Biomedical Signal Processing and Control 123 (2026) 110574

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

# Multi-modal learning for automatic breast cancer diagnostics from mammography and ultrasound

Yhonatan Kvich [a,*,1], Adi Kalamaro [a,1], Shachar Ashkenasy [a], Adi Wegerhoff [a], Yishai M. Elyada [b], Ahuva Grubstein [c,d], Eli Atar [c,d], Yonina C. Eldar [a]

[a] Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel
[b] Mobileye Vision Technologies, Ltd., Jerusalem, Israel
[c] Radiology Department, Beilinson Campus, Rabin Medical Center, Petah Tikva, Israel
[d] Gray Faculty of Medicine and Health Sciences, Tel Aviv University, Tel Aviv, Israel

ARTICLE INFO

*Keywords:*
Breast cancer (BC)
Multi-Modality (MM) learning
Mammography imaging
Ultrasound imaging
Automated diagnostics
Artificial intelligence (AI) in healthcare

ABSTRACT

Screening mammography (MG) and ultrasound (US) are used for breast cancer (BC) diagnostic, offering complementary diagnostic information. Existing classification approaches, including artificial intelligence (AI), primarily focus on distinguishing between benign and malignant cases, overlooking the critical distinction between normal, benign, and malignant classes—thereby assuming the presence of abnormalities. Additionally, many methods depend on physician input, such as manually cropping regions of interest (ROIs) or contouring areas in US, which limits automation and scalability. We propose a fully automatic patient-level Multi-Modality (MM) screening classification without physician input. Our learning framework that integrates MG and US data to improve classification without manual physician input. Our approach is the first to leverage C-view MG, a synthesized MG format, within a large dataset of 1250 patients with paired MG and US images. The framework employs an ArcFace layer to enhance tumor class separation. We evaluated two fusion techniques — feature concatenation and cross-attention — to effectively combine MG and US. Our MM model achieves Area Under the Curve (AUCs) of 95.5%, 97.5%, 98.9% for distinguishing normal, benign, and malignant cases, respectively. The evaluations highlight benefits of C-view MG and ArcFace, showing substantial improvements over models that exclude components or use architectures designed for cropped images and binary (benign/malignant) classification in prior studies. This work highlights the potential of MM approaches in BC classification, especially when combining complementary modalities, including C-View MG, with advanced imaging techniques and feature fusion strategies. By eliminating the need for manual physician input, our framework represents a significant step toward efficient, automated diagnostic solutions.

## 1. Introduction

Breast cancer (BC) is one of the leading causes of mortality in women worldwide, making early detection crucial, particularly for younger patients. Traditionally, 2D mammography (MG) has been the primary screening tool, as it enables the detection of smaller tumors, thereby contributing to reduced mortality rates [1]. However, its effectiveness is limited in women with dense breast tissue, where a "masking effect" occurs due to overlapping parenchyma. To address this limitation, supplemental imaging methods such as ultrasound (US) are recommended for high-risk individuals. Additionally, Digital Breast Tomosynthesis (DBT) MG offers a three-dimensional view of the breast, helping to overcome the tissue overlap challenge inherent in 2D MG [2]. From DBT imaging, C-view MG images are synthesized to create a 2D representation. This process reduces radiation exposure by eliminating the need for separate 2D scans, effectively addressing the limitations of standard MG.

Despite these technological advancements, the increased workload for radiologists remains a significant concern, particularly given global shortage of specialists [3]. Additionally, interpreting results from multiple imaging modalities poses challenges, often leading to considerable variability among radiologists [4,5]. Developing a fully automatic patient-level multi-modality screening system without physician input is especially important in general screening settings, where no prior indications are available. Although multiple imaging modalities may

* Corresponding author.
*E-mail address:* yonatan.kvich@weizmann.ac.il (Y. Kvich).
[1] Equal contribution

https://doi.org/10.1016/j.bspc.2026.110574
Received 31 August 2025; Received in revised form 7 April 2026; Accepted 5 May 2026
Available online 13 May 2026

be used, effectively combining their complementary strengths in an automated framework remains challenging. In recent years, deep learning models have demonstrated significant potential for improving BC detection, particularly through the application of MG based models [6–8]. The authors in [9] show an Area Under the Curve (AUC) of 0.84 for classification between benign and malignant cases. In [10], the authors exclusively used US scans and fine-tuned several well-known pre-trained architectures, including ResNet50, VGG-12, VGG-16, and VGG-19, for the task of classifying normal, benign, and malignant cases. Among these, ResNet50 achieved the highest accuracy of 77.77%. The methods in [8–10] did not require additional manual input from physicians.

To address the inherent limitations of MG and US, several studies have explored the combination of those modalities for BC classification with promising results compared to using MG alone [11–14]. Those studies focus predominantly on binary classification between benign and malignant, limiting their ability to effectively differentiate normal tissue from pathological findings [12–16]. For instance, an ensemble method that combined classical machine learning algorithms such as SVM, naïve Bayes, and K-NN for both MG and US images achieved an AUC of 0.89 in distinguishing between malignant and benign cases [14]. Healthy cases present a significant challenge in classification, as benign features — such as dense tissue or benign masses — can closely resemble abnormalities like calcifications, masses, and distortions, making it difficult to differentiate normal from suspicious findings [17]. Limiting the model to benign or malignant cases reduces its applicability, as it assumes prior diagnosis of a tumor condition. Effective handling of cases where healthy individuals are suspected of cancer is essential to ensure broader diagnostic accuracy.

Another challenge arises from the dependence on manual tumor localization techniques, such as annotations and region of interest (ROI) cropping, increasing the workload for radiologists and introducing variability into the diagnostic process as in [12–15]. The authors in [13] applied a pre-trained GoogleNet model for a Multi-Modality (MM) BC classification using manually annotated tumor regions. This approach achieved an AUC of 0.94 for classification between benign and malignant cases on a limited dataset of 153 patients. Another study [15] utilized a dataset of 31 patients to develop a MM model by employing SVM. The study used manually cropped ROI images, processed under the guidance of expert radiologists, and achieved an AUC of 0.99 for binary tumor classification. Similarly, Chen et al. [16] explored a MM approach for binary breast lesion classification. To minimize manual effort, they used YOLOv8 for automatic ROI detection, achieving an AUC of 0.97.

