniformity across the dataset and supported reliable processing performance.

2.3 The Proposed Network

In this study, we propose a stacked ensemble of models to diagnose detected breast calcification. The base model comes from the YOLOv8 architecture and its variations. The primary contributions of our work are summarized as follows:
1) Utilizing the lightweight as YOLOv8 to reduce the computational resources required for detecting large images, thereby enhancing the system's operational efficiency and real-time performance.
2) To strengthen the detection capability for small targets, a small objects detection layer is added and the feature fusion network is improved following the BiFPN link idea.
3) BRA module is introduced to improve the ability to capture multi-scale contextual information effectively.



(a) YOLOv8

(b) The proposed network

Figure 1. The overall architecture: (a) YOLOv8 integrates the CSPNet backbone and an enhanced FPN+PAN neck, divided into four parts: Input, Backbone, Neck, and Head, along with detailed descriptions of core modules and mechanisms. (b) The proposed network is a multi-scale model based on YOLOv8 with self-attention mechanism.

2.4 Multi-Level Feature Fusion Networks

The backbone of our proposed network extracts feature maps from the raw input images at scales $P_2 - P_5$, with the added small object detection layer P_2 .

YOLOv8 employs a dual approach with a top-down Feature Pyramid Network (FPN)[4] and a bottom-up Pyramid Attention Network (PANet)[5] to effectively combine both shallow and deep feature information. In our work, we introduce the Bi-directional Feature Pyramid Network (BiFPN),[6] which features an additional pathway connecting input and output feature matrices that have the same dimensions. This extra route facilitates the integration of fundamental feature data from shallow layers without significantly increasing computational costs. As a result, the BiFPN improves the accuracy of detecting small objects. The structuresof of the FPN, PAN and BiFPN network is depicted in Figure 2.

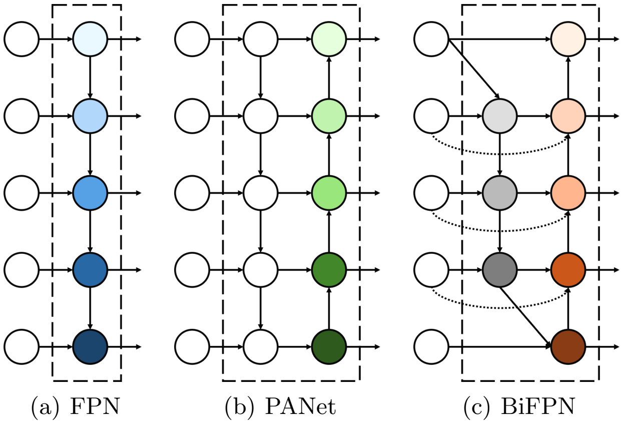


Figure 2. Structures of feature fusion networks: (a) FPN combines multi-scale features with a top-down pathway; (b) PANet introduces an additional bottom-up pathway on top of FPN; (c) BiFPN simplifies the architecture by removing nodes with a single input edge and adding shortcut connections between input and output nodes at the same level.

2.5 Multi-Scale Features Mechanism

Adopting a single-stage approach, in contrast to a two-stage framework, complicates the recognition of fine-grained features, thereby exacerbating the complexity of fittings detection. We incorporated the Bi-Level Routing Attention (BRA)[7] mechanisms to address this issue. BRA is a dynamic sparse attention mechanism that enables efficient information processing by selectively focusing on relevant features across multiple scales. Figure 3 depicts the architecture of the BRA mechanism. The detailed explanation of BRA is as follows:

Step 1, given a 2D input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ divided into $S \times S$ non-overlapped regions and reshaped $\mathbf{X}\ as\ \mathbf{X}^r \in \mathbb{R}^{S^2 \times \frac{HW}{S^2} \times C}$ such that each region contains $\frac{HW}{S^2}$ feature vectors. The query, key, and value tensor, $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V} \in \mathbb{R}^{S^2 \times \frac{HW}{S^2} \times C}$ are derived through linear projections:

$$\mathrm{Q} = \mathrm{X}^r \mathrm{W}^q,\ \mathrm{K} = \mathrm{X}^r \mathrm{W}^k,\ \mathrm{V} = \mathrm{X}^r \mathrm{W}^v, \tag{1}$$

where $\mathbf{W}^q$, $\mathbf{W}^k$, $\mathbf{W}^v \in \mathbb{R}^{C \times C}$ are projection weights for the query, key, value, respectively.

