2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3315-8618-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/EMBC58623.2025.11254317

Enhanced Multiple Instance Learning for Breast Cancer Detection in Mammography: Adaptive Patching, Advanced Pooling, and Deep Supervision

Fareeha Sarwar[1], Nuno Miguel de Figueiredo Garrido[1], Pedro Sebastião[1], Margarida Silveira[2]

***Abstract*— This paper addresses the challenge of weakly supervised learning for breast cancer detection in mammography by introducing an Enhanced Embedded Space MI-Net model with deep supervision. The framework integrated adaptive patch creation, convolution feature extraction, and pooling methods -max, mean, log-sum-expo, attention, and gated attention pooling - evaluated in three MIL models, Instance Space mi-Net, Embedded Space MI-Net and Enhanced Embedded Space MI-Net. A key contribution is the incorporation of deep supervision, improving feature learning across network layers and enhancing bag-level classification performance. Experimental results on the CBIS / DDSM dataset demonstrate that the Enhanced MI-Net model achieves the highest AUC of 86% with attention pooling. This work addresses the gap in leveraging MIL techniques for high-resolution medical imaging without requiring detailed annotations, offering a robust and scalable solution for breast cancer detection.**

***Clinical Relevance*—This study highlights the potential of MIL-based models with attention pooling to accurately detect breast cancer in mammographic images without requiring detailed ROI annotations, offering a scalable and efficient diagnostic tool for clinical practice.**

I. INTRODUCTION

Among the most diagnosed cancers worldwide breast cancer holds the highest prevalence among women. In 2020, it was the leading cancer globally, with an estimated 2.3 million cases, making up 11.7% of all cancer diagnoses, and contributing to 6.9% of cancer-related deaths worldwide [1]. Studies demonstrate that early breast cancer identification is essential for the management of breast cancer, enhancing prognosis and survival rates. Full field digital mammography is a two-dimensional breast X-ray imaging technique that has demonstrated efficacy in decreasing breast cancer mortality and efficiency in breast cancer prediction [2]. In consideration of the significant success of Deep Neural Networks (DNN's) in computer vision, multiple deep learning techniques have been explored to address the issue of efficient lesion detection in mammography [3]. Numerous suggested approaches address the challenge as a segmentation or classification task, incorporating the identification of Region of Interest (ROI) and the delineation of lesion areas in phases. Annotating mammograms requires domain-savvy radiologists, which adds workload and makes it difficult to find enough for huge data sets. Reducing annotation needs in deep learning mammographic image analysis systems is difficult. Popular Convolutional Neural Network (CNN) approaches in deep learning are primarily geared toward classifying typical image data sets. Most deep networks for mammographic image interpretation use resized images. Resizing a large mammogram to a smaller one may make undetected tumors harder to identify, affecting model performance [4]. To cope with these challenges, recent research has demonstrated the significance of the Multiple-Instance Learning (MIL) paradigm for medical image analysis applications [5]. However, to date, research employing MIL algorithms for medical image processing remains constrained due to challenges such as insufficient annotated data and variability in medical images. To generalize supervised classification, MIL group images or image patches into bags, where each bag contains multiple instances and is assigned a single class label instead of labelling individual instances. This makes MIL a weakly supervised method [6]. In this study, the problem of breast cancer detection from mammography images is approached as a binary classification task, where the goal is to determine whether a given image contains malignant lesions. This method only requires the label of mammographic images as normal/benign or malignant, eliminating the need for more detailed annotations. This study introduces an enhanced MI-Net model that incorporates deep supervision layers and advanced pooling techniques to improve feature aggregation and classification performance in MIL significantly. Unlike conventional approaches, this method eliminates the need for segmentation and ROI annotations, offering a more adoptable and efficient solution for high-resolution imaging tasks. To demonstrate it's effectiveness, the enhanced MI-Net is systematically compared against mi-Net, which processes individual instances before combining them for classification, and MI-Net, aggregate instance level features into bag for better classification. Furthermore, the study explores five pooling strategies—max pooling, average pooling, log-sum-exp pooling, attention pooling and gated attention pooling—within these frameworks. By integrating deep supervision with adaptive bag creation, the proposed approach enhances patch selection, refines feature represen-