By focusing exclusively on ROI inputs, such approaches risk omitting relevant spatial information present in the full breast image, potentially missing critical diagnostic indicators, such as those associated with multifocal cancer [18]. Additionally, the reliance on small datasets poses a major limitation in many studies, underscoring the need for larger and more diverse datasets to enhance model performance and ensure broader applicability. Overall, the studies mentioned emphasize the need for automated and scalable methods capable of handling both pathological and healthy cases effectively.

In this paper, we utilize neural networks for three scenarios: two single-modality cases — one for MG and one for US — and a MM integration network. Our dataset includes 1250 patients, each providing four MG images captured from two distinct views, encompassing both 2D and 3D information, as well as a corresponding single US image per case. The MG modality includes various breast angles, integrating domain knowledge through C-view MG. The classification is always between three categories: normal, benign, or malignant. In the screening setting, women who are not clarified as having a suspicious findings, and relying on such priors would require physician input and shift the task toward diagnosis rather than screening. Moreover, our method does not assume pre-annotated or cropped regions of interest, which would again depend on expert labeling. Therefore, the ability to perform fully automatic patient-level classification across the three classes is particularly valuable for screening workflows, where efficient, large-scale assessment of normal, benign, and malignant cases is essential. A concise comparison between prior approaches and our proposed framework is summarized in Table 1.

**Table 1**
Comparison between prior work and this work.

| Prior work | This work |
|---|---|
| Lesion diagnosis | Screening triage |
| Cropped ROI | Full image |
| Binary classification | Realistic three-class workflow |
| Physician-assisted | Fully automated |

To enhance the dataset and address the challenges posed by limited instances, we expanded the number of instances by pairing the same four MG images with different US views from each patient. For each patient, several US images were available, allowing us to generate multiple instances by coupling the fixed MG images with a different US image each time. This approach ensures greater variability and robustness in the dataset. A pre-trained ResNet50 [19] serves as the backbone for feature extraction. The ArcFace layer [20] enhances class separability by introducing an additive angular margin to the softmax loss function, effectively enforcing a more discriminative feature space. This is particularly useful in our classification task, as it reduces intra-class variability and increases inter-class distances in the embedding space, leading to more robust decision boundaries. For MM integration, we implement two fusion strategies, feature concatenation and cross-attention mechanisms [21,22], effectively combining the complementary strengths of MG and US modalities. The fused features are then passed through a fully connected layer for final classification. To address class imbalance, which arises naturally from the statistical distribution of individuals undergoing diagnostic tests, the networks are trained using a weighted cross-entropy loss function. US excels at detecting normal tissue patterns, while MG is particularly sensitive to malignant features, this sensing story is reflected in the single-modality results. By fusing features from both modalities, the MM framework leverages their complementary strengths to perform well across classes.

To evaluate diagnostic accuracy at the patient level, we conducted multiple assessments using different US images for each patient while keeping the same four MG images. The final predictions were aggregated using a majority voting mechanism, combining results from all generated instances per patient. By integrating the complementary strengths of both modalities, our framework improves the BC diagnostic process, ensuring accurate classification, including the detection of normal cases, all without reliance on physician input. Our best network, a MM architecture utilizing concatenation for feature fusion, with C-View MG data and ArcFace, achieved AUCs of 95.5% for normal, 97.5% for benign cases, and 98.5% for malignant cases. To ensure robust evaluation, we partition the dataset into 10 equally sized folds, rotating through them such that each patient appears exactly once in the validation set. All reported results reflect validation performance averaged across these folds.

Our proposed model was compared to other approaches and outperformed them, demonstrating the effectiveness of the MM approach. The comparison included several baselines to assess the impact of different components. First, we evaluated single-modality networks, including a standalone US model and an MG model trained with and without C-View MG. This allowed us to isolate the benefits of using C-View MG in classification performance. We further compared against MM architectures that excluded key components, such as ArcFace and C-View MG. The MM model without ArcFace demonstrated weaker class separation, highlighting the importance of our discriminative transformation. Similarly, excluding the C-View MG images in the MM model resulted in lower performance, reinforcing the advantage of leveraging C-View MG. Additionally, we implemented the approach

from [13], which utilized a pre-trained GoogleNet for BC classification. We trained this model both with and without C-View MG to evaluate its effectiveness under different conditions. In both cases, our proposed MM model surpassed the performance of the GoogleNet-based method, further validating the advantages of our additional discriminative layer and feature fusion.

The remainder of this paper is organized as follows: Section 2 describes our clinical dataset and the preprocessing steps applied. Section 3 details the methods used, including both single-modality and MM architectures. Section 4 presents the results, including evaluation, analysis, and comparison. Finally, Section 5 concludes the work and suggests potential future directions.

## 2. Clinical dataset

For this study, data was extracted from digital records at Rabin Medical Center - Belinson Hospital, focusing on women who underwent MG and US within a six-month period. To ensure reliable outcomes, cases classified as benign or healthy required a minimum two-year follow-up, while cancer diagnoses had to be confirmed within six months of the MG exam. The Breast Imaging Reporting and Data System (BI-RADS), developed by the American college of radiology, provides a standardized lexicon, assessment categories, and management recommendations to ensure consistent interpretation, reporting, and outcome monitoring of breast imaging across modalities. To maintain dataset integrity and avoid bias, inclusion criteria were applied. MG images containing surgical clips, pacemakers, or scar markers were excluded. Similarly, US images with annotations or Doppler features, were excluded to prevent the model from receiving unintended cues. This filtering was performed manually through visual inspection, though not by a clinical expert. Radiologists reviewed and categorized the dataset into three groups: malignant (136 cases), benign (446 cases), and normal (668 cases). As data for both breasts was not consistently available, we analyzed only one breast per patient, using its corresponding diagnostic information. We received images containing relevant findings. Our objective is to utilize our clinical dataset from both modalities to develop a robust MM classification for patient diagnosis. All data were collected from a single medical center, which may limit the diversity of imaging protocols and patient populations.

The MG images in this dataset were predominantly sourced from Hologic systems, with a smaller proportion from Siemens, and encompass three key types

1. The Mediolateral Oblique (MLO) view, captures the breast at an angle from the upper outer to the lower inner region.
2. The Craniocaudal (CC) view, captures the breast tissue from above, providing a direct view of the central and inner regions
3. The C-View synthesized 2D images derived from 3D DBT, offering clear tissue visualization with reduced radiation, elevating the benefits of 3D imaging.

