# Mammo-Clustering: A Weakly Supervised Multi-view Global-Local Context Clustering Network for Detection and Classification in Mammography

Shilong Yang, Chulong Zhang, Qi Zang, Juan Yu, Liang Zeng, Xiao Luo, Yexuan Xing, Xin Pan, Qi Li, Xiaokun Liang, and Yaoqin Xie

Abstract—Breast cancer has long posed a significant threat to women's health, making early screening crucial for mitigating its impact. However, mammography, the preferred method for early screening, faces limitations such as the burden of double reading by radiologists, challenges in widespread adoption in remote and underdeveloped areas, and obstacles in intelligent early screening development due to data constraints. To address these challenges, we propose a weakly supervised multi-view mammography early screening model for breast cancer based on context clustering. Context clustering, a feature extraction structure that is neither CNN nor transformer, combined with multi-view learning for information complementation, presents a promising approach. The weak supervision design specifically addresses data limitations. Our model achieves state-of-the-art performance with fewer parameters on two public datasets, with an AUC of 0.828 on the Vindr-Mammo dataset and 0.805 on the CBIS-DDSM dataset. Our model shows potential in reducing the burden on doctors and increasing the feasibility of breast cancer screening for women in underdeveloped regions.

Index Terms—Artificial intelligence, Breast cancer, Deep Learning, Mammography, Medical imaging.

### I. INTRODUCTION

**B**REAST cancer remains the most prevalent malignant tumor among women worldwide [1]. In today's global context, annually, more than several million women are diagnosed with breast cancer, which constitutes approximately 25% of all cancer cases among women [2]. Recent studies have highlighted that breast cancer has surpassed cardiovascular diseases as a leading cause of premature mortality globally [3] [4]. However, breast cancer is also one of the malignancies for which prevention and treatment strategies are clearly effective and efficacious [5].

Early detection is vital for reducing breast cancer mortality rates [6] [7]. Early detection can also adopt less aggressive treatment plans to improve breast retention rates and reduce

Shilong Yang, Chulong Zhang, Xiaokun Liang, and Yaoqin Xie are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, 518055.

Qi Zang is with the Qingdao University, Qingdao, China, 266000.

Juan Yu, Liang Zeng, Xiao Luo, Yexuan Xing, Xin Pan, Qi Li are with the Department of Radiology, The First Affiliated Hospital of Shenzhen University, Health Science Center, Shenzhen Second People's Hospital, 3002 SunGangXi Road, Shenzhen, 518035, China.

These authors contributions are equal: Shilong Yang and Chulong zhang corresponding authors: Xiaokun Liang and Yaoqin Xie (e-mail: xk.liang@siat.ac.cn and yq.xie@siat.ac.cn)

the physical and psychological burden on patients. Early detection can also allow for less aggressive treatment options, reducing the physical and psychological burden on patients [8]. Additionally, it can help in identifying high-risk individuals who may benefit from preventive measures, ultimately contributing to improved overall public health outcomes.

Mammography, a low-dose X-ray technique [9], is crucial for early breast cancer detection, improving breast-conserving treatment rates and reducing mortality by over 20% [10]. It effectively identifies tumors too small to be felt, facilitating early intervention. Studies show significant mortality reduction in women aged 50 to 74 after 7 to 9 years, regardless of screening frequency [11]. Its non-invasive nature and detailed imaging enhance diagnosis and treatment planning.

However, mammography is not without its limitations, particularly concerning the risk of misdiagnosis [12]. Double reading has been proposed as a solution to reduce missed diagnoses, yet this approach significantly increases the workload for radiologists [13]. Given that the majority of mammography results are normal, the repetitive and resource-intensive nature of double reading poses a substantial burden on healthcare systems [14] [15]. Furthermore, the high costs and resource demands associated with traditional radiologist-led mammography screenings often restrict access to these services in less developed regions [16]. In resource-limited settings, it is crucial to identify which interventions are most effective and feasible in reducing overall breast cancer mortality. Some researchers are dedicated to exploring hybrid screening methods that are more suitable for women in impoverished developing countries [17]. Additionally, related studies have proposed the concept of mobile mammography services [18].

In response to these challenges, substantial efforts have been made to incorporate computer-assisted detection systems to alleviate the burden on radiologists [19]. The integration of artificial intelligence (AI) for autonomous breast cancer prevention and early detection is becoming increasingly mainstream.

Relevant studies have initiated the development of breast cancer detection and classification models based on deep transfer learning [12].

Some studies have attempted to integrate digital mammography and digital breast tomosynthesis in order to detect suspected cancerous regions on mammography images [20].

However, some researchers recently collected mammographic images obtained from different breast X-ray imaging systems and investigated their performance in intelligent early breast cancer screening tasks. They found that the effectiveness of early breast cancer screening systems is often significantly constrained by the quality, accessibility, and standardization of mammographic imaging data [21] [22] [23]. Addressing these challenges is crucial for enhancing the efficiency and accessibility of breast cancer screening across diverse populations. In addition, the inherent complexity and high resolution of mammography images pose substantial challenges for AI algorithms. These images contain intricate anatomical details that are difficult for neural networks to comprehensively extract and integrate into a unified representation.

Traditional approaches to breast cancer detection and classification from mammography images predominantly rely on supervised learning techniques [24] [25] [26]. These methods typically employ object detection networks to extract feature information from mammography images, followed by a series of operations such as feature fusion, prediction, localization, and classification.

Due to the influence of mammography's feature complexity, and breast cancer detection and classification are constrained by the quality of object detection networks, often resulting in issues such as missed detections and bounding box deviations. Moreover, the scarcity of annotated data poses significant challenges to traditional supervised learning methods in this field.

In response to these limitations, the GMIC (Global and Local Mammography Image Classification) [27] architecture introduces a novel approach based on weakly supervised learning for mammography detection and classification. This framework circumvents the constraints imposed by data scarcity, offering a promising alternative to traditional methodologies.

