Domain Generalization for Mammogram Classification by Suppressing Domain-Specific Features

Jiqun Chen[1], Luhao Sun[2], Wenzong Jiang[3], Weifeng Liu[1], Chao Li[2], Zhiyong Yu[2], and Baodi Liu[1*]

[1] The College of Control Science and Engineering, China University of Petroleum (East China), China
[2] Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, China
[3] The College of Oceanography and Space Informatics, China University of Petroleum (East China), China

Abstract. Mammogram is the gold standard for early breast cancer screening, and its integration with deep learning-based computer-aided diagnosis (CAD) models has demonstrated significant advantages in improving the accuracy of breast cancer diagnosis. However, due to differences in mammography acquisition protocols and scanner models, significant inter-domain variations exist in images obtained from different mammography devices. As deep learning models tend to overfit to domain-specific feature representations during training, models trained on source domain often experience notable performance degradation when applied to cross-domain data, hindering their deployment in dynamic clinical settings. Therefore, this paper proposes a novel domain generalization approach for mammogram classification by suppressing domain-specific features (MC-SDS). MC-SDS first employs an adaptive channel filter to identify and drop channels that have a tendency to capture domain-specific features to suppress domain-specific features. Then, by perturbing the low-frequency components, the model is encouraged to learn from the high-frequency parts, further suppressing the domain-specific features present in the low-frequency components. Experiments conducted on a public dataset and two internal datasets demonstrate that MC-SDS outperforms other benchmark methods.

Keywords: Computer-aided diagnosis · Mammogram classification · Domain generalization.

1 Introduction

In recent years, the prevalence of breast cancer has surpassed that of lung cancer, becoming the leading type of cancer globally [21]. Breast cancer accounts for

* Corresponding author: Baodi Liu, email: thu.liubaodi@gmail.com

nearly 30% of all cancers diagnosed in women, and recent data indicates a continuous upward trend in its occurrence. Early detection and swift intervention of breast cancer cases are of paramount importance and are key to significantly reducing patient mortality [18]. Mammogram is characterized by its rapid imaging, sharp resolution, and an elevated signal-to-noise ratio, and is considered by imaging experts as the preferred imaging method for early screening and diagnosis [2]. However, screening mammograms are typically assessed by radiologists whose diagnostic accuracy is profoundly influenced by their professionalism [1]. Additionally, the examination procedure is not only cumbersome but also time-intensive, resulting in superfluous expenditures and redundant allocation of resources [15]. To address these challenges, numerous computer-aided diagnostic (CAD) systems have been engineered to bolster radiologists' effectiveness [20, 8, 9].

However, a major obstacle to deploying current deep learning systems into medical imaging diagnostics lies in their lack of robustness to distribution shifts between internal and external cohorts [3, 12, 11], which often exist across multiple mammography devices due to variations in the protocols for capturing images. For instance, there may be significant differences in the appearance of images obtained from different mammography devices. This disparity can deteriorate the performance of trained models.

To tackle this challenge, domain generalization (DG) emerges as a more demanding yet realistic approach. It endeavors to develop a model trained across various distinct yet interconnected source domains, with the objective of achieving robust performance across any unfamiliar target domain [6, 7]. To achieve domain generalization in mammogram classification, a natural approach is to capture only features that are clinically relevant to the disease (i.e., domain-invariant features). However, current domain generalization methods primarily impose constraints on the entire network to supervise the model in learning domain-invariant features. This method acts on the network's prediction layer, overlooking the possibility that the intermediate layers of the network may still learn excessive irrelevant information.

To address the above challenges, this paper proposes a novel domain generalization approach for mammogram classification by suppressing domain-specific features (MC-SDS). The main contributions of this work include: **(1)** MC-SDS suppresses the erroneous impact of domain-specific features on the model, thereby reducing the risk of incorrect predictions in new mammogram data domains. **(2)** A dropout-based adaptive channel filter (ACF) is proposed to identify channels that tend to capture domain-specific features and adaptively drop these channels. **(3)** A low-frequency perturbation module (LFPM) is introduced to further suppress redundant domain-specific features within the low-frequency information by perturbing the low-frequency components. **(4)** Experiments conducted on a public dataset (INbreast) and two internal datasets (InH1 and InH2) demonstrate that MC-SDS outperforms other benchmark methods.