When referring to the MG of a patient, we are specifically discussing the use of a series of 4 images : MLO 2D, CC 2D, MLO C-View, CC C-View. By using four different MG images, we offer a new combination, by expanding domain knowledge in our dataset, with the integration of C-VIEW images enhancing the classification of abnormalities that standard 2D MG might miss [23].

US images were collected from a wide range of equipment, mostly Siemens ACUSON S2000 and Supersonic MACH 30, ensuring diverse representation of imaging technologies [23]. This variety enhances the dataset's robustness, enabling comprehensive evaluation of the model's performance across different imaging systems. See Table 2 for a summary of dataset metadata including demographics, and imaging devices. The dataset, free from prior interventions or annotations, allows the model to learn directly from the medical images. US images from over five vendors, with varying dimensions and extraneous noise, were included to ensure dataset quality. Vendor-specific cropping procedures were applied to generate clean, focused images, minimizing noise and irrelevant details. For each patient, multiple US images were captured, each from a different angle, further enriching the dataset's diversity.

**Table 2**
Dataset metadata summary.

| | |
|---|---|
| Total number of patients | 1814 |
| Age (years) | Min: 31, Max: 89, Mean: 52.43 |
| Age distribution | 30–35: 14, 35–40: 96, 40–45: 406, 45–50: 394, 50–55: 234, 55–60: 38, 60–65: 42, 65–70: 144, 70–75: 129, 75–80: 66, 80–85: 26, 85–90: 11 |
| Sex | All female |
| US device | S2000: 940, Aixplorer MACH30: 665 iU22: 129, LOGIQE9: 76 Sequoia: 9, ABVS Workplace: 8 Aixplorer Ultimate: 4, Other: 7 |

We applied standard image augmentation techniques, including random horizontal flip and random rotation, to increase dataset variability while maintaining the natural clinical appearance of US and MG images [24]. For each patient, multiple US images were available, each treated as a separate entity to generate individual data instances. For every unique US image, we paired it with the same four MG images from the corresponding patient, creating a new data instance. This approach significantly increased the total number of data instances, allowing for training on a larger-scale dataset. Pairing the same MG examination with multiple US views reflects routine clinical practice, where MG provides global structure and US captures localized variability. This design leverages their complementary roles rather than introducing representation bias.

We divided the dataset into training (70%), testing (20%), and validation (10%) subsets, maintaining consistent class distributions and ensuring that each patient appeared in only one subset. To assess generalization, we implemented a 10-fold cross-validation scheme in which the dataset was split into ten equal parts, and each subset served once as the validation set while the remaining were used for training and testing. This ensured that every patient contributed to the evaluation exactly once. This cross-validation strategy ensures that the vendor distribution is also mixed across folds, as each fold's validation set is a randomly selected subset drawn from the entire dataset, including all vendors.

In the next section we present our single modalities and MM networks.

## 3. Methods

In this section, we begin by describing the single-modality models used for MG and US classification, followed by an explanation of the fusion strategies employed in the MM framework. Our single-modality MG model processes each view using a pre-trained ResNet50 encoder, which extracts high-level visual features from the four MG views. During training, the first two layers of the ResNet architecture are kept fixed, while the remaining layers are fine-tuned. This process generates four feature vectors, each of length 2048. These vectors are passed through two fully connected (FC) layers with LeakyReLU activations, reducing their dimensionality to 1024 and then to 512. To further refine the feature representation, a cross-attention mechanism [22] is applied, aggregating information across the four MG views into a single vector of length 512. This mechanism selectively emphasizes the most relevant features by computing weighted scores through a scoring function, improving classification accuracy by focusing on critical regions within the images.

The resulting cross-attention weighted vector is processed using the decoupling ArcFace method [20]. This approach leverages a contrastive

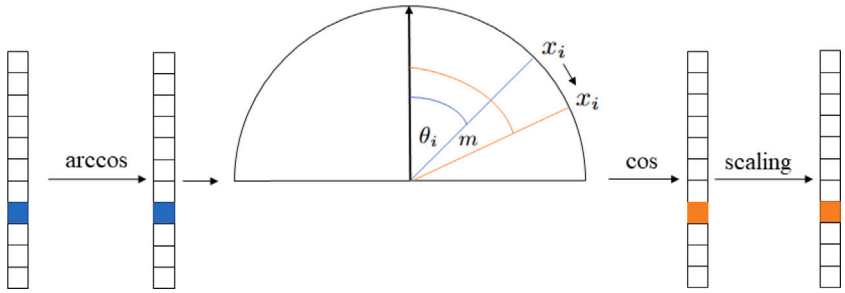


**Fig. 1.** ArcFace operator with angular margin for improved class separation.

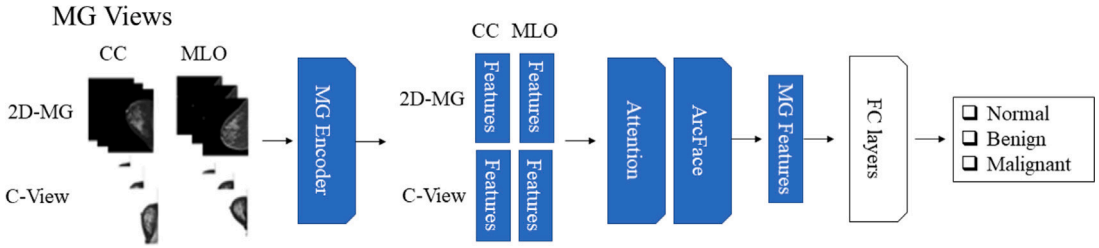


**Fig. 2.** Single-modality MG Architecture.

learning layer to construct a more distinct feature space, significantly improving class separation. It introduces an additive angular margin $m$ to the classification decision boundaries, enabling clearer distinctions between categories in the *cosine* space. This transformation is calculated by the following trigonometric identity

$$\cos(\theta_i + m) = \cos(\theta_i)\cos(m) - \sin(\theta_i)\sin(m) \tag{1}$$

where $\theta_i$ is the angle between the feature vector of the input and the weight vector corresponding to class $i$. By adding the margin $m$, the decision boundary is pushed further in angular space, making the model more confident in distinguishing between classes. This results in the enhanced cosine similarity, represented as $\cos(\theta_i + m)$.

Conceptually, ArcFace enforces angular margin separation in the embedding space, which is particularly advantageous in MM settings where features originate from heterogeneous modalities and class differences may be subtle. By promoting well-aligned and discriminative embeddings prior to fusion, it facilitates more reliable separation between normal, benign, and malignant cases. Fig. 1 presents a schematic of the ArcFace operator, illustrating its role in enforcing angular margin penalties to improve class separation.