[1]Fareeha Sarwar, Nuno Miguel de Figueiredo Garrido and Pedro Sebastião are with Instituto de Telecomunicações, IT-IUL, fsraa1@iscte-iul.pt, nuno.garrido@iscte-iul.pt, pedro.sebastiao@iscte-iul.pt

[2]Margarida Silveira is with the Institute for Systems and Robotics, LARSyS, Instituto Superior Técnico, Universidade de Lisboa, Portugal, msilveira@isr.tecnico.ulisboa.pt

MS supported by FCT (DOI: 10.54499/LA/P/0083/2020, 10.54499/UIDP/67250009/2020, 10.54499/UIDB/50009/2020). F.S. gratefully acknowledges the support of the ISCTE-IUL Merit Scholarship for funding her PhD studies.

tation, and strengthens bag-level predictions, contributing to the advancement of MIL-based classification.

II. RELATED WORK

Several investigations have explored the use of CNN's to automatically diagnose breast cancer from mammograms. Dhungel et al. [7] devised a detection system employing a multistage methodology for the classification of mammograms. Nevertheless, their methodology necessitated an expensive ROI detection step, and the multistage training process restricted the DNN's ability to perform end-to-end learning, limiting its full potential. Shu et al. [8] Introduced Region-based Group max-Pooling(RGP) and Global Group max-Pooling (GGP) architectures for convolutional neural networks as substitutes to traditional pooling methods, which split images into regions and choose a subset with a high probability of malignancy to represent a full mammographic image. Their pooling approach increases the effectiveness of the models on mammographic imaging data. However, their approach was limited to using single-view images, failing to capture the connections between different image perspectives of the patients.

Numerous recent studies have applied MIL to analyze mammographic images for breast cancer detection. Sánchez De la Rosa et al. employed the MIL paradigm [9]. In the proposed approach, the breasts are adaptively segmented into distinct regions, and textual or mass/micro classification are subsequently extracted from each region. Elmoufidi et al. [10] applied a computer-assisted detection technique for mammography. Quellec et al. [11] also developed mammography classification algorithms. Both works involved dividing breast images into regions, extracting features, and input them into MIL algorithms. They showed that anomaly detectors and classifiers can be beneficial. Nevertheless, their tasks need meticulously built features tailored to individual data. Before adding lesion ROI into models, it was necessary to determine their types, and varied approaches were employed for lesion detection in mass and classification images or concentrated on mass classification. Elmoufidi et al.[12] also worked on MIL model for breast cancer detection where features from segmented ROI and its sub-layers extracted and put into bags. Limitation of this work involves use of handcrafted features and their proposed architecture performance dependability to segmentation accuracy. Li et al. developed a novel region label assignment technique that utilizes all areas in a patient's mammogram by assigning labels and loss for each region specifically [13]. They also proposed an Area Under the Curve (AUC)-based optimization method for selecting mini-batches, aiming to improve the model's ability to differentiate between classes especially where data imbalance exists. Zhu et al. [14] investigated three methods for designing a deep MIL network for mammography classification. They used sparse theory to implement the general MIL assumption in label assignment issues. The strategy only explored a tiny portion of regions for model training, potentially losing information from unexplored regions. Different pooling strategies including max, average and attention-based pooling were used in [15] to aggregate features for histopathology breast cancer dataset. Their work emphasized that attention-based pooling strategy is essential as the pertinent regions (lesion zones) are unknown beforehand and must be discerned from the complete high-resolution image. Whole slide image classification is frequently formulated within the MIL paradigm. Although deep supervision has been employed to enhance feature representation [16], most existing methods rely on fully supervised settings. In this work, we incorporate deep supervision into an MIL framework to more effectively address the challenges of weakly supervised learning. Similarly, attention-based MIL combined with domain-specific feature extraction has been shown to improve performance in high-resolution pathology images [17], supporting our choice of advanced pooling and custom feature extraction. These works collectively motivate our enhanced MIL approach, which integrates adaptive patching, deep supervision, and robust pooling to improve breast cancer detection in mammography.