In traditional machine learning, each sample has a distinct label, but in multiple instance learning (MIL), samples are grouped into "bags" with known labels, while individual "instance" labels are unknown. A bag is labeled positive if at least one instance is positive; otherwise, it is negative [28]. MIL enhances generalization by training on diverse examples, reduces overfitting by exposing models to various scenarios, and improves prediction accuracy by aggregating information across instances. It is also effective for learning from weakly labeled data, useful when precise annotations are costly or difficult to obtain [29] [30].

In mammographic images, an image can be considered a "bag" containing multiple "instances," which are different image regions or patches. Multi-Instance Learning (MIL) methods allow training with bag-level labels without precise annotations. Even if we do not know which specific instance is the lesion area, the entire bag is labeled as positive if at least one instance within it is positive [31] [27]. We can use this approach to avoid being constrained by the complexity of dataset construction. Our model draws on the concept of multi-instance learning, specifically focusing on selecting important patches and features.

Multi-view learning is a machine learning paradigm that utilizes multiple distinct sets of features, or "views," to improve

learning performance. Each view provides complementary information about the data, enhancing the model's ability to understand complex patterns [32] [33]. Multi-view learning is employed in various domains such as image and video analysis, natural language processing, and bioinformatics. It is particularly useful in scenarios where data can be naturally divided into different perspectives [34] [35]. The primary advantages of multi-view learning include improved generalization performance, robustness to noise, and the ability to leverage complementary information from different views. By integrating diverse perspectives, multi-view learning can often achieve better accuracy and more insightful models compared to single-view approaches [32] [33]. The relevant study introduces the concept of multi-view learning into the classical generalized eigenvalue proximal support vector machine (GEPSVM) to demonstrate that multi-view learning can enhance model performance by coordinating the complementarity and consistency between different views [36]. The challenges of multi-view learning include effectively integrating information from different perspectives, as data from various viewpoints may require alignment or synchronization, which can be complex in practical applications, particularly when some perspectives may contain noise or misleading information [33].

Given that mammographic images are typically captured from different angles, such as craniocaudal (CC) and mediolateral oblique (MLO) views, and that the symmetry between the left and right breasts in the same view is considered a criterion for health in clinical diagnosis, a multi-view learning strategy is often employed. This approach leverages the complementary information from these different angles to enhance classification performance [37] [38].

Convolutional networks, as the most common feature extraction paradigm, are renowned for their ability to capture spatial hierarchies and local patterns in data, making them highly popular in image processing tasks [39] [40]. However, recently, Vision Transformers (ViT) based on attention mechanisms have posed a challenge to convolutional networks [41]. By employing global self-attention operations to adaptively integrate information from patches, they have achieved impressive results.

Recent work addressing the shortcomings of these two feature extraction paradigms has significantly advanced the field of computer vision. Individuals often divide into two factions, each supporting either the convolutional or the Vision Transformer (ViT) paradigm, and focusing on their respective optimizations. Some studies have demonstrated that ResNet, with appropriate training schemes and minor modifications, can perform on par with ViT [42]. The relevant study employs an asymmetric encoder-decoder architecture to randomly mask patches of the input image and reconstruct the missing pixels, enabling efficient and effective training of large models [43]. As the competition between these two paradigms evolves, there is also a continuous integration of both [44]. Recent research, inspired by Vision Transformers (ViT), has achieved more powerful models by replacing a series of small kernels in convolution with several large convolutional kernels.

Nonetheless, beyond convolution and attention mechanisms,

there are more possibilities in feature extraction tasks. The success of graph networks has also demonstrated additional potential [45]. Related work employs an anatomy-aware graph convolutional network (AGN) for early breast cancer screening in mammography, achieving promising results [46]. Recent related work directly considers the use of MLP layers for spatial interactions in the representation of image features [47].

Although clustering, as a traditional feature extraction paradigm, has gradually faded from prominence with the development of deep learning [48]. The relevant study segments the image into multiple regions by grouping a set of pixels with common characteristics. The sparsity and simplicity required for clustering demonstrate its satisfactory inter-pretability [49]. Additionally, it can connect representations of point clouds and images, reflecting its remarkable generalization capability.

We propose a Weakly Supervised Multi-instance Multiview Mammography Classification Network, which effectively utilizes unlabeled or partially labeled data for breast cancer detection and classification.

- Our work is the first to apply a non-CNN and nonattention mechanism image feature extraction method, namely Context Clustering, to the early screening task of mammography for breast cancer.
- Our work achieves state-of-the-art accuracy in the early screening task of breast cancer mammography while maintaining the lowest number of parameters.
- We propose a novel fusion mechanism that integrates global information, feature-based local information, and patch-based local information, placing greater emphasis on local details compared to previous methods.

# II. METHOD

Our research aims to develop a rapid and reliable breast cancer early screening system to alleviate the burden on medical professionals and provide screening opportunities for women's health in underdeveloped regions.

In this section, we describe the weakly supervised multiinstance multi-view network architecture we propose. In the first subsection, we formulate the steps involved in early breast cancer screening through mammography. In the second subsection, we provide a detailed description of the framework we propose. In the third subsection, we provide a detailed introduction and evaluation of the dataset used. In the four subsection, we introduce the evaluation metrics used to assess the effectiveness of early breast cancer screening.

# A. Overall Framework

The proposed model for mammography classification can be formulated as follows:

For each image I in the given view, we enhance all points into 5-dimensional information points containing color and position data to obtain the set of points  $S \in \mathbb{R}^{5 \times w \times h}$ , where  $w \times h$  is equal with the number of points.

1) Global Information Extraction: The set S is input to the first point clustering network to obtain global information  $F_g$  and Saliency-map  $I_{map}$ :

$$F_q, I_{map} = f_{\mathbf{global}}(S) \tag{1}$$

Where  $f_{\text{global}}$  represents the point clustering network for global information extraction.  $F_g$  is extracted global information from global network. The saliency-map extracted by the global network, denoted as  $I_{map}$ .

$$P = f_{roi}(I_{map})$$

The P is a set of location information, representing the positions of n example patches selected by the ROI selection function  $f_{roi}$ :

$$P = \{p_1, p_2, \dots, p_n\}$$

where  $p_n$  is a coordinate representing a position, written as  $(x_n, y_n)$ .