Finally, the feature vector is passed through three additional FC layers, each utilizing LeakyReLU activations, progressively reducing its size from 256 to 128, then to 64, and ultimately to 3, corresponding to the classification output. Refer to Fig. 2 for a visual representation of the architecture. The final classification, using the softmax function, defined as

$$\Pr[y = i \mid x] = \frac{e^{s\cdot\cos(\theta_i + m)}}{\sum_{j=1}^{C} e^{s\cdot\cos(\theta_j)}} \tag{2}$$

where $s$ is a scaling factor used to stabilize the training process.

Our single-modality US model processes a single US view as input, following a pipeline similar to the MG single-modality model. Feature extraction is performed using a pre-trained ResNet50 encoder, capturing high-level visual features. The extracted feature vector, initially of length 2048, is reduced to 1024 and then to 512 using two FC layers with LeakyReLU activations, similar to the MG model. To enhance feature discrimination, the ArcFace method is applied, further reducing the feature vector's dimensionality from 512 to 256. Finally, the resulting vector passes through three FC layers with LeakyReLU activations, reducing its size from 256 to 3 for the classification output. This dimensionality reduction process mirrors the approach used in the single-modality MG model, maintaining consistency across modalities.

For our MM, we utilized the single-modality architectures as feature extractors, generating high-dimensional feature vectors from each modality after their ArcFace layers, each of dimension 256. We investigate two fusion methods: concatenation and cross-attention. The first approach involves concatenating the embedded features. This is highly effective for distinguishing subtle differences in more discriminative feature space. The concatenated vector represents the joint information from both modalities and serves as the input to subsequent layers, which are designed to perform the final classification with fully connected layers. By leveraging concatenation, this approach preserves the unique characteristics of each modality while enabling the model to learn complementary relationships across modalities. This is followed by FC layers with LeakyReLU activations, reducing the vector to a dimensionality of 3. Fig. 4 provides a visual depiction of the architecture.

The second fusion method uses a cross-attention mechanism [21] on the extracted feature vectors. This technique allows the model to focus on the most relevant features. Unlike concatenation, which passively aggregates features across modalities, cross-attention actively facilitates the dynamic interaction and selective alignment of features. This mechanism enables the model to adaptively emphasize modality-specific information, such as refining representations from MG based on US and vice versa, optimizing inter-modal feature integration for enhanced learning. This process follows the standard multi-head cross-attention mechanism, where each feature vector is used to derive key ($K$), value ($V$), and query ($Q$) representations. Cross-attention model is then applied, with each modality receiving the query from the other modality

$$\hat{f}_{\mathrm{US}} = Att(Q_{\mathrm{MG}}, K_{\mathrm{US}}, V_{\mathrm{US}}), \quad \hat{f}_{\mathrm{MG}} = Att(Q_{\mathrm{US}}, K_{\mathrm{MG}}, V_{\mathrm{MG}}) \tag{3}$$

where $\hat{f}_{\mathrm{US}}$ and $\hat{f}_{\mathrm{MG}}$ are the updated feature representations. The general cross-attention function $Att$ is defined as:

$$Att(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{C}}\right)V \tag{4}$$

where $C$ is the number of channels. The cross-attention fusion process is illustrated in Fig. 3, showing the exchange of queries between modalities, the use of multi-head cross-attention, and the concatenation of $\hat{f}_{\mathrm{US}}$ and $\hat{f}_{\mathrm{MG}}$.

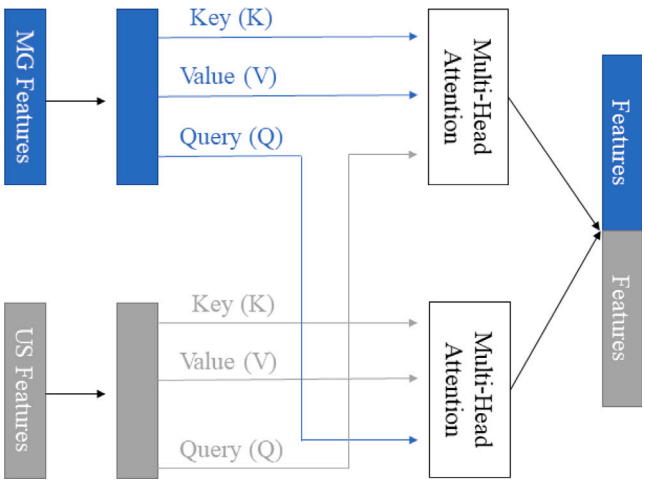


**Fig. 3.** Cross-attention fusion with multi-head cross-attention and concatenation.

Like in the previous fusion method, we use FC layers to get a vector of length 3. The effectiveness of these fusion strategies is evaluated in the results section, where concatenation approach is shown to perform better, highlighting the benefits of direct feature aggregation. Training was performed using PyTorch with Python 3.9.7 on an NVIDIA L40 GPU, while inference was conducted using an NVIDIA A40 GPU, also with PyTorch All networks, including single-modalities and MM models, were optimized using weighted cross-entropy to address class imbalances. Optimization was performed with Stochastic Gradient Descent (SGD), employing a learning rate of 0.001 over 100 epochs.

In the clinical setting, the proposed fully automated pipeline operates on paired US and MG images from a patient without any prior clinical information. Following a simple preprocessing step, as described above, which can be performed by a technician and does not require physician input, the data are directly fed into the network to produce a three-class (normal, benign, malignant) classification.

## 4. Results

Recall that each US image was treated as a separate instance and paired with the same four MG images for each patient. The evaluation for each instance is done by taking the class with the higher value. To reflect the clinical objective of diagnosing at the patient level, we evaluate the model by aggregating predictions across all paired US and MG images from the same patient using a majority voting mechanism. This ensures that the final diagnosis accounts for the full multi-modal input per patient, rather than individual image instances. In this section, we compare the performance of single-modality models, MM architectures — with and without C-View MG and the ArcFace layer — and previously used GoogleNet models trained with and without C-View MG. All evaluation are performed on the validation dataset.