III. DATASET AND METHODS

A. Dataset

This study uses the CBIS-DDSM dataset, which is an improved and curated version of the original DDSM. It contains high-resolution mammography images in DICOM format. While the dataset provides annotations such as ROI segmentations and bounding boxes for visual reference, these annotations are not used in model training. Instead, only image-level labels are used for training the MIL models.The dataset includes a total of 2,620 images with 1,196 images labelled as 0 class(benign) and 1,424 images labeled as 1 class (malignant) [18]. For this research, the data is divided into three subsets: 70% for training, 20% for validation, and 10% for testing. The split is performed at the patient level to prevent images from the same patient from appearing in both training and test subsets, ensuring data independence and reducing the risk of leakage.

B. Preprocessing

To improve the visual quality of the images and help the model better identify important patterns, several preprocessing techniques are applied. One key method used is Contrast Limited Adaptive Histogram Equalization (CLAHE) [19], which is known for its ability to enhance local contrast and reveal subtle structures in medical images. In this study, CLAHE is applied with a clip limit of 2, which helps to control over-amplification of noise while still improving contrast. After enhancement, all images are normalized to the [0,1] range [20]. Unlike many deep learning pipelines, the original image resolution was maintained, and no resizing was performed on the full mammogram images.

C. Adaptive Patch Extraction

The mammography image (the bag) is divided into smaller overlapping patches (the instances) using a fixed stride value for analysis. This way, intricate characteristics are effectively captured without losing critical information at

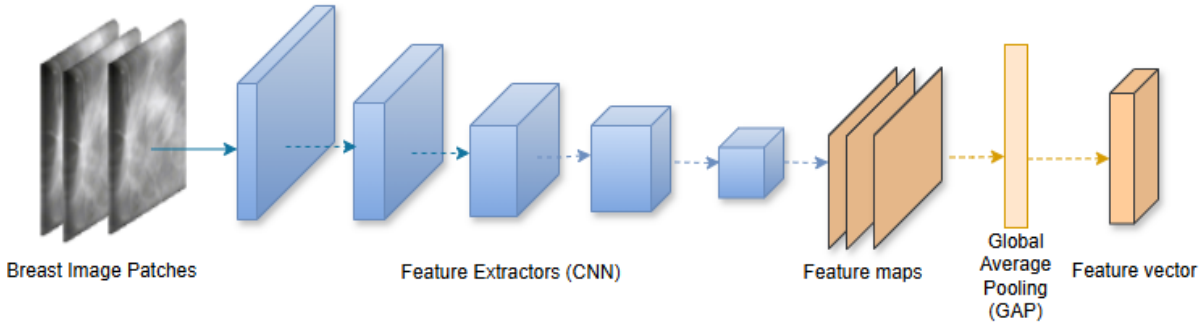

Fig. 1. Feature Extraction Module

the image peripheries [21]. For this study, patches of size 256x256 were extracted with a stride of 64 from the full-size mammographic image. The mammographic dataset had artifacts like projection labels and black backgrounds, which were reduced using techniques like Gaussian Blur, Median Filtering, and Bilateral Filtering, through these sometimes cause the loss of important features. Also, some suitable techniques are costly in terms of computational resources and memory. For this reason, breast masking is adopted as the first step, focusing on the breast region and reducing non-informative background noise. In addition, patches are also chosen based on variance calculations carrying substantial variability and informative content. In this way, low-variance patches are discarded.