With  $p_n$ , we can extract n patches  $\tilde{I}$  from the original image  $I_i$  and extract n feature-based local information  $F_{fl}$  from the global feature information  $F_q$ .

$$F_{fl}^n = F_g(x_n, y_n) \tag{2}$$

where  ${\cal F}^n_{fl}$  is means n-th feature-based local information.

2) Local Information Extraction: Each selected patch  $I_n$  is treated as a new image, re-enhanced based on each point to obtain its five dimensional point set  $\tilde{S}_n$ . And processed through a second point clustering network  $f_{\text{local}}$  to obtain patch-based local information  $F_{pl}^n$ :

$$F_{pl}^{n} = f_{\text{local}}(\tilde{S}_{n}) \tag{3}$$

where  $f_{local}$  represents the point clustering network for local information extraction, similar to global clustering network  $f_{global}$ .

3) Information Fusion and Attention Mechanism: The local information  $F_l$  from all patches is fused with all feature-based local information and all patch-based local information:

$$F_l = F_{fl} \oplus F_{pl} \tag{4}$$

where Operation  $\oplus$  overlays two features.

after that processed through an attention mechanism to enhance relevant features:

$$F_a = f_{attention}(F_l) \tag{5}$$

Then, the attention-enhanced information is fused with the original global information, resulting in multi-instance fusion information from single-view:

$$F_f = f_{\text{fuse}}(F_a, F_a) \tag{6}$$

4) View Fusion and Classification: Process the images  $I_{view}$  from the four views (bilateral craniocaudal (CC) and mediolateral oblique (MLO)) using the aforementioned procedure to obtain single-view fusion information. This information is then integrated for multi-view fusion, which is used for the final binary classification, resulting in the early screening model's output:

$$F_{fusion} = f'_{fuse}(F_f^{lcc}, F_f^{lmlo}, F_f^{rcc}, F_f^{rmlo})$$
 (7)

where  $F_f^{\rm lcc}$  represents the fusion feature of LCC images, other similar situations.  $f_{fuse}'$  is another fusion structure

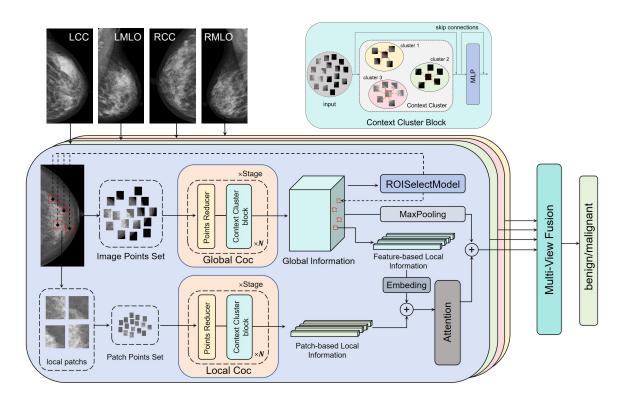


Fig. 1. Architecture of the proposed model. Images from four perspectives are enhanced into point sets and processed via a multi-level Context Clustering module, Global Coc, to extract global information. This module includes point reducers and Context Cluster Blocks. The ROISelectModel utilizes this global information to select patch-based images, which are processed through another Context Clustering module, Local Coc, to generate patch-based local information. This is fused with feature-based local information derived from the Global information to produce Local information. Subsequently, Local and Global information are combined to create single-view fusion information. Fusion information from each perspective is integrated across views and regressed to produce the final output.

different from  $f_{\text{fuse}}$ . In addition to integrating features  $F_{fusion}$ ,  $f'_{fuse}$  also merges  $F_g$  and  $F_l$  from different views.

$$F_{global} = f_{fuse}^{\prime}(F_g^{\rm lcc}, F_g^{\rm lmlo}, F_g^{\rm rcc}, F_g^{\rm rmlo})$$

$$F_{local} = f_{fuse}^{\prime}(F_l^{\rm lcc}, F_l^{\rm lmlo}, F_l^{\rm rcc}, F_l^{\rm rmlo})$$

Finally,  $F_{fusion}$  performs the final classification through  $f_{cls}$ .

$$res = f_{cls}(F_{fusion}) \tag{8}$$

This formulation encapsulates the entire process of the weakly supervised multi-instance multi-view network for mammography classification.

### B. Detailed Network Structure

1) From Image to Set of Points: The scale of an image can be expressed as (3, h, w), where 3 represents the RGB channels, and h and w are its height and width. We enhance each pixel by considering it as a 5-dimensional data point containing color and positional information (r, g, b, x, y). After this enhancement, the image can be represented as a set of  $h \times w$  5-dimensional data points, with a scale of  $(h \times w, 5)$ . We can then perform feature extraction through simple clustering. From a global perspective, the image is viewed as a collection of unordered discrete data points with color and positional information. Through clustering, all points are grouped into

clusters, each containing a centroid. Since each point in the set includes color and positional information, this clustering implicitly incorporates spatial and image information.

2) Context Cluster Block: We employ context cluster blocks for hierarchical feature extraction, a paradigm similar to convolutional networks. At the beginning of each stage, we utilize a point reducer to decrease the number of points, thereby enhancing computational efficiency. Subsequently, a series of context cluster blocks are used to extract deep features and adaptively assign aggregated features to each uniformly selected anchor point within the cluster based on similarity, and connect and fuse the nearly points through linear projection. Finally, we perform a point-wise averaging operation on the output of the last layer.

As illustrated in Figure 2, the input image undergoes point set transformation, and then, in step  $\mathbf{a}$ , n central anchor points are uniformly selected in the space. the method is similar to those in SuperPixel [50] and SLIC [51].

The selected central anchor points are highlighted with red boxes in the figure. In step  $\mathbf{b}$ , for each central anchor point, k neighbors are identified, indicated by arrows in the figure. The value of k can be 4 or 8, as determined manually, and it can also be the four neighbors in the up, down, left, and right directions, in which case k equals 4.