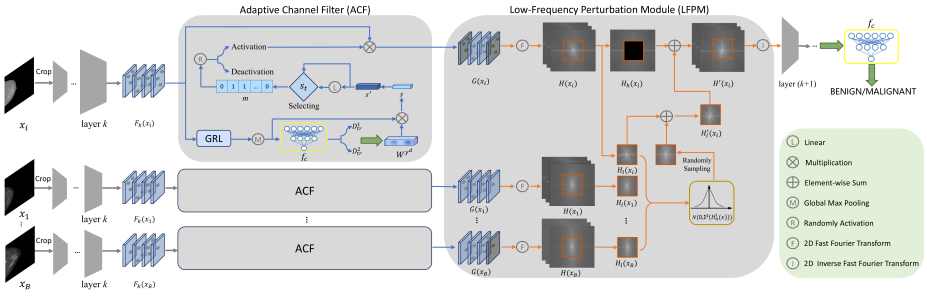

Fig. 1. Overview of the MC-SDS. The framework consists of two key components, namely ACF and LFPM.

2 Methods

Problem Setup In the context of domain generalization, we work with a setup involving N distinct source domains, denoted as $\mathcal{D}_{tr} = D_{tr}^1, D_{tr}^2, \ldots, D_{tr}^N$, each characterized by its own distribution. The objective is to leverage these diverse source domains $\mathcal{D}_{tr}$ to develop a model F that will exhibit robust performance when applied to a previously unseen target domain, referred to as D_{te}.

We assume that an iamge contains both domain-invariant features V_i and domain-specific features V_s. Specifically, V_i could be features such as the edge, shape, and curvature of a mass, which correspond to the characteristics observed by radiologists. Beyond domain-invariant features, we incorporate V_s to bridge the gap between domains in image data. V_s refers to features that appear only in specific datasets or environments. Specifically, V_s is introduced by different mammography devices during the imaging acquisition process. When the model makes predictions in a new data domain, these features lose their effectiveness.

2.1 Proposed Architecture

The whole architecture of MC-SDS is shown in Fig. 1, including the following steps: **(1)** Use Otsu's [16] segmentation method to identify the boundary of the breast, and then crop to remove the background. This preprocessing is fully automated, and it has been proven that it will not lead to incorrect cropping [26, 19]. **(2)** The cropped image is fed into the feature extractor, which utilizes the ResNet-50 backbone to generate the feature map. **(3)** Insert ACF in multiple network middle layers, and for each sample, only one randomly activated ACF is used to identify and drop channels that are prone to capturing domain-specific features, thereby suppressing domain-specific features. **(4)** After the activated ACF, LFPM is used to perform low-frequency reconstruction on the samples, further suppressing the redundant domain-specific features in the low-frequency information. **(5)** Use the fully connected layer to generate the final classification result.

2.2 Adaptive Channel Filter

The motivation behind ACF is that the features captured by the model's channels can be categorized into domain-specific features and domain-invariant features. Therefore, identifying and dropping those channels that tend to capture domain-specific features can effectively suppress domain-specific features, thereby reducing the probability of the model making incorrect predictions on new data domains. Since each domain has its own unique domain-specific features, these features serve as the basis for the domain classification task.

For a sample x_i, we initially derive the feature representation $F_k(x_i) \in R^{C\times H\times W}$ from the k-th intermediate layer. C represents the channel count, and H, W are the height and width of the feature map. $F_k(x_i)$ is fed into an ACF. To counteract the potential adverse effects of the ACF on primary network's learning, we introduce a gradient reversal layer (GRL) [13], strategically placed prior to ACF. To identify channels that are rich in domain-specific features, we rely on the ACF's performance in the middle layers as a measure of their significance. We assume that channels which are most influential for domain classification are potential carriers of domain-specific features. The correlation between individual channels and domain-specific traits is determined by computing the weighted activations that lead to accurate domain identification. For a given input $F_k(x_i)$, $W^{y^d} \in R^C$ is the FC layer weight of the ACF for the true domain label, we can compute the scores s for all channels. Specifically, the importance of the j-th channel within the feature map $F_k(x_i)$ is is scored by the expression:

$$s_j = W_j^{y^d} \cdot \mathrm{GMP}(F_k(x_i))_j \tag{1}$$

where $W_j^{y^d}$ is the j-th element of W^{y^d} , and GMP is global max pooling. A higher activation value suggests that the channel has a more significant impact on domain prediction.