D. Feature Extraction

A custom feature extraction module was designed to process the extracted patches using convolutional and pooling layers. The network begins with a convolutional block (conv1) of 64 feature channels. This is followed by max pooling to reduce spatial dimensions. The four subsequent blocks (conv2-conv5) incrementally increase feature channels to 128, 256, 512 and 1024, each incorporating max pooling layers, batch normalization and Rectified Linear Unit (RELU) activation function to enhance feature learning. Residual connections using 1x1 convolutions with a stride of 2, are added to mitigate gradient vanishing and improve convergence. The feature extraction module ends with a Global Average Pooling(GAP) layer, producing a 1024-dimensional feature vector for bag-level classification in the MIL model. Global average pooling was also added to make patch extraction more modular and efficient. While pre-trained models offer faster convergence, a custom-designed feature extractor was chosen for primarily two reasons. First, pre-trained models such as those trained on ImageNet, often fail to capture unique mammogram patterns, as they are optimized for natural images rather than medical images, leading to reduced accuracy for domain-specific tasks [22]. Second, a custom architecture provides flexibility to adapt to specific input sizes, resolutions and domain-specific requirements, enabling better adaptation to the intricacies of mammographic data [23]. Fig. 1 shows the CNN feature extraction module.

E. Multiple Instance Neural Networks

As in MIL definition, instances are grouped into bags and during the training and testing process, only bag labels are assigned, while individual instance labels remain unavailable. Two MIL constraints are:

- If bag X_i is negative, then all instances in bag X_i are also negative, i.e., if $X_i = 0$ then $x_{ij} = 0$.
- If bag X_i is positive, then at least one instance in bag X_i must be positive, that is, if $X_i = 1$ then $x_{ij} \geq 1$.

In this work, we focus on two main MIL architectures known as mi-Net and MI-Net [24]. mi-Net focuses on predicting instance-level probabilities, which are aggregated via pooling to obtain a bag-level prediction. On the other hand, MI-Net focuses on obtaining instance embeddings using a shared network, but no instance-level predictions are made, the instance embeddings are aggregating to obtain a global bag-level representation which is used to directly predict the bag label.

1) Instance Space MIL Algorithm (mi-Net): Fig. 2 depicts our model which incorporates two Fully Connected (FC) layers that progressively reduce the feature dimensions from 1024 to 512 and then, to 256 as well as to refine instance features. Dropout is applied after each FC layer (with a rate of 0.5), helping to prevent overfitting, and the final classification FC layer computes the instance probability. The instance-level probabilities are then aggregated using a pooling function to compute a single probability score for the entire bag. Consider a setting where p_{ij} denotes the probability of the j-th instance of i-th bag, M represents the pooling function to aggregate instance-level probability to bag-level representation p_i and m_i is the total number of instances in the i-th bag. fc-final represents the final FC layer of the network. Maps the probability of instance p_i to the bag representation. In summary, mi-Net can be stated in (1) as

$$\begin{aligned} P_i &= M\left(p_{ij} \mid j = 1, \ldots, m_i\right) \\ y_i' &= \text{fc-final}(P_i) \end{aligned} \tag{1}$$

2) Embedded Space MIL Algorithm (MI-Net): Fig. 3 depicts our model MI-Net, including two FC layers (FC1, FC2) progressively reducing the feature dimensions from 1024 to 512 and then, to 256. Dropout is applied after each FC layer (with rate of 0.5), to prevent overfitting. These FC layers

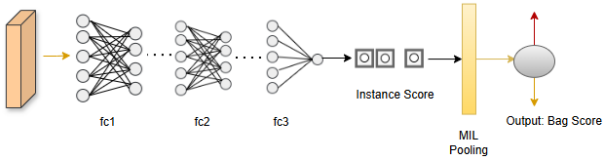

Fig. 2. mi-Net MIL model with fully connected layers. The last layer is the pooling layer that accepts instance probability as input and outputs bag probability