Step  $\mathbf{c}$  involves calculating the features of a central anchor point determined by itself and its k neighboring points, illus-

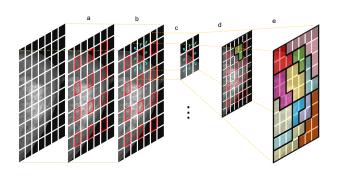


Fig. 2. The figure illustrates the clustering visualization following five steps: selecting central anchor points, identifying neighbors for each anchor, calculating features for each anchor, performing similarity analysis based on these anchors, and representing all clusters on the chart.

trated in the figure for the case where the number of neighbors is 8. The calculation process is:

$$P^x = \frac{\left(P^x + \sum_{q=1}^k p_q^x\right)}{k+1}$$

where  $P^x$  represents a x dimension of the central anchor point,  $x \in \{r, g, b, h, w\}$ . And  $p_q^x$  represents the n-th neighbor point in the x dimension,  $q \in \{0, 1, 2, \dots, k\}$ .

After computing the features for all central anchor points, a similarity analysis is conducted between all points in the point set and each central anchor point's features in step **d**. Each point is assigned to the cluster of the central anchor point with which it has the highest similarity. The steps for the similarity analysis is conducted by computing the pairwise cosine similarity matrix between a point and set of central points:

$$f(P_i, P'_j) = \frac{P_i \cdot P'_j}{|P_i||P'_j|}$$

where  $f(P_i, P'_j)$  is pairwise cosine similarity compute.  $P_i$  is i-th central anchor point,  $i \in \{0, 1, 2, \dots, n\}$ .  $P'_j$  is j-th point in the image point set.

Finally, in step **e**, all clusters are combined, resulting in the desired clustering outcome for the entire image.

Since each point contains both feature and positional information, the calculation implicitly emphasizes the point's distance (locality) and feature similarity. Each point is then assigned to the most similar center, resulting in c clusters. During this process, clusters may contain varying numbers of points. In extreme cases, some clusters may contain zero points, rendering them redundant.

*3) ROI Selection Module:* Firstly, the dimensions of the region of interest mapping are required to be consistent with the dimensions of the saliency-map.

$$h_{crop} = h_{patch} \times h_{map} \div h_I$$

$$w_{crop} = w_{patch} \times w_{map} \div w_I$$

where  $h_{patch}$  and  $w_{patch}$  are user-defined for the patch image size.  $h_{map}$  and  $w_{map}$  are the size of the saliency-map.  $h_I$  and

 $w_I$  are the size of the original image data.  $h_{crop}$  and  $w_{crop}$  are dynamically adjusted based on the original image and saliency map scales. Subsequently, the saliency-map is normalized and divided into regions with a height of  $h_{crop}$  and a width of  $w_{crop}$  regions for greedy ROI search. In each iteration, the algorithm greedily identifies each region and selects the one with the largest total weight, as determined by average pooling, among all current regions. The coordinates of this region are added to a list, and a mask flag is applied to the region to prevent redundant selection. The coordinates of these regions will be mapped to the size of the original image to obtain patch-based images.

Figure 3 visualizes the patch-based images selected by the ROISelectModule, along with the patches' positions on the source image and their comparison with the locations of suspicious lesions.

4) Attention Module: In our model, we use two different attention mechanisms to separately fuse multi-instance and multi-view information.

Similar Multi-instances information fusion: After the global network, the ROI select module to choose k patch-based images, a number set manually. This implies not all patch-based images carry beneficial information, and some may be redundant. Considering and integrating all the information from these patch-based images could significantly impair our network. Therefore, an attention module is added before integrating local and global information, allowing the model to learn how to filter out irrelevant local information.

The attention mechanism receives feature representation of patch-based images  $F_l$ , shaped as (batchsize, k, dim), where batchsize and k are manually set parameters; the former defines the batch size during training, while the latter specifies the number of patches required for multiple instance learning. The size of dim varies depending on the model.

First, we use a neural network layer with simple linear transformations  $f_{weights}$  and softmax function to compute attention weights  $W_{att}$ .

$$W_{att} = \operatorname{softmax}(f_{weights}(F_l))$$
$$F_a = F_l \odot W_{att}$$

where  $\odot$  represents the stationary point multiplication algorithm. Subsequently, the attention weights  $W_{att}$  are multiplied pointwise with the feature representation of patch-based images  $F_l$  is performed to obtain the final implicit representation  $F_a$ .

**Multi-view information fusion:** In multi-view learning, not all information from each view is necessarily classified as malignant. However, If a single view exhibits malignant characteristics, the instance should be classified as malignant. Therefore, we introduce an attention mechanism to enable the model to autonomously filter out irrelevant view information, enhancing classification accuracy.

The attention mechanism processing is largely consistent with multi-instance fusion attention. However, in multi-view attention, this attention module processes not only the  $F_a$  fused by the multi-instance attention module but also  $F_g$  and  $F_l$ . This is because all three features are considered in the loss function for loss calculation.

	Vindr-Mammo			CBIS-DDSM			
	AUC	ACC	F1 score	AUC	ACC	F1 score	Params
Single-View Res [37]	$0.727 \pm 0.02$	0.783	0.619	$0.719 \pm 0.02$	0.591	0.558	1477025
Single-View Swin-transformer	$0.731 \pm 0.02$	0.651	0.594	$0.724 \pm 0.02$	0.601	0.599	14184625
GMIC [27]	$0.793 \pm 0.02$	0.624	0.518	$0.778 \pm 0.02$	0.712	0.705	22487298
Multi-View res [37]	$0.740 \pm 0.02$	0.753	0.567	$0.731 \pm 0.02$	0.676	0.630	6128546
MaMVT [38]	$0.770 \pm 0.02$	0.918	0.647	$0.749 \pm 0.02$	0.649	0.649	30730082
Multi-View GMIC [52]	$0.797 \pm 0.02$	0.637	0.521	$0.781 \pm 0.02$	0.719	0.699	22686871
Mammo-Clustering(ours)	$\textbf{0.828} \pm 0.02$	0.919	0.694	$\textbf{0.805} \pm 0.02$	0.709	0.709	9805459

PERFORMANCE OF EACH MODEL ON TWO DATASETS

- 5) Embedding Module: The embedding module in the model primarily aligns feature-based local information  $F_{fl}$  with patch-based local information  $F_{pl}$  before their integration. Here, we employ a trainable MLP to align the scales. We designed relevant ablation experiments to verify its effectiveness.
- 6) Maxpooling Module: We employ max-pooling to fold and align global information  $F_g$ , facilitating better integration with local information fused through the attention module.