The single modality MG model demonstrates high performance in detecting malignant cases, achieving an AUC of 95.4%, a recall of 95.5%, and a precision of 98.8%, highlighting its robustness in identifying critical cases. It achieves lower AUC for normal cases, with values of 67.5%, a recall of 68.0%, and a precision of 65.1%. Its performance in classifying benign cases is also weak, as reflected by a lower AUC of 70.0%, a recall of 70.1%, and a precision of 73.0%. Notably, the use of MG imaging as the modality in this study underscores its significant strength in enabling high precision, particularly in detecting malignant conditions. The single-modality US model accurately identifies normal cases, achieving an AUC of 91.3%, a recall of 91.6%, and a precision of 94.9%. However, for benign cases, the model shows a lower performance, with an AUC of 84.6%, a recall of 84.9%, and a precision of 67.9%. For malignant cases, the model presents a low AUC of 57.5%, a recall of 46.4%, and precision of 57.4%. While the US modality excels in detecting normal cases, it struggles with benign and malignant classifications. See Table 3 for the AUC values of the models across the different classes, reported as the mean ± standard deviation over the 10-fold cross-validation.

While each modality demonstrates individual strengths, their combination in MM has the potential to enhance overall performance. For instance, US excels in detecting normal cases, but shows weaker performance in identifying malignant cases. In contrast, MG achieves superior performance in detecting malignant tumors but falls short in classifying normal case. When fused through MM approaches like concatenation or cross-attention, these strengths are combined, leading to improved overall performance. The concatenation method, in particular, balances the strengths of both modalities, achieving high and consistent AUC values across all classes. This complementarity demonstrates the importance of leveraging both modalities, as one modality compensates for the limitations of the other, leading to a more robust classification. The concatenation method yields strong performance, achieving an AUC of 95.5%, 97.5%, and 98.5% for normal, benign, and malignant cases, respectively. There is also a low standard deviation across the different folds with values of 5.2, 3.4, and 3.0 for normal, benign and malignant class, receptively. this indicate that the methods achieve consistent results, showing generalization. This demonstrates an improvement in benign detection over the single-modality US (84.6%) and MG (70.0%) models. Meanwhile, cross-attention methods shows slightly lower AUC values for the benign and malignant classes (96.2% and 87.5%, respectively). In contrast, the model without ArcFace achieves a higher AUC for the normal class (98.0%) but exhibits substantial variability across folds for the benign and malignant classes, as indicated by the larger standard deviations. These results suggest that while cross-attention may aid in distinguishing normal cases, the concatenation-based fusion offers more robust and consistent performance across all classes and folds.

To further explore whether cross-attention performance could be improved, we evaluated lightweight alternatives aimed at stabilizing the fusion process. First, we applied L2 normalization to the embeddings after ArcFace and prior to fusion. This resulted in degraded performance and increased variability across folds, with mean AUCs of $0.70 \pm 0.45$, $0.60 \pm 0.40$, and $0.80 \pm 0.50$ for the normal, benign, and malignant classes, respectively. A possible explanation is that, in a multimodal setting, the magnitude of the feature embeddings carries meaningful information about the relative contribution of each modality. Enforcing a fixed norm may therefore remove important modality-specific cues, particularly given that MG and US exhibit different strengths across classes. We also evaluated a single-head attention mechanism as a simplified alternative to multi-head cross-attention. This approach likewise resulted in inferior performance and high variability, with mean AUCs of $0.70 \pm 0.40$, $0.87 \pm 0.45$, and $0.79 \pm 0.47$ for the normal, benign, and malignant classes, respectively. This may be attributed to the limited representational capacity of a single attention head, which constrains the ability to capture the complex relationships between modalities.

The concatenation MM model achieves high AUC for both normal and malignant cases compared to single-modality models (US for normal, MG for malignant)—but excels by performing well on both simultaneously. It also notably improves AUC for benign cases, where single-modality models struggle, offering strong performance across all classes. Table 3 shows the AUC of the various models and Fig. 5 (left) for the Receiver Operating Characteristic (ROC) curves of the concatenation MM model for one of the folds. Importantly, unlike semantic MM settings such as image–text pairs, US and MG represent physically complementary measurements rather than semantically aligned modalities; therefore, simple feature concatenation provides a more stable and appropriate fusion strategy than cross-attention-based mechanisms that assume explicit cross-modal alignment.