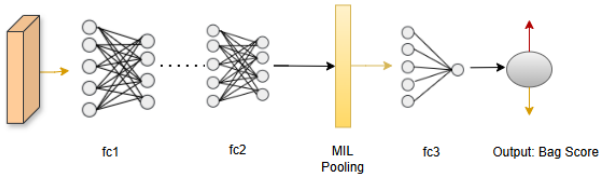

Fig. 3. MI-Net MIL model with fully connected layers and a pooling layer that accepts instance features as input and outputs bag representation. The classification layer converts bag representation into bag predictions

refine the instance level features before aggregating features as well help in better understanding semantic and bag content relationships as well as dimensionality reduction in that way models will learn more abstract and useful features. After refining the features a pooling function aggregates the instance-level features to form a bag-level representation. Lastly, a final FC3 layer used to predict the bag label. [25]. In this research, we used this model for calculating bag prediction directly from instance features. In (2) x_{ij} is the feature vector of the j-th instance in the i-th bag. P is the pooling function that aggregates instance features into bag representation X_i and classification layer (fc-final) transforms the bag representation into a bag prediction.

$$\begin{aligned} X_i &= P\left(x_{ij} \mid j = 1, \ldots, n_i\right) \\ y_i' &= \text{fc-final}(X_i) \end{aligned} \tag{2}$$

3) MI-Net with Deep Supervision: MI-Net model was enhanced with deep supervision by leveraging the outputs from intermediate convolutional layers at different depths (conv3, conv4 and conv5). The output from intermediate layers is processed through auxiliary fully connected layers for additional supervision. During training, these auxiliary outputs ensure that early layers receive more direct guidance, while testing the score from these layers provides a more robust bag classification. In this enhancement we also use a residual connection in patch extraction module to preserve information flow. The aim of enhancing MI-Net is the use of multilevel supervision strategy that will improve the network's ability to learn deeper level features and increase robustness. In (3), the pooling function P is applied across all instances in the bag to generate a single bag-level representation, where k denotes the specific level or layer of pooling in the network. For example, k=3 may correspond to pooling at an intermediate layer, while k=4 corresponds to pooling at a deeper layer.

$$\begin{aligned} X_i^k &= P^k\left(x_{ij}' \mid j = 1, \ldots, n_i\right) \quad \text{for} \quad k = \{3, 4, 5\} \\ y_i^k &= \text{fc\_aux}^k(X_i^k) \end{aligned} \tag{3}$$

F. MIL Pooling Methods

As previously mentioned, we employed a pooling layer to aggregate patch scores and patch representation. In this research, we compare five MIL pooling methods: max pooling, average pooling, log-sum-exp pooling, attention-based pooling and gated attention pooling methods. Max pooling focuses on the most significant instances by selecting the maximum feature value, highlighting the most critical part of the bag. Average pooling averages the features of all instances, providing a generic overview of the bag features. Log-sum-exp pooling offers a smooth approximation to max pooling by weighing all instances exponentially, which balances the influence of larger values while considering the smaller ones. Attention Pooling has been used in MIL models for various medical tasks[26], [27]. Attention pooling learns to assign varying weights to instances based on their importance to bag classification, allowing the model to focus dynamically on relevant instances. Gated attention Pooling [28] is an enhancement of attention pooling. It adds a gated mechanism that controls instance contribution, enabling more nuanced and context aware representation.

G. Experimental Configuration

The experimental setup for this study is detailed in Table I. To address the class imbalance in the dataset, oversampling was applied to the training data. The models were trained for binary classification using binary cross-entropy with logits, and the Adam optimizer was employed to ensure stable convergence. A learning rate scheduler was implemented to reduce the learning rate by a factor of 0.1 every 20 epochs, balancing rapid convergence with gradual fine-tuning. The enhanced MI-Net model included supervision layers (conv3, conv4, and conv5) to improve performance and enable visualization of critical features. Multi-output loss was computed by combining the main loss with auxiliary losses from intermediate layers, with specific weights determined through a greedy search by minimizing validation loss.