After scoring the channels, in order to reduce the domain-specific information in the feature map, ACF choose and delete the most sensitive domain channels during the training process. Specifically, to make the channel scores more discriminative, we need to weight the scores s through two steps. First, through normalization methods, each element s_j of the score s is converted into probability values p_j, making the sum of probabilities for all channels equal to 1:

$$p_j = \frac{s_j}{\sum_{c=1}^{C} s_c} \tag{2}$$

where C is the number of channels. Then, an exponential amplification is performed on p_j to obtain the weighted scores s'.

The higher the weighted score, the more likely the channel is to contain domain-specific information. The weighted score is passed through a linear layer to obtain a threshold score s_t. A binary drop mask m is then created to select channels for dropping, with each element of m is generated:

$$m_j = \begin{cases} 0, & s'_j > s_t \\ 1, & s'_j \leq s_t \end{cases} \tag{3}$$

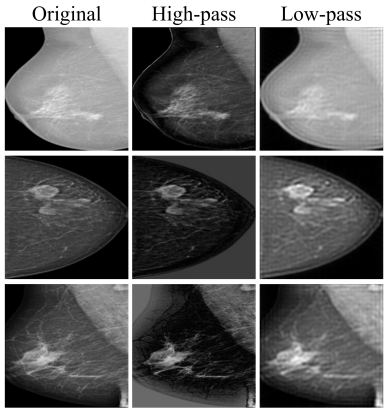

Fig. 2. Filtered image.

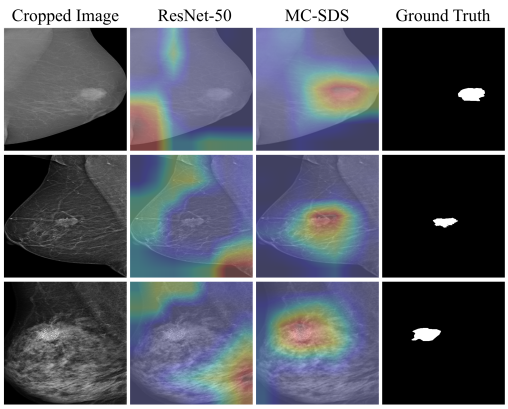

Fig. 3. Visualization results.

where j is the channel index.

Considering that the simultaneous use of dropout at multiple layers may result in the loss of too many features and hinder the learning process of the model,we introduce a multi-layer stochastic activation strategy. For each sample, we randomly select a different network layer to activate dropout. The binary drop mask only takes effect when the ACF in this layer is activated, and it guides the ACF to generate the output $G(x_i)$. If the ACF is in a deactivation state, it will not perform any processing on the input. Please note that ACF is only active during training and disabled during inference.

2.3 Low-Frequency Perturbation Module

The motivation behind LFPM is the nature of spectral properties. As shown in Fig. 2, we conducted high-pass and low-pass filtering experiments on mammogram images. It can be observed that the high-frequency information describes the edge structure of objects. The low-frequency components retain the smooth structure and stylistic information of objects. For breast cancer diagnosis, information such as the edges, shapes, and curvatures of lesions is crucial. Furthermore, since the low-frequency part retains most of the image's energy (i.e., the positional information of the target), it is inappropriate to directly remove the low-frequency components. LFPM can suppress domain-specific features in the low-frequency information by perturbing the low-frequency components, helping the model to emphasize domain-invariant features in the high-frequency components.

Given an input feature map $G(x_i) \in R^{C\times H\times W}$ that has been processed by ACF, We initially derive the Fourier transforms of the input features:

$$H(x_i)(u,v,c) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} G(x_i)(c,h,w)e^{-j2\pi\left(\frac{h}{H}u+\frac{w}{W}v\right)} \tag{4}$$

where j represents the imaginary unit, it is standard in our experimental setup to center the low-frequency components. Subsequently, we incorporate a binary mask M, an element of $R^{r\times r}$, which is null everywhere except for the central region where it takes a non-zero value:

$$M_{u,v} = \begin{cases} 1, & \text{if } \max\left(\left|u-\frac{H}{2}\right|, \left|v-\frac{W}{2}\right|\right) \leq \frac{r\cdot\min(H,W)}{2} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where r denotes the ratio governing the dimensions of M, which serves to differentiate between the high-frequency and low-frequency components within the spectrum. Utilizing this mask, we isolate the low-pass filtered frequency components as $H_l(x_i)$ and the high-pass filtered components as $H_h(x_i)$:

$$H_l(x_i) = M \otimes H(x_i) \tag{6}$$

$$H_h(x_i) = H(x_i) - H_l(x_i) \tag{7}$$

where symbol $\otimes$ represents the element-wise multiplication operation.