# C. Loss Function

We chose a composite loss function to achieve targeted optimization of different components.

$$L_{total} = \alpha \cdot L_{global} + \beta \cdot L_{local} + \gamma \cdot L_{fusion} + \delta \cdot L_{map}$$

After the multi-view fusion module, we retain not only the fused information for regression but also intermediate features such as global information, local information, and saliency maps. These features are used to compute a composite loss function for precise optimization of each part of the network. And we determine the sensitivity of the loss function to different types of data through component analysis.

 $L_{qlobal}$  is calculated using the global information obtained from multi-view fusion and the ground truth values. The loss function chosen here is BCELoss. And  $L_{map}$  will be calculated from the saliency-map, it is the weighted average intensity of the saliency-map under the L1 norm. The  $L_{qlobal}$ , combined with  $L_{map}$ , indicates the quality of the Global Network and further to determining the adjustment magnitude for the Global Network to enhance the accuracy of locating patch-based images. BCEWithLogitsLoss function is used for both  $L_{local}$  and  $L_{fusion}$ .  $L_{local}$  represents the quality of the local network, calculated from local information and ground truth values, determining the adjustment magnitude for the Local Network to enhance the feature extraction capability of the Local Network.  $L_{fusion}$  represents the model's final classification error, driving the overall model training. The weights  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$  represent the proportion of each loss, all manually set.

# III. EXPERIMENT AND RESULT

# A. Datasets

1) Vindr-Mammo: The Vindr-Mammo [53] dataset is a large-scale, annotated collection of digital mammographic images aimed at advancing breast cancer detection and diagnosis through machine learning. It includes thousands of images sourced from diverse populations, with detailed annotations

such as lesion types, BI-RADS categories, and precise lesion locations. This dataset is designed to support the development of robust AI models by providing a wide variety of cases, including both normal and abnormal findings, thus enhancing the generalizability and accuracy of diagnostic algorithms.

2) CBIS-DDSM: The CBIS-DDSM (Curated Breast Imaging Subset of the Digital Database for Screening Mammography) [54] dataset is a widely used resource in the field of breast cancer research. It comprises digitized film mammograms, which have been meticulously annotated with information such as lesion boundaries, types (e.g., calcifications, masses), and pathology-confirmed labels (benign or malignant). The dataset also includes patient metadata and additional clinical information, making it an invaluable tool for training and validating computer-aided detection and diagnosis systems. Its comprehensive nature and established use in the research community make it a benchmark for evaluating the performance of mammography-based AI models.

	V	indr-Mammo	)	(	CBIS-DDSM	
	benign	malignant	total	benign	malignant	total
train	3614	385	3999	629	660	1289
test	904	96	1000	185	146	331
total	4518	481	4999	814	806	1620
			TABLE	II		

THE COMPOSITION OF DATA FOR THE TWO DATASETS.

# B. Evaluating Indicator

In breast cancer early screening models, several evaluation metrics are commonly used to assess the performance of the classification models. Here are the definitions and significance of each metric along with their respective formulas:

1) AUC (Area Under the Curve): AUC means the area under the receiver operating characteristic (ROC) curve. The ROC curve uses the true positive rate for mammography benign-malignant classification as the y-axis and the false positive rate as the x-axis. It provides an aggregate measure of performance across all possible classification thresholds. A higher AUC value indicates a better model performance, with 1 representing a perfect model and 0.5 a random guess.

$$AUC = \int_0^1 TPR \left( FPR^{-1}(x) \right) dx$$

where TPR(t) is the true positive rate at threshold t, and FPR(t) is the false positive rate at threshold t

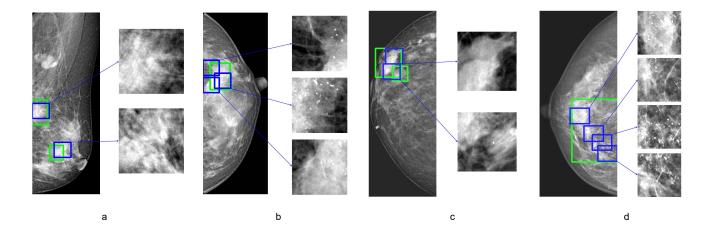


Fig. 3. Figures a to d provide a visualization of patch-based images extracted by the model. The green box on the mammography indicates the location of the suspicious lesion, while the blue box represents the patch-based images selected by the model. We can observe that the model's extracted patch-based images perform exceptionally well, and the magnified images clearly show calcifications and masses.

2) ACC (Accuracy): Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. It gives a straightforward measure of how often the classifier is correct.

$$\label{eq:acc} \text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

*3)* F1 Score: The F1 Score is the weighted average of Precision and Recall. This score takes both false positives and false negatives into account. It is particularly useful when the class distribution is uneven.

F1 Score = 
$$2 \cdot \frac{Precision \cdot Recall}{Precisionn + Recall}$$

The F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

These metrics collectively provide a comprehensive evaluation of the performance of breast cancer screening models, helping to understand their strengths and weaknesses in various aspects of classification.

# C. Comparative Experiment

In this study, we evaluated several models on two datasets: Vindr-Mammo and cbis-ddsm. The performance metrics considered were AUC, ACC, and F1 score.