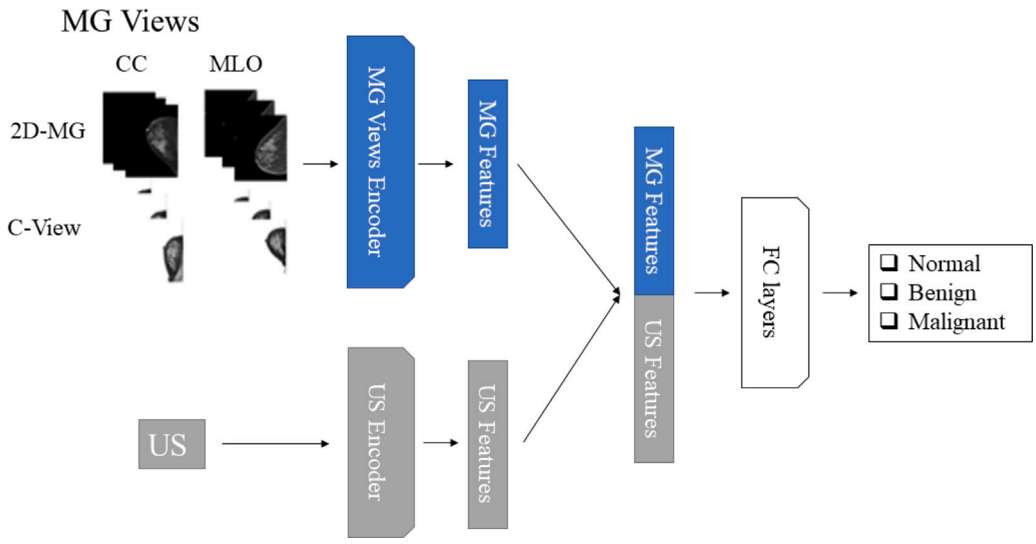


**Fig. 4.** Concatenation MM architecture.

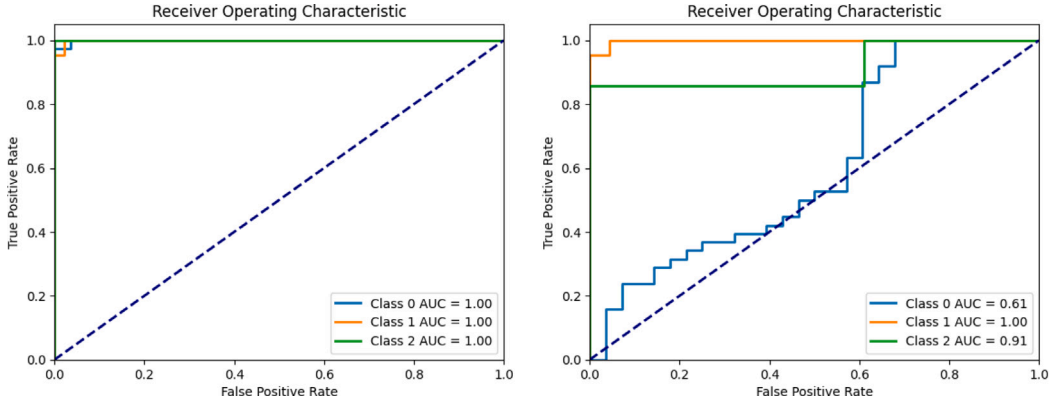


**Fig. 5.** ROC curves for the concatenation MM network on the validation dataset. Normal — Class 0, Benign — Class 1, Malignant — Class 2. The left panel shows results with ArcFace incorporated, and the right panel shows results without ArcFace.

**Table 3**
AUC performance comparison (in %) for all models across the three classification categories: normal, benign, and malignant. Results are averaged over 10-fold cross-validation and reported as mean ± standard deviation.

| | C-View usage | ArcFace incorporation | Normal | Benign | Malignant |
|---|---|---|---|---|---|
| Single-modality US | No | Yes | 91.3 ± 4.7 | 84.6 ± 4.8 | 57.5 ± 24.0 |
| Single-modality MG | Yes | Yes | 67.5 ± 7.1 | 70.0 ± 6.2 | 95.4 ± 7.2 |
| | No | Yes | 57.9 ± 19.1 | 36.7 ± 15.2 | 86.5 ± 17.5 |
| **Concatenation MM** | **Yes** | **Yes** | **95.5 ± 5.2** | **97.5 ± 3.4** | **98.5 ± 3.0** |
| | No | Yes | 50.0 ± 29.7 | 82.5 ± 19.5 | 50.0 ± 11.2 |
| | Yes | No | 82.2 ± 11.7 | 71.4 ± 11.5 | 97.2 ± 4.6 |
| Cross-attention MM | Yes | Yes | 85.3 ± 29.2 | **96.2 ± 4.5** | 79.0 ± 39.8 |
| | Yes | No | **98.0 ± 1.7** | 75.0 ± 39.2 | 87.5 ± 29.2 |
| GoogleNet | Yes | No | 93.0 ± 4.0 | 81.6 ± 9.4 | 96.8 ± 7.3 |
| | No | No | 87.0 ± 5.6 | 57.0 ± 9.9 | 90.0 ± 9.1 |

The confusion matrix in Fig. 6 illustrates the performance of MM with concatenation fusion, aggregated across all 10 folds. Each patient appeared exactly once as validation, making this a true patient-level analysis. The model achieved a recall of 95.7% and a precision of 98.5% for normal class, correctly identifying 335 out of 350 cases. For the benign class, the model achieved a recall of 97.7% and a precision of 92.4%, correctly identifying 208 out of 213 cases. For the malignant class, the model achieved a recall of 97.8% and perfect precision of 100%, correctly identifying 91 out of 93 cases. These metrics highlight strong performance for normal and malignant cases and slightly lower performance for the benign cases. The high recall and precision for malignant cases highlights the model's effectiveness in

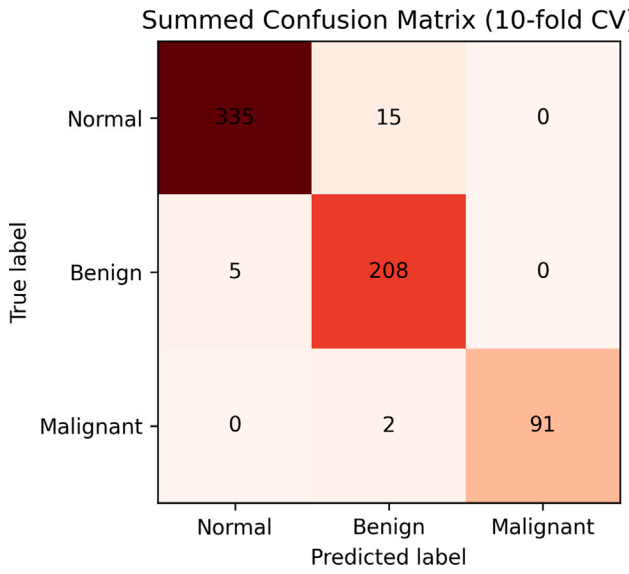


**Fig. 6.** Confusion matrix for the concatenation MM at the patient level, aggregated over the validation sets of all 10 cross-validation folds, with each patient appearing exactly once as validation.

detecting cancerous tumors, though distinguishing between normal and benign cases remains a challenge. This challenge is also well recognized in clinical practice, as radiologists often find differentiating benign findings from normal tissue more difficult than identifying malignant lesions, due to subtle and overlapping imaging characteristics.

The method in [10] achieved 77.77% accuracy for single-modality US but did not report a full confusion matrix or AUC values. In contrast, our ArcFace-enhanced US model reached 86.3% patient-level accuracy across the folds, demonstrating its effectiveness. Our concatenation MM model further improves performance, achieving higher AUC values and an 96.6% accuracy, as shown in the confusion matrix in Fig. 6.

As part of an ablation analysis, we evaluate the contribution of C-View MG input and ArcFace incorporation to the model's performance. We examined the effect of incorporating the C-View MG into the networks. For the single-modality MG approach, including C-View MG led to an AUC of 67.5% for normal, 70.0% for benign, and 95.4% for malignant cases. When C-View MG was not included, performance dropped across all classes, with AUC values of 57.9% for normal, 36.7% for benign, and 86.5% for malignant cases, demonstrating a reduction in performance when C-View was excluded. There is also an increasing in performance variably across the folds, as indicated in higher standard deviation of the AUCs. Our MM approach performed worst without the use of C-View MG, lowering the AUCs to 50.0% for normal, 82.5% for benign, and 50% for malignant cases. We also implemented a model from a previous study designed for binary BC classification (benign and malignant) using a pre-trained GoogleNet refined on clinical data [13]. This model relied on manually annotated MG and US images, requiring physician input for localizing regions of interest. We adapted their approach to classify normal, benign, and malignant cases without using manual localization. This adaptation allowed us to evaluate its performance in a fully automated setting, aligning with the goals of our study. The results demonstrate the benefits of our approach over models dependent on manual intervention. The results were lower compared to our approach, with AUC values of 93.0% for normal, 81.6% for benign, and 96.8% for malignant cases. To evaluate the impact of C-View MG in this MM approach, we also refined the pre-trained GoogleNet on the clinical data excluding C-View MG. In this case, the AUC values were 87.0% for normal cases, 57% for benign cases, and 90.0% for malignant cases. These findings reveal a consistent pattern, showing that excluding C-View MG leads to a decline in performance. Overall, these results demonstrate that our MM approach outperforms the previous study in the absence of annotated data while emphasizing the significant impact of C-View MG information.