LFPM assumes that the low-frequency spectra across various samples follow a Gaussian distribution. We use this to generate new low-frequency spectra through resampling to substitute the initial spectra. For a set of input features $\{G(x_n)\}_{n=1}^{B}$, where B is the batch size, we initially compute their Fourier transforms and subsequently extract the low-frequency components. To streamline the notation, we refer to the low-frequency and high-frequency components as $\left\{H_n^l\right\}_{n=1}^{B}$ and $\left\{H_n^h\right\}_{n=1}^{B}$, respectively. Subsequently, we characterize the distribution of each element within the low-frequency spectrum using a multivariate Gaussian model. This model is anchored at the original element's value, and its variance is derived from the values of the same element across various samples:

$$\Sigma^2(H_n^l(u,v,c)) = \frac{1}{B}\sum_{n=1}^{B}\left[H_n^l(u,v,c) - E[H_n^l(u,v,c)]\right]^2 \tag{8}$$

The magnitude of variance Σ^2 reflects the variation in intensity of elements, accounting for potential domain shifts. We then resample each element's probability within the low-frequency spectra from the estimated distribution:

$$H_l'(x_i) = H_l(x_i) + \epsilon \cdot \Sigma(H_n^l), \quad \epsilon \sim N(0,1) \tag{9}$$

Finally, $H_l'(x_i)$ and $H_h(x_i)$ are combined into a new frequency representation $H'(x_i)$, and $H'(x_i)$ is converted back to the spatial representation through the 2D inverse fast Fourier transformation.

3 Experiments and Results

3.1 Dataset and Implementation details.

We used three Mammogram datasets: INbreast with 410 images, InH1 with 513 images from Siemens MAMMOMAT inspiration GE Senographe Essential, and

Table 1. Model's average accuracy across three datasets for different r-values.

r	0.1	0.2	0.3	0.4	**0.5**	0.6	0.7	0.8	0.9	1.0
Acc	77.23	77.47	77.63	77.87	**78.15**	77.90	77.75	77.66	77.50	77.31

InH2 with 956 images from HOLOGIC Selenia Dimensions. Each dataset was split into 90% training and 10% validation sets, with the entire dataset used for testing. The input size was 224×224 and the model was implemented in PyTorch on an RTX 3090 GPU. We used a pre-trained ResNet-50 on ImageNet, trained with SGD over 50 epochs, a batch size of 64, and an initial learning rate of 0.001, which decreased by 10% at the 80% mark. Model performance was assessed using accuracy, with the optimal model selected based on validation set performance and results averaged over five trials.

3.2 Results and Analysis

Hyperparameters Equation (5) introduces a variable r. It controls the degree of alteration applied to the low-frequency components within the image. Increasing r expands the perturbation range for low-frequency components. We have tried different values of r and the results are shown in Table 1. We found that manipulating the low-frequency segment of feature representations significantly bolsters the model's resilience against domain shifts. Furthermore, we note a performance dip with higher mask ratios, implying that extensive alterations to high-frequency image aspects may impede the acquisition of domain-invariant features. Notably, the MC-SDS approach achieves optimal results at a mask ratio r of 0.5. We select the value of r that optimizes the model's performance.

Comparison with SOTA Methods To ensure fairness, several domain generalization methods based on ResNet-50 were selected for comparison with MC-SDS. We compare with several augmentation based methods (i.e., MixStyle [25], EFDMix [23], CSU [10]), feature decorrelation methods (i.e., CausEB [4], I^2-ADR [14],). As shown in Table 2, MC-SDS achieves the best results under unknown domains in all settings. Specifically, we achieved an average performance of 78.15% across three datasets. It is worth noting that MC-SDS significantly outperformed the domain-invariant representation-based method I^2-ADR by a margin of 5.81%. This is because MC-SDS can explicitly suppress domain-specific features in every intermediate layer of the network, rather than relying on superficial statistical relationships.