Step 2, averaged $\mathbf{Q}$ and $\mathbf{K}$ for each region to obtain region-level queries and keys $\mathbf{Q}^r$, $\mathbf{K}^r \in \mathbb{R}^{S^2 \times C}$. Then, derives the adjacency matrix $\mathbf{A}^r \in \mathbb{R}^{S^2 \times S^2}$ of the region-to-region affinity graph via matrix multiplication between $\mathbf{Q}^r$ and transposed $\mathbf{K}^r$:

$$\mathrm{A}^r = \mathrm{Q}^r (\mathrm{K}^r)^T. \tag{2}$$

Step 3, prune the affinity graph by keeping only top-k connections for each region. We derive a routing index matrix $\mathbf{I}_r \in \mathbb{N}^{S^2 \times k}$, with the row-wise top-$k$ operator:

$$\mathrm{I}^r = topkIndex(\mathrm{A}^r), \tag{3}$$

such that the i^{th} row of $\mathbf{I}^r$ contains k indices of most relevant regions for the i^{th} region.

Step 4, compute fine-grained token-to-token attention with the routing index matrix $\mathbf{I}^r$. All key-value pairs residing in the union of k routed regions indexed with $\mathbf{I}^r_{(i,n)}, n = 1, 2, \ldots, k$ for each query token in region i.

To address the challenge that modern GPUs rely on coalesced memory operations that load blocks of dozens of contiguous bytes at once, we first gather the key and value tensors, i.e.,

$$\mathbf{K}^g = \text{gather}(\mathbf{K}, \mathbf{I}^r), \quad \mathbf{V}^g = \text{gather}(\mathbf{V}, \mathbf{I}^r), \tag{4}$$

where $\mathbf{K}^g, \mathbf{V}^g \in \mathbb{R}^{S^2 \times \frac{kHW}{S^2} \times C}$ are gathered key and value tensor. Then, apply attention on the gathered key-value pairs as:

$$\mathbf{O} = \text{Attention}(\mathbf{Q}, \mathbf{K}^g, \mathbf{V}^g) + \text{LCE}(\mathbf{V}). \tag{5}$$

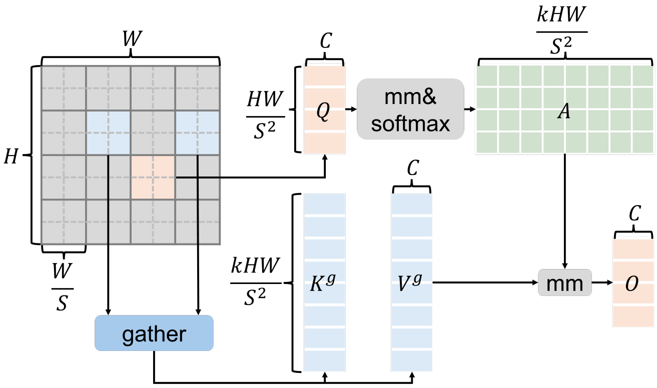


Figure 3. The bi-level routing attention structure. The mechanism combines key-value pairs and uses sparse operations to avoid calculations in less relevant areas, leading to reductions in both parameters and computational resource usage.

3. EXPERIMENT AND ANALYSIS

3.1 Experimental Details

All experiments in this paper were conducted on a high-performance DL workstation with the following overall configuration: CPU: Intel Core i5-10400 CPU @ 2.90GHz × 12; GPU: NVIDIA GeForce RTX 3090; Memory: 32GB; Framework: Pytorch YOLO; Operating system: Ubuntu 20.04.6 LTS. In the experiment, we used a unified training parameter setting, including input image size of 640×640, BatchSize of 4, an initial learning rate of 0.001, the final learning rate at the last epoch for training of 0.0005 and optimized with Adam optimizer. To ensure the stability and reliability of the model, we trained each model for 300 epochs using 5-fold cross-validation.