Table 3 also demonstrates the significant impact of applying the ArcFace layer on the performance of our MM methods. For the concatenation fusion, AUC values with ArcFace are 95.5% (normal), 97.5% (benign), and 98.5% (malignant), compared to lower values of 82.2%, 71.4%, and 97.2% without ArcFace. For the cross-attention fusion, AUC values with ArcFace are 85.3 (normal), 96.2% (benign), and 87.5 (malignant), while without ArcFace the AUC for normal and malignant cases increase to 98.0% and 87.5%, respectively. The AUC for benign dropped to 75%. Note that for cross-attention fusion there are big variability thought the folds. Fig. 5 displays the ROC curves for the concatenation-based MM, comparing performance with ArcFace (left) and without ArcFace (right) on the same cross-validation fold. The absence of ArcFace leads to notably poorer AUCs, particularly in correctly identifying normal cases. While the AUC for normal cases in this specific fold is lower than the average AUC over all folds without ArcFace, this illustrates the increased variability across folds when ArcFace is not used. Unlike the conventional approach of integrating ArcFace into the loss function, we deliberately applied the ArcFace layer to the high-dimensional feature vectors generated by each modality before fusion. This mapping to a more discriminative manifold enhances the separability of features, improving the overall classification.

To interpret the model's predictions, we applied Grad-CAM [25], which highlights the image regions most influential to the classification decision. This is clinically significant for interpretability, as the model is expected to focus on meaningful anatomical regions such as lesions in benign or malignant cases across both modalities, rather than on irrelevant image areas or artifacts that are not clinically relevant. Given a set of input images from a single patient, $I = \{I_{\mathrm{US}}, I_{\mathrm{MG}_1}, I_{\mathrm{MG}_2}, I_{\mathrm{MG}_3}, I_{\mathrm{MG}_4}\}$, consisting of one US and four MG views, Grad-CAM generates class-discriminative heatmaps by computing the gradient of the class score $y^c$ with respect to the final convolutional feature maps $A^k$:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{5}$$

where $Z$ is the spatial size of the feature map, and $\alpha_k^c$ represents the importance of feature map $A^k$ for class $c$. The heatmap is then computed as:

$$L^c = \mathrm{ReLU}\left( \sum_k \alpha_k^c A^k \right). \tag{6}$$

We applied Grad-CAM to the single-modality branches within our MM concatenation model, focusing on the final residual block of the deepest convolutional layer in ResNet50. This layer captures high-level spatial features while maintaining localization accuracy. Fig. 7 presents Grad-CAM visualizations for a malignant patient, showing the US, MG, and their respective heatmaps. For the benign case (Fig. 8), the MG clearly highlights a localized region of interest corresponding to the lesion. Meanwhile the US focuses on a region in the upper part of the image and extension to the finding, marked in the figure. For the normal case (Fig. 9), the MG shows mild sensitivity to an area slightly lower and outside the central breast region, whereas the US emphasizes a region outside the chest wall, clearly not representing a breast cancer—consistent with the model's strong performance on normal cases. The visualizations correspond to the malignant class, illustrating that the model primarily attends to suspicious regions, reinforcing its interpretability and clinical relevance.

The results highlight the distinctiveness of our proposed architecture, where the creation of a discriminative space via the ArcFace transformation, combined with the inclusion of C-View MG, enables high performance without reliance on manual localization techniques from physicians. This approach streamlines the diagnostic process, achieving significantly high AUC values for a more complex classification problem that includes normal cases as well. These advancements underscore the robustness and efficiency of our method in handling MM information.

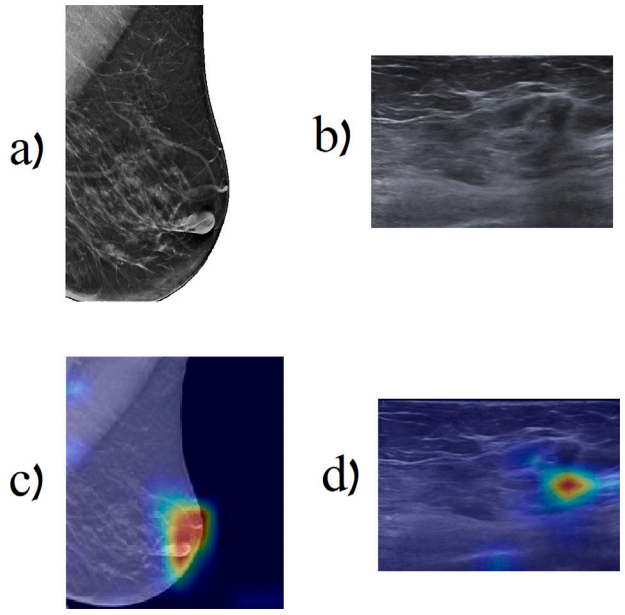


**Fig. 7.** Visualizations for a malignant patient. (a) MG image, (b) US image, (c) Grad-CAM for the MG image, and (d) Grad-CAM for the US image, highlighting the suspicious area identified by the model.