It can also be observed that introducing ACF alone or introducing LFPM alone can both improve the model's accuracy. This indicates that both dropping channels that tend to capture domain-specific features and perturbing low-frequency components can suppress domain-specific features and enhance the model's generalization ability. Furthermore, using ACF and LFPM simultaneously can further enhance the model's performance.

Table 2. Comparisons with state-of-the-art domain generalization approaches: Acc evaluation (%) of three datasets on unseen domains. ◯ represents test domains, • denotes training domains.

InH1 InH2 INbreast	backbone	◯ • •	• ◯ •	• • ◯	Avg.
ResNet-50	—	72.90	74.48	75.61	74.33
MixStyle [25]	ResNet-50	73.34	76.80	71.14	73.76
FACT [22]	ResNet-50	73.65	75.27	77.07	75.33
I^2-ADR [14]	ResNet-50	74.02	74.98	68.02	72.34
EFDMix [23]	ResNet-50	71.82	76.46	70.06	72.78
MODE [5]	ResNet-50	73.80	75.39	76.97	75.39
CSU [10]	ResNet-50	71.73	73.74	75.61	73.69
XDomainMix [24]	ResNet-50	74.00	76.40	76.40	75.60
CausEB [4]	ResNet-50	75.30	76.40	75.20	75.63
ACF	ResNet-50	**74.85**	**78.13**	**78.54**	**77.17**
LFPM	ResNet-50	**73.89**	**77.24**	**78.59**	**76.57**
MC-SDS	ResNet-50	**75.82**	**78.87**	**79.76**	**78.15**

Table 3. Classification accuracy under different ratios of available training data.

Training data ratio	20%	40%	60%	80%	100%
EFDMix [23]	49.10	71.50	74.50	75.20	76.46
MODE [5]	67.15	72.38	74.16	74.90	75.39
MC-SDS	**71.23**	**73.95**	**77.00**	**77.82**	**78.87**

Limited Source Data Experiment Due to the scarcity of medical data, it is necessary to evaluate model's average classification accuracy with limited training data. We evaluate MC-SDS on target domains (InH2) using partial training data from source domains (INbreast and InH1). The classification accuracy on the target domain is reported in Table 3. We observe that MC-SDS consistently outperform EFDMix and MODE. Even with minimal data, MC-SDS outperforms other methods. This is because when there is a limited amount of data, the model tends to overfit to domain-specific features (such as low-frequency background noise from specific equipment). MC-SDS can suppress these features and force the model to focus on domain-invariant features.

Visual Explanation To intuitively verify the claim that MC-SDS has learned features that are invariant across domains, we used Grad-CAM [17] to compute and visualize the regions of interest in three mammograms from the test set, as shown in Fig. 3. The shade of color in the graph reflects the network's malignant probability towards the region. It can be seen that by adding ACF and LFPM, the network corrected the previous misidentification of the lesion areas.

4 Conclusions

To address the issue of insufficient generalization capability of deep learning models in mammogram classification, this paper proposes a novel domain generalization method for the classification of mammogram. MC-SDS integrates ACF and LFPM. ACF suppresses domain-specific features by identifying and dropping channels that tend to capture domain-specific features, reducing reliance on device-style information. LFPM further inhibits domain-specific features within the low-frequency components by perturbing the low-frequency part, helping to emphasize domain-invariant features in the high-frequency components. Experiments conducted on three datasets (INbreast, InH1 and InH2) demonstrate that MC-SDS outperforms other benchmark methods in terms of performance. Considering the class imbalance issue often present in mammogram datasets, which may affect the model's generalization ability for rare cases, addressing the class imbalance problem will be a direction for future improvement.

Acknowledgments. This work was supported by the Shandong Natural Science Foundation under Grants ZR2024MF102, 05J24050006, ZR2023MF008, the National Natural Science Foundation of China under Grant 62372468, the Major Basic Research Projects in Shandong Province under Grant ZR2023ZD32, the Qingdao Natural Science Foundation under Grant 23-2-1-161-zyyd-jch.

Disclosure of Interests. The authors have no competing interests to declare relevant to this article's content.