1) Vindr-Mammo: For the Vindr-mammo, ours model achieved the highest performance across all metrics, with an AUC of  $0.828 \pm 0.02$ , ACC of 0.919, and F1 score of 0.694. For single-view task models, the GMIC with a Global-local structure achieves an AUC of 0.793, significantly outperforming ResNet and Swin-Transformer, which had AUCs

- of  $0.727 \pm 0.02$  and  $0.731 \pm 0.02$ , respectively. For multiview task models, our model achieves an AUC of 0.828, significantly surpassing other models. Multi-View GMIC also showed competitive performance with an AUC of  $0.797 \pm 0.02$ , validating the effectiveness of the Global-local architecture, but its ACC and F1 score were lower at 0.637 and 0.512.
- 2) cbis-ddsm: On the cbis-ddsm dataset, the ours model again demonstrated superior performance with an AUC of  $0.805 \pm 0.02$ , ACC of 0.709, and F1 score of 0.709. Multi-View GMIC, which had an AUC of  $0.781 \pm 0.02$  and an F1 score of 0.699, obtained the highest ACC on this dataset. In this dataset, the advantages of the Global-Local architecture are more pronounced, with ResNet and Swin-Transformer showing significant disadvantages in AUC.
- 3) Model Complexity: In terms of model complexity, measured by the number of parameters, the ours model had 98.05 million parameters. Other smaller networks often cannot achieve the accuracy of our model and show a significant gap. This is efficient compared to the MaMVT with 30.73 million parameters and the Multi-View GMIC with 22.68 million parameters, considering the performance gains achieved.
- 4) ROC curve: The ROC curve in figure 4 provides insights that cannot be obtained from tables alone. Analyzing the ROC curve, we observe that most models, except ours, exhibit a concave shape in the middle. This is due to class imbalance in the data, further validating the effectiveness of our model's architecture.

Overall, our model offers a robust and efficient approach, achieving state-of-the-art performance on both datasets, surpassing the second-best AUC by over 0.02, with fewer parameters. The Global-Local architecture proves effective for both multi-view and single-view models. Additionally, the multi-view learning approach enhances model performance.

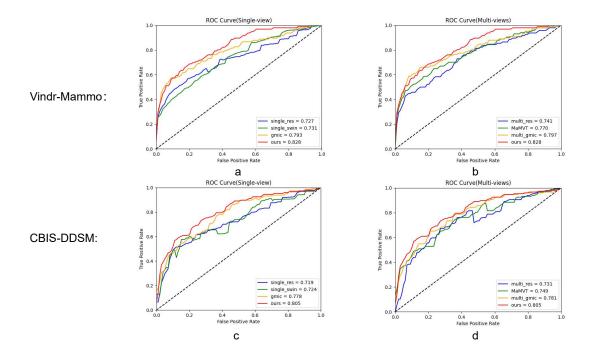


Fig. 4. Figures a and b compare the ROC curves of our model with other Single-view and Multi-view architectures on the Vindr-mammo dataset. Figures c and d present the ROC curves comparison on the CBIS-DDSM dataset.

# D. Ablation Experiment

1) Different Information Fusion Method: This ablation study aims to demonstrate the effectiveness of our proposed weakly supervised architecture on mammography.

	AUC	ACC	F1 score
Single-view Coc	$0.762 \pm 0.02$	0.711	0.627
Multi-View Coc	$0.783 \pm 0.02$	0.815	0.658
Ours(mv+gl)	$0.828 \pm 0.02$	0.919	0.694
	TARLEIII		

PERFORMANCE OF DIFFERENT INFORMATION FUSION METHODS ON THE VINDR-MAMMO

The table clearly demonstrates the superiority of our architecture, achieving the highest AUC as well as optimal ACC and F1 scores, indicating its balance in mammography tasks. mv represents a multi-view learning approach, and gl refers to our proposed weakly supervised framework.

2) Different Local Information: We identified two distinct sources of local information: patch-based local information and feature-based local information. Moreover, this feature-based local information has been overlooked in existing work.

This ablation study aims to verify the effectiveness of the mechanism integrating feature-based local information with patch-based local information.

$0 \pm 0.02$	0.873	0.670
- U - U - U - U	0.675	0.678
$06 \pm 0.02$	0.895	0.659
$28 \pm 0.02$	0.919	0.694
		$8 \pm 0.02$ 0.919

PERFORMANCE OF DIFFERENT LOCAL INFORMATION ON THE VINDR-MAMMO

We found that focusing on only one type of local information doesn't yield better results. The AUC for patch-based local information is 0.810, and for feature-based local information, it's 0.806. However, combining both achieved the best result, with an AUC of 0.828.

### IV. DISCUSSION

Figure 5 reveals that lesions in mammography predominantly appear in clustered forms. The Coc feature extraction paradigm utilizes clustering of point sets based on color and location information, making it adept at identifying the shape and position of such lesions. This paradigm offers simplicity, resulting in excellent interpretability and generalizability. Therefore, we believe that the Coc network has a strong capability to learn the prior structure of lesions in mammography images.

	Single-View GMIC	Multi-View GMIC	Ours			
MDR	0.291	0.277	0.177			
TABLE V						

THE PERFORMANCE OF DIFFERENT MODELS IN TERMS OF MISSED DETECTION RATE.

We introduce an additional evaluation metric, the missed detection rate (MDR):

$$\mathit{MDR} = \frac{N_{miss}}{N_{qt}}$$

MDR is defined as the percentage of the number of unrecognized suspicious lesion areas  $N_{miss}$  relative to the total number of suspicious lesion areas  $N_{gt}$ . Because, in clinical practice, we are more concerned about lesions being undetected, i.e., false negatives, rather than false positives.

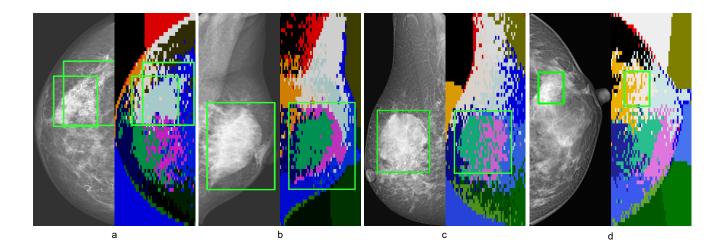


Fig. 5. From figure a to d, the left half of each image shows the original mammogram with annotated suspicious lesions, while the right half presents our contextual clustering visualization, akin to a CNN heatmap and a VIT attention map, with the suspicious lesion locations also outlined. This figure clearly shows that our Context Clustering approach effectively identifies and groups suspicious lesion areas in mammography.