IV. Results and Discussion

Table II presents a comparative analysis of mi-Net, MI-Net and Enhanced MI-Net models using five pooling strategies. The leftmost column lists the pooling methods applied, while the right section displays the corresponding accuracy and AUC values for each model. Among the tested approaches, Enhanced MI-Net with attention pooling achieved the highest performance, with 75% accuracy and 86% AUC. The MI-Net with attention pooling also performed well, attaining 78% accuracy and 85% AUC. Max and gated attention pooling in MI-Net also shows competitive results. Enhanced MI-Net demonstrates a better ability to distinguish between positive and negative cases, as reflected by its higher AUC. Overall, attention-based pooling methods led to improved

TABLE II

TEST SET PERFORMANCE OF MIL MODELS WITH VARIOUS POOLING METHODS. RED AND BLUE REPRESENT THE BEST AND SECOND BEST RESULT, RESPECTIVELY.

Pooling Methods	mi-Net		Mi-Net		Enhanced MI-Net	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Log-Sum-Expo	70	78	73	80	71	79
Mean	67	74	70	76	68	74
Max	69	77	76	84	70	82
Attention	71	83	78	85	75	86
Gated Attention	73	81	75	83	73	80

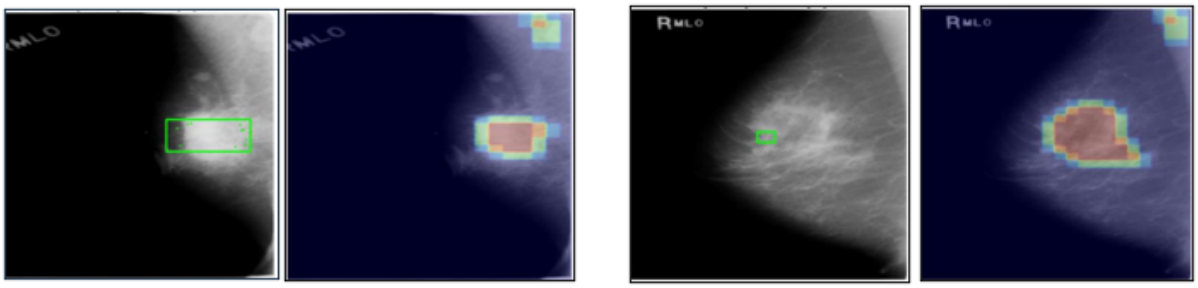

Fig. 4. Visualization of heatmaps showing the model's focus areas, for two test cases: a cancerous region (left) and a benign region (right). In each example, the ROI annotation is shown on the left and the attention heatmap on the right.

TABLE I

HYPERPARAMETERS CONFIGURED FOR MODEL TRAINING

Hyperparameter	Configuration
Loss Function	Binary Cross-Entropy with Logits
Optimizer	Adam
Initial Learning Rate	0.0001
Weight Decay	Applied (value not specified)
Learning Rate Scheduler	StepLR (γ = 0.1, every 20 epochs)
Dropout Rate	0.5 (in fully connected layers)
Activation Function	Sigmoid

TABLE III

RESULTS COMPARISON WITH OTHER STATE OF THE ART ON CBIS/DDSM DATASET. RED AND BLUE REPRESENT THE BEST AND SECOND BEST RESULT, RESPECTIVELY.

Methodology	Accuracy(%)	AUC(%)
Deep MIL [14]	71.70	77.16
RGP Pooling MIL [8]	76.00	82.20
MCLRA [13]	76.55	82.37
Attention Pooling Enhanced MI-Net	75.43	86.35
Attention Pooling MI-Net	78.00	85.00