References

1. Bae, M.S., Moon, W.K., Chang, J.M., Koo, H.R., Kim, W.H., Cho, N., Yi, A., La Yun, B., Lee, S.H., Kim, M.Y., et al.: Breast cancer detected with screening us: reasons for nondetection at mammography. Radiology **270**(2), 369–377 (2014)
2. Bruno, A., Ardizzone, E., Vitabile, S., Midiri, M.: A novel solution based on scale invariant feature transform descriptors and deep learning for the detection of suspicious regions in mammogram images. Journal of Medical Signals & Sensors **10**(3), 158–173 (2020)
3. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nature Communications **11**(1), 3673 (2020)
4. Chen, L., Zhang, Y., Song, Y., Zhang, Z., Liu, L.: A causal inspired early-branching structure for domain generalization. International Journal of Computer Vision pp. 1–21 (2024)
5. Dai, R., Zhang, Y., Fang, Z., Han, B., Tian, X.: Moderately distributional exploration for domain generalization. arXiv preprint arXiv:2304.13976 (2023)
6. Guo, J., Qi, L., Shi, Y.: Domaindrop: Suppressing domain-sensitive channels for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19114–19124 (2023)
7. Guo, J., Wang, N., Qi, L., Shi, Y.: Aloft: A lightweight mlp-like architecture with dynamic low-frequency transform for domain generalization (2023), https://arxiv.org/abs/2303.11674

8. Hupse, R., Karssemeijer, N.: Use of normal tissue context in computer-aided detection of masses in mammograms. IEEE Transactions on Medical Imaging **28**(12), 2033–2041 (2009)
9. Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C.I., Mann, R., den Heeten, A., Karssemeijer, N.: Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis **35**, 303–312 (2017)
10. Liu, Y., Zou, Y., Qiao, R., Liu, F., Lee, M.L., Hsu, W.: Cross-domain feature augmentation for domain generalization. arXiv preprint arXiv:2405.08586 (2024)
11. Lu, C., Ahmed, S.R., Singh, P., Kalpathy-Cramer, J.: Estimating test performance for ai medical devices under distribution shift with conformal prediction. arXiv preprint arXiv:2207.05796 (2022)
12. Ma, W., Chen, C., Zheng, S., Qin, J., Zhang, H., Dou, Q.: Test-time adaptation with calibration of medical image classification nets for label distribution shift pp. 313–323 (2022)
13. Matsuura, T., Harada, T.: Domain generalization using a mixture of multiple latent domains. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11749–11756 (2020)
14. Meng, R., Li, X., Chen, W., Yang, S., Song, J., Wang, X., Zhang, L., Song, M., Xie, D., Pu, S.: Attention diversification for domain generalization. In: European conference on computer vision. pp. 322–340. Springer (2022)
15. Myers, E.R., Moorman, P., Gierisch, J.M., Havrilesky, L.J., Grimm, L.J., Ghate, S., Davidson, B., Mongtomery, R.C., Crowley, M.J., McCrory, D.C., et al.: Benefits and harms of breast cancer screening: a systematic review. Jama **314**(15), 1615–1634 (2015)
16. Otsu, N., et al.: A threshold selection method from gray-level histograms. Automatica **11**(285-296), 23–27 (1975)
17. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
18. Selvi, R.: Breast diseases: imaging and clinical management. Springer (2014)
19. Shu, X., Zhang, L., Wang, Z., Lv, Q., Yi, Z.: Deep neural networks with region-based pooling structures for mammographic image classification. IEEE transactions on medical imaging **39**(6), 2246–2255 (2020)
20. Sun, Z., Jiang, H., Ma, L., Yu, Z., Xu, H.: Transformer based multi-view network for mammographic image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 46–54. Springer (2022)
21. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
22. Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A fourier-based framework for domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14383–14392 (2021)
23. Zhang, Y., Li, M., Li, R., Jia, K., Zhang, L.: Exact feature distribution matching for arbitrary style transfer and domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8035–8045 (2022)
24. Zhang, Z., Wang, B., Jha, D., Demir, U., Bagci, U.: Domain generalization with correlated style uncertainty. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2000–2009 (2024)

25. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. arXiv preprint arXiv:2104.02008 (2021)
26. Zhu, W., Lou, Q., Vang, Y.S., Xie, X.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. pp. 603–611. Springer (2017)