3.2 Evaluation Metrics

The performance of YOLO object detectors in object detection task was evaluated with several metrics,[8] including Precision, Recall, F1-score, and mean average precision (mAP, i.e., mAP@0.5 and mAP@[0.5:0.95]).

4. RESULTS AND DISCUSSION

4.1 K-fold Cross-validation

Experiments are conducted to evaluate machine learning models and validate their generalization ability on independent test datasets. In this K-fold cross-validation experiment, samples were divided into five folds. At each fold, the dataset was partitioned into three subsets: 1,280 images for training, 320 images for validation, and the same 206 images for testing in every fold. The abnormalities (i.e., calcifications) detection performance throughout the 5-fold cross-validation is reported in Table 1.

The results show the robustness of our proposed network on detecting the calcification position in mammograms with an overall precision of 98.7%. The highest performance achieved in the k-fold cross-validation was a precision rate of 100.0%, a recall rate of 87.8%, a mAP@0.5 of 92.1%, a mAP@[0.5:0.95] of 82.0%, and an F1-score of 93.11%. The lower mAP@[0.5:0.95] values than mAP@0.5 is because the former used higher IoU thresholds (implying more stringent criteria) for the AP calculation.

Table 1. Result of the proposed network throughout the k-fold cross validation. (K=5)

Fold Number	Precision	Recall	mAP50	mAP50-95	F1 Score
1	**1.0000**	0.8140	0.8730	0.7570	0.8975
2	0.9870	0.8550	0.8970	0.7780	0.9163
3	0.9910	**0.8780**	**0.9210**	**0.8200**	**0.9311**
4	**1.0000**	0.8620	0.9130	0.8050	0.9259
5	0.9880	0.8410	0.8890	0.7760	0.9086
Average Value	0.9932	0.8500	0.8986	0.7872	0.9159
Standard Deviation	0.0064	0.0241	0.0191	0.0251	0.0135

4.2 Contrastive Experiment

For comparison, all models listed below are modified versions based on the YOLOv8 architecture. The proposed network is a YOLOv8-based model enhanced with BiFPN and BRA. Both YOLOv8 and YOLOv10[9] serve as the original versions. The backbone of YOLOv8 has been replaced with the Swin Transformer,[10] as well as with ConvNeXt V2[11] in separate implementations.

We summarize the performance of those models as an average of the 5-fold cross-validation results in terms of evaluation metrics. Table 2 summarizes the performance results of all the YOLO detection models. On the whole, the accuracy of all models in detecting calcification classes varies considerably. The F1-score values ranged from 0.14 % obtained by YOLOv8 with a ConvNeXtv2 backbone to 91.59 % by the proposed network. In terms of the proposed network, which achieves a precision rate of 99.32%, a recall rate of 85.0%, a mAP@0.5 of 89.86%, a mAP@[0.5:0.95] of 78.72%, and an F1-score of 91.59%, all the metrics values outperform other models.

Table 2. Average performance comparison across several network with multi-level features BiFPN.