From Table V, we can see that our model achieved the lowest missed detection rate of 0.177, surpassing other model with rates close to 0.1, demonstrating its potential. The other two models, like ours, are weakly supervised methods capable of identifying lesion locations using only classification labels, making them suitable for comparison.

# V. CONCLUSIONS

In this study, we developed a novel weakly supervised multi-view model for early breast cancer screening using mammography images. Unlike conventional feature extraction paradigms such as CNNs and Transformers, our approach employs a context clustering-based method. This strategy allows for the integration of feature-based local information with patch-based local information, enhancing the model's ability to capture nuanced image details. Furthermore, by incorporating multi-view mammography image features, our model effectively leverages complementary information from different perspectives. This comprehensive approach addresses the limitations of single-view analysis and improves diagnostic accuracy. The model's performance was rigorously evaluated on two publicly available datasets, Vindr-Mammo and CBIS-DDSM, where it achieved state-of-the-art accuracy, demonstrating its potential as a robust tool for early breast cancer detection.

### REFERENCES

- E. Nolan, G. J. Lindeman, and J. E. Visvader, "Deciphering breast cancer: from biology to the clinic," *Cell*, vol. 186, no. 8, pp. 1708– 1728, 2023.
- [2] B. W. Stewart, P. Kleihues, et al., World cancer report, vol. 57. IARC press Lyon, 2003.
- [3] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: a cancer journal for clinicians, vol. 71, no. 3, pp. 209–249, 2021.

- [4] F. Bray, M. Laversanne, E. Weiderpass, and I. Soerjomataram, "The ever-increasing importance of cancer as a leading cause of premature death worldwide," *Cancer*, vol. 127, no. 16, pp. 3029–3030, 2021.
- [5] J. L. Kelsey and L. Bernstein, "Epidemiology and prevention of breast cancer," *Annual review of public health*, vol. 17, no. 1, pp. 47–67, 1996.
- [6] M. M. Althobaiti, A. A. Ashour, N. A. Alhindi, A. Althobaiti, R. F. Mansour, D. Gupta, and A. Khanna, "[retracted] deep transfer learning-based breast cancer detection and classification model using photoa-coustic multimodal images," *BioMed Research International*, vol. 2022, no. 1, p. 3714422, 2022.
- [7] L. Wang, "Early diagnosis of breast cancer," *Sensors*, vol. 17, no. 7, p. 1572, 2017.
- [8] O. Ginsburg, C.-H. Yip, A. Brooks, A. Cabanes, M. Caleffi, J. A. Dunstan Yataco, B. Gyawali, V. McCormack, M. McLaughlin de Anderson, R. Mehrotra, et al., "Breast cancer early detection: A phased approach to implementation," Cancer, vol. 126, pp. 2379–2393, 2020.
- [9] E. D. Pisano and M. J. Yaffe, "Digital mammography," *Radiology*, vol. 234, no. 2, pp. 353–362, 2005.
- [10] L. Tabár, P. B. Dean, C. S. Kaufman, S. W. Duffy, and H.-H. Chen, "A new era in the diagnosis of breast cancer," *Surgical oncology clinics of North America*, vol. 9, no. 2, pp. 233–277, 2000.
- [11] K. Kerlikowske, D. Grady, S. M. Rubin, C. Sandrock, and V. L. Ernster, "Efficacy of screening mammography: a meta-analysis," *Jama*, vol. 273, no. 2, pp. 149–154, 1995.
- [12] T. Hovda, S. R. Hoff, M. Larsen, L. Romundstad, K. K. Sahlberg, and S. Hofvind, "True and missed interval cancer in organized mammographic screening: a retrospective review study of diagnostic and prior screening mammograms," *Academic Radiology*, vol. 29, pp. S180–S191, 2022.
- [13] E. Azavedo, S. Zackrisson, I. Mejàre, and M. Heibert Arnlind, "Is single reading with computer-aided detection (cad) as good as double reading in mammography screening? a systematic review," *BMC medical imaging*, vol. 12, pp. 1–12, 2012.
- [14] M. C. Posso, T. Puig, M. J. Quintana, J. Solà-Roca, and X. Bonfill, "Double versus single reading of mammograms in a breast cancer screening programme: a cost-consequence analysis," *European radiology*, vol. 26, pp. 3262–3271, 2016.
- [15] J. H. Yoon, F. Strand, P. A. Baltzer, E. F. Conant, F. J. Gilbert, C. D. Lehman, E. A. Morris, L. A. Mullen, R. M. Nishikawa, N. Sharma, et al., "Standalone ai for breast cancer detection at screening digital mammography and digital breast tomosynthesis: a systematic review and meta-analysis," Radiology, vol. 307, no. 5, p. e222639, 2023.
- [16] J. G. Elmore, K. Armstrong, C. D. Lehman, and S. W. Fletcher, "Screening for breast cancer," *Jama*, vol. 293, no. 10, pp. 1245–1256, 2005.