feature representation and model performance across all three models. Unlike the multi-scale deep learning approach by Quintana et al. [21], which employs varying patch sizes and resolutions, this research focuses on adaptive fixed-size patches along with advanced pooling strategies. Furthermore, unlike the methodology proposed by Li et al. [13], which assigns region-specific labels to mammograms to enhance lesion detection and optimize the model using AUC to handle class imbalance, or Shu et al. [8], which reduces annotation dependency through RGP/GGP, the current studies involve a more flexible and robust approach tailored specifically to high-resolution mammographic images. Experimental results in Table III demonstrate a high AUC of 86%, surpassing several state-of-the-art methods on the same CBIS/DDSM dataset. Unlike other models, which often emphasize multi-resolution analysis or region-specific labeling, this study offers a more adaptable solution designed for analyzing mammographic images effectively.: Inference Benchmarks reveals that all three models are lightweight and well-suited for near real-time for clinical use. The deep supervision model leads with the fastest inference time (0.0434s per bag/image), closely followed by MI-Net (0.0526s) and mi-Net (0.0538s). Comparable classification performance was also achieved using ResNet-based pre-training, though those results are not shown here. Impressively, Deep supervision achieves this speed without extra memory cost, its peak GPU usage (2595.04 MB) is nearly identical to the others. This combination of computational efficiency and rapid inference (benchmarking on GPU (Quadro RTX 8000, 48GB) positions the model as a promising candidate for deployment in time-sensitive clinical environment. Fig. 4 shows heat maps that illustrate the model focus areas for prediction in two test cases: a cancerous region (left) and a benign region (right).

V. CONCLUSION

This research introduces an advanced computer-aided detection and diagnosis system leveraging the MIL paradigm to identify breast cancer in mammography images. By combining adaptive patch creation, CNN feature extraction, and advanced pooling techniques, the proposed approach

enhances the performance of breast cancer detection using MIL. Advanced pooling methods, such as attention and gated attention significantly improve the model's ability to focus on critical image regions, achieving state-of-art performance on the CBIS/DDSM dataset. Unlike traditional approaches, this framework eliminates the need for detailed ROI annotations, making it both efficient and scalable. Furthermore, deep supervision enhances feature learning, ensuring robust bag-level predictions and achieving an AUC of 86%. This study highlights the potential of MIL framework in advancing medical image diagnosis, particularly in handling high-resolution images where conventional convolutional methods may fall short.

In the future, transformer-based attention pooling methods can be explored to assess richer spatial and contextual features. Additionally, evaluating this model on 3D breast cancer modality dataset could enhance generalization across diverse clinical settings.

ACKNOWLEDGMENT

F. Sarwar gratefully acknowledges the invaluable support of ISCTE-IUL and Instituto de Telecomunicações in advancing her research and academic pursuits.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] M. Broeders, S. Moss, L. Nyström, S. Njor, H. Jonsson, E. Paap, N. Massat, S. Duffy, E. Lynge, and E. Paci, "The impact of mammographic screening on breast cancer mortality in Europe: a review of observational studies," *Journal of medical screening*, vol. 19, no. 1_suppl, pp. 14–25, 2012.

[3] M. Bahl, "Detecting breast cancers with mammography: will AI succeed where traditional CAD failed?" pp. 315–316, 2019.

[4] D. A. Zebari, H. Haron, D. M. Sulaiman, Y. Yusoff, and M. N. M. Othman, "CNN-based deep transfer learning approach for detecting breast cancer in mammogram images," in *2022 IEEE 10th Conference on Systems, Process & Control (ICSPC)*. IEEE, 2022, pp. 256–261.

[5] J. Buler, R. Buler, M. Bobowicz, M. Ferlin, M. Rygusik, A. Kwasigroch, and M. Grochowski, "Interpretable deep learning approach for classification of breast cancer-a comparative analysis of multiple instance learning models," in *2023 27th International Conference on Methods and Models in Automation and Robotics (MMAR)*. IEEE, 2023, pp. 105–110.

[6] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[7] N. Dhungel, G. Carneiro, and A. P. Bradley, "The automated learning of deep features for breast mass classification from mammograms," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 106–114.