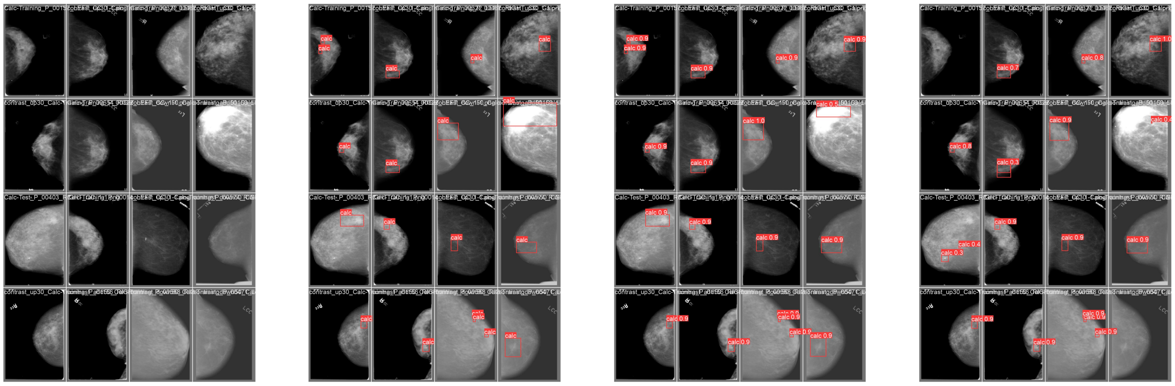
Backbone	Attention	Precision	Recall	mAP50	mAP50-95	F1 Score
YOLOv8	**BRA**	**0.9932**	**0.8500**	**0.8986**	**0.7872**	**0.9159**
YOLOv8	✗	0.9316	0.7906	0.8408	0.7252	0.8552
YOLOv10	✗	0.7154	0.4714	0.5416	0.2532	0.5679
SwinTransformer	✗	0.4492	0.3396	0.3446	0.2434	0.3834
ConvNeXtv2	✗	0.0007	0.0489	0.0012	0.0004	0.0014

4.3 Instances Detection

YOLO[12] generates confidence probability for each potential ROI (indicated as a box) that represents the calcifications position. In Figure 4, true detection cases represent those have the detected boxes with confidence probability higher than 0.6 or IOU less than 0.2. These detection results are achieved by comparing the IOU of each detected boxes with the ground truth.

5. CONCLUSIONS

To address the problem of breast calcification detection in mammographic images, this study designed a deep learning target detection algorithm based on the YOLO architecture, enhanced with a self-attention guidance mechanism. Building upon the Bi-Level Routing Attention mechanisms, an adaptive channel attention BRA module was incorporated. The capability for shallow feature extraction is improved by adjusting the number of detection output layers and incorporating the BiFPN architecture. Utilizing images from the CBIS-DDSM dataset for calcifications in mammographic images, the results demonstrate that the proposed network can effectively improve accuracy, recall rate, and stability, while also accelerating the training process. The model is able to detect breast calcifications more efficiently while meeting the real-time requirements of medical inspection.



(a) Inputs

(b) Inputs with ground truth

(c) Prediction by the proposed networks

(d) Prediction by YOLOv8

Figure 4. Examples of mammographic images with predicted bounding boxes.

REFERENCES

[1] Henson, D. E., Ries, L., Freedman, L. S., and Carriaga, M., "Relationship among outcome, stage of disease, and histologic grade for 22,616 cases of breast cancer. the basis for a prognostic index," *Cancer* **68**(10), 2142–2149 (1991).

[2] Bonfiglio, R., Granaglia, A., Giocondo, R., Scimeca, M., and Bonanno, E., "Molecular aspects and prognostic significance of microcalcifications in human pathology: a narrative review," *International Journal of Molecular Sciences* **22**(1), 120 (2020).

[3] Sawyer-Lee, R., Gimenez, F., Hoogi, A., and Rubin, D., "Curated breast imaging subset of digital database for screening mammography (cbis-ddsm)," (2016). Data set.

[4] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S., "Feature pyramid networks for object detection," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 2117–2125 (2017).

[5] Wang, K., Liew, J. H., Zou, Y., Zhou, D., and Feng, J., "Panet: Few-shot image semantic segmentation with prototype alignment," in [*proceedings of the IEEE/CVF international conference on computer vision*], 9197–9206 (2019).

[6] Tan, M., Pang, R., and Le, Q. V., "Efficientdet: Scalable and efficient object detection," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 10781–10790 (2020).

[7] Zhu, L., Wang, X., Ke, Z., Zhang, W., and Lau, R. W., "Biformer: Vision transformer with bi-level routing attention," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 10323–10333 (2023).

[8] Hossin, M. and Sulaiman, M. N., "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process* **5**(2), 1 (2015).

[9] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., and Ding, G., "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458* (2024).

[10] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 10012–10022 (2021).

[11] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., and Xie, S., "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 16133–16142 (2023).

[12] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., "You only look once: Unified, real-time object detection," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 779–788 (2016).