- [17] J. Li and Z. Shao, "Mammography screening in less developed countries," *Springerplus*, vol. 4, pp. 1–12, 2015.
- [18] U. Trivedi, T. S. Omofoye, C. Marquez, C. R. Sullivan, D. M. Benson, and G. J. Whitman, "Mobile mammography services and underserved women," *Diagnostics*, vol. 12, no. 4, p. 902, 2022.
- [19] M. Posso, T. Puig, M. Carles, M. Rué, C. Canelo-Aybar, and X. Bonfill, "Effectiveness and cost-effectiveness of double reading in digital mammography screening: a systematic review and meta-analysis," *European journal of radiology*, vol. 96, pp. 40–49, 2017.
- [20] C. F. de Vries, S. J. Colosimo, R. T. Staff, J. A. Dymiter, J. Yearsley, D. Dinneen, M. Boyle, D. J. Harrison, L. A. Anderson, G. Lip, et al., "Impact of different mammography systems on artificial intelligence performance in breast cancer screening," Radiology: Artificial Intelligence, vol. 5, no. 3, p. e220146, 2023.
- [21] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser, T. H. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, et al., "Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists," JNCI: Journal of the National Cancer Institute, vol. 111, no. 9, pp. 916–922, 2019.
- [22] K. Loizidou, R. Elia, and C. Pitris, "Computer-aided breast cancer detection and classification in mammography: A comprehensive review," *Computers in Biology and Medicine*, vol. 153, p. 106554, 2023.
- [23] J. Logan, P. J. Kennedy, and D. Catchpoole, "A review of the machine learning datasets in mammography, their adherence to the fair principles and the outlook for the future," *Scientific Data*, vol. 10, no. 1, p. 595, 2023.
- [24] M. T. Mustapha, D. U. Ozsahin, I. Ozsahin, and B. Uzun, "Breast cancer screening based on supervised learning and multi-criteria decisionmaking," *Diagnostics*, vol. 12, no. 6, p. 1326, 2022.
- [25] F. R. Cordeiro, W. P. d. Santos, and A. G. Silva-Filho, "Analysis of supervised and semi-supervised growcut applied to segmentation of masses in mammography images," *Computer Methods in Biomechanics* and Biomedical Engineering: Imaging & Visualization, vol. 5, no. 4, pp. 297–315, 2017.
- [26] B. Bektaş, İ. E. Emre, E. Kartal, and S. Gulsecen, "Classification of mammography images by machine learning techniques," in 2018 3rd International conference on computer science and engineering (UBMK), pp. 580–585, IEEE, 2018.
- [27] Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S. G. Kim, L. Moy, K. Cho, et al., "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," Medical image analysis, vol. 68, p. 101908, 2021.
- [28] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artificial Intelligence*, vol. 176, no. 1, pp. 2291– 2320, 2012.
- [29] O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification.," in *ICML*, vol. 98, pp. 341–349, Citeseer, 1998.
- [30] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proceedings of the IEEE* conference on computer vision and pattern recognition, pp. 3460–3469, 2015.
- [31] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE reviews in biomedical engineering*, vol. 10, pp. 213–234, 2017.
- [32] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 10, pp. 1863–1883, 2018.
- [33] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.
- [34] Y. Liang, H. Li, B. Guo, Z. Yu, X. Zheng, S. Samtani, and D. D. Zeng, "Fusion of heterogeneous attention mechanisms in multi-view convolutional neural network for text classification," *Information Sciences*, vol. 548, pp. 295–312, 2021.
- [35] H. Fu, Y. Geng, C. Zhang, Z. Li, and Q. Hu, "Red-nets: Redistribution networks for multi-view classification," *Information Fusion*, vol. 65, pp. 119–127, 2021.
- [36] Q. Ye, P. Huang, Z. Zhang, Y. Zheng, L. Fu, and W. Yang, "Multiview learning with robust double-sided twin svm," *IEEE transactions on Cybernetics*, vol. 52, no. 12, pp. 12745–12758, 2021.
- [37] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzębski, T. Févry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L. L. Y. Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. Toth, K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S. G. Kim, L. Heacock, L. Moy, K. Cho, and K. J. Geras, "Deep neural networks improve radiologists' performance

- in breast cancer screening," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1184–1194, 2020.
- [38] F. Manigrasso, R. Milazzo, A. S. Russo, F. Lamberti, F. Strand, A. Pagnani, and L. Morra, "Mammography classification with multiview deep learning techniques: Investigating graph and transformerbased architectures," *Medical Image Analysis*, p. 103320, 2024.
- [39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision* and pattern recognition, pp. 770–778, 2016.
- [40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
- [41] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
- [42] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on* computer vision and pattern recognition, pp. 11976–11986, 2022.
- [43] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.
- [44] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," Advances in neural information processing systems, vol. 34, pp. 3965–3977, 2021.
- [45] K. Han, Y. Wang, J. Guo, Y. Tang, and E. Wu, "Vision gnn: An image is worth graph of nodes," *Advances in neural information processing* systems, vol. 35, pp. 8291–8303, 2022.
- [46] Y. Liu, F. Zhang, C. Chen, S. Wang, Y. Wang, and Y. Yu, "Act like a radiologist: towards reliable multi-view correspondence reasoning for mammogram mass detection," *IEEE Transactions on Pattern Analysis* and Machine Intelligence, vol. 44, no. 10, pp. 5947–5961, 2021.
- [47] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al., "Mlp-mixer: An all-mlp architecture for vision," Advances in neural information processing systems, vol. 34, pp. 24261–24272, 2021.
- [48] B. Jähne, Digital image processing. Springer Science & Business Media, 2005.
- [49] Ren and Malik, "Learning a classification model for segmentation," in Proceedings ninth IEEE international conference on computer vision, pp. 10–17, IEEE, 2003.
- [50] Ren and Malik, "Learning a classification model for segmentation," in Proceedings ninth IEEE international conference on computer vision, pp. 10–17, IEEE, 2003.
- [51] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.
- [52] S. Pathak, J. Schlötterer, J. Geerdink, O. D. Vijlbrief, M. van Keulen, and C. Seifert, "Weakly supervised learning for breast cancer prediction on mammograms in realistic settings," arXiv preprint arXiv:2310.12677, 2023.
- [53] H. T. Nguyen, H. Q. Nguyen, H. H. Pham, K. Lam, L. T. Le, M. Dao, and V. Vu, "Vindr-mammo: A large-scale benchmark dataset for computeraided diagnosis in full-field digital mammography," medRxiv, 2022.
- [54] R. Sawyer-Lee, F. Gimenez, A. Hoogi, and D. Rubin, "Curated breast imaging subset of digital database for screening mammography (cbisddsm)[skup podataka]," *The cancer imaging archive*, 2016.