[8] X. Shu, L. Zhang, Z. Wang, Q. Lv, and Z. Yi, "Deep neural networks with region-based pooling structures for mammographic image classification," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 2246–2255, 2020.

[9] R. S. de la Rosa, M. Lamard, G. Cazuguel, G. Coatrieux, M. Cozic, and G. Quellec, "Multiple-instance learning for breast cancer detection in mammograms," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 7055–7058.

[10] A. Elmoufidi, K. El Fahssi, S. Jai-andaloussi, A. Sekkaki, Q. Gwenole, and M. Lamard, "Anomaly classification in digital mammography based on multiple-instance learning," *IET Image Processing*, vol. 12, no. 3, pp. 320–328, 2018.

[11] G. Quellec, M. Lamard, M. Cozic, G. Coatrieux, and G. Cazuguel, "Multiple-instance learning for anomaly detection in digital mammography," *Ieee transactions on medical imaging*, vol. 35, no. 7, pp. 1604–1614, 2016.

[12] A. Elmoufidi, "Deep multiple instance learning for automatic breast cancer assessment using digital mammography," *IEEE transactions on instrumentation and measurement*, vol. 71, pp. 1–13, 2022.

[13] D. Li, L. Wang, T. Hu, L. Zhang, and Q. Lv, "Deep multiinstance mammogram classification with region label assignment strategy and metric-based optimization," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 4, pp. 1717–1728, 2021.

[14] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "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*. Springer, 2017, pp. 603–611.

[15] G. Li, C. Li, G. Wu, D. Ji, and H. Zhang, "Multi-view attention-guided multiple instance detection network for interpretable breast cancer histopathological image diagnosis," *IEEE Access*, vol. 9, pp. 79 671–79 684, 2021.

[16] L. Qu, Y. Ma, X. Luo, Q. Guo, M. Wang, and Z. Song, "Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[17] L. Godson, N. Alemi, J. Nsengimana, G. P. Cook, E. L. Clarke, D. Treanor, D. T. Bishop, J. Newton-Bishop, A. Gooya, and D. Magee, "Immune subtyping of melanoma whole slide images using multiple instance learning," *Medical Image Analysis*, vol. 93, p. 103097, 2024.

[18] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017.

[19] S. Bagchi, K. G. Tay, A. Huong, S. K. Debnath *et al.*, "Image processing and machine learning techniques used in computer-aided detection system for mammogram screening-a review," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 3, pp. 2336–2348, 2020.

[20] G. M. El-Banby, N. S. Salem, E. A. Tafweek, and E. N. A. El-Azziz, "Automated abnormalities detection in mammography using deep learning," *Complex & Intelligent Systems*, vol. 10, no. 5, pp. 7279–7295, 2024.

[21] G. I. Quintana, Z. Li, L. Vancamberg, M. Mougeot, A. Desolneux, and S. Muller, "Exploiting patch sizes and resolutions for multi-scale deep learning in mammogram image classification," *Bioengineering*, vol. 10, no. 5, p. 534, 2023.

[22] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[23] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[24] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern recognition*, vol. 74, pp. 15–24, 2018.

[25] M. U. Oner, J. M. S. Kye-Jet, H. K. Lee, and W.-K. Sung, "Studying the effect of MIL pooling filters on MIL tasks," *arXiv preprint arXiv:2006.01561*, 2020.

[26] M. Waqas, M. A. Tahir, S. Al-Maadeed, A. Bouridane, and J. Wu, "Simultaneous instance pooling and bag representation selection approach for multiple-instance learning (MIL) using vision transformer," *Neural Computing and Applications*, vol. 36, no. 12, pp. 6659–6680, 2024.

[27] Z. Han, B. Wei, Y. Hong, T. Li, J. Cong, X. Zhu, H. Wei, and W. Zhang, "Accurate screening of covid-19 using attention-based deep 3d multiple instance learning," *IEEE transactions on medical imaging*, vol. 39, no. 8, pp. 2584–2594, 2020.

[28] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.