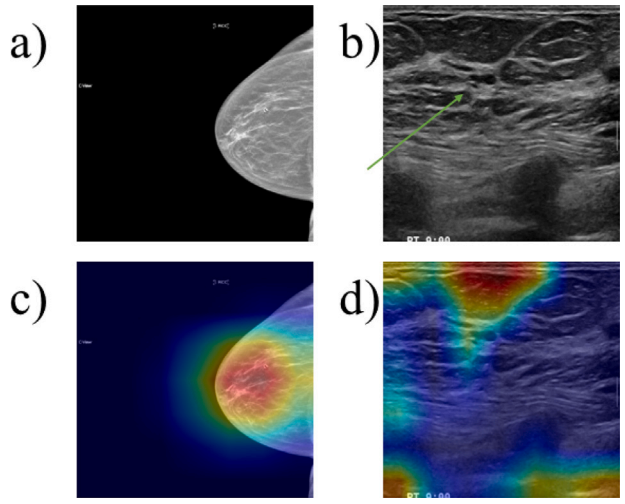


**Fig. 8.** Visualizations for a benign patient. (a) MG image, (b) US image with green arrow indicating the radiologist-marked finding, (c) Grad-CAM for MG, (d) Grad-CAM for US, showing model cross-attention on suspicious regions.

## 5. Conclusion

In this study, we introduced a fully automated MM network for BC classification, leveraging MG and US images without requiring any physician input. Our approach classifies cases into three distinct categories: normal, benign, and malignant, offering a comprehensive diagnostic solution. A key innovation of our work is the integration of C-View MG, a synthesized MG format not previously utilized in BC classification. Additionally, we incorporated a contrastive layer to enhance class separation, further improving the model's discriminative capabilities. Our MM framework effectively combines the strengths of both modalities: MG excels at identifying malignant cases, while US is particularly effective in detecting normal cases. This synergy enabled our MM model, using concatenation for feature fusion, to outperform the single-modality at the classification task. Our method is also capable of accurately detecting benign cases better than single-modalities.

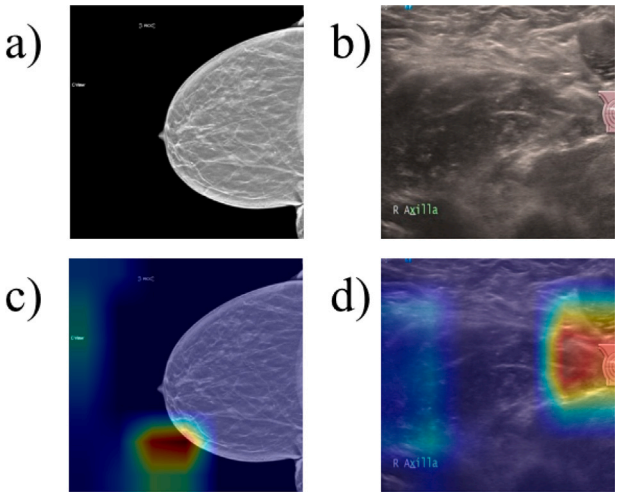


**Fig. 9.** Visualizations for a patient without findings (normal case). (a) MG image, (b) US image, (c) Grad-CAM for MG, (d) Grad-CAM for US, showing model cross-attention in low-suspicion regions.

The proposed MM approach demonstrates significant potential for real-world applications, paving the way for fully automated and reliable BC diagnostics.

A limitation of this study is that the dataset was collected from a single medical center, which may affect generalizability to other institutions. However, the dataset includes variability across imaging devices and acquisition settings, which partially mitigates this limitation.

## CRediT authorship contribution statement

**Yhonatan Kvich:** Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis. **Adi Kalamaro:** Validation, Software, Methodology, Investigation. **Shachar Ashkenasy:** Visualization, Validation, Software. **Adi Wegerhoff:** Validation, Project administration. **Yishai M. Elyada:** Validation. **Ahuva Grubstein:** Validation, Data curation, Conceptualization. **Eli Atar:** Validation, Data curation, Conceptualization. **Yonina C. Eldar:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yonina Eldar reports was provided by Israel Science Foundation. Yonina Eldar reports financial support was provided by European Research Council. Yonina Eldar reports financial support was provided by Swiss Society Institute for Cancer Prevention Research at the Weizmann Institute of Science, Rehovot, Israel. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This research was supported in part by the Israel Precision Medicine Partnership (IPMP), in part by the Israel Science Foundation (ISF) under Grant No. 3805/21, in part by European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program Grant No. 101000967 and in part by Swiss Society Institute for Cancer Prevention Research at the Weizmann Institute of Science, Rehovot, Israel.

## Data availability

Data will be made available on request.

## References


[1] M. Fenichel, American cancer society changes breast cancer screening guidelines to reflect analysis of benefits and harms, JNCI: J. Natl. Cancer Inst. 108 (2) (2016) http://dx.doi.org/10.1093/jnci/djw022.

[2] N. Houssami, P. Skaane, Overview of the evidence on digital breast tomosynthesis in breast cancer detection, Breast 22 (2) (2013) 101–108.

[3] A. Gulland, Staff shortages are putting UK breast cancer screening "at risk," survey finds, 2016.

[4] E. Lazarus, M.B. Mainiero, B. Schepps, S.L. Koelliker, L.S. Livingston, BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value, Radiology 239 (2) (2006) 385–391.

[5] P.D. Trieu, N. Borecky, T. Li, P.C. Brennan, M.L. Barron, S.J. Lewis, The impact of prior mammograms on the diagnostic performance of radiologists in early breast cancer detection: A focus on breast density, lesion features and vendors using wholly digital screening cases, Cancers 15 (4) (2023) 1339.

[6] A.D. Lauritzen, M. Lillholm, E. Lynge, M. Nielsen, N. Karssemeijer, I. Vejborg, Early indicators of the impact of using AI in mammography screening for breast cancer, Radiology 311 (3) (2024) e232479.

[7] J.L. Raya-Povedano, S. Romero-Martín, E. Elías-Cabot, A. Gubern-Mérida, A. Rodríguez-Ruiz, M. Álvarez-Benito, AI-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation, Radiology 300 (1) (2021) 57–65.

[8] K. Dembrower, Y. Liu, H. Azizpour, M. Eklund, K. Smith, P. Lindholm, F. Strand, Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction, Radiology 294 (2) (2020) 265–272.

[9] Y. Shen, N. Wu, J. Phang, J. Park, G. Kim, L. Moy, K. Cho, K.J. Geras, Globally-aware multiple instance classifier for breast cancer screening, in: Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10, Springer, 2019, pp. 18–26.

[10] J. Ellis, K. Appiah, E. Amankwaa-Frempong, S.C. Kwok, Classification of 2D ultrasound breast cancer images with deep learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024, pp. 5167–5173.

[11] S.M. Ha, M.-j. Jang, I. Youn, H. Yoen, H. Ji, S.H. Lee, A. Yi, J.M. Chang, Screening outcomes of mammography with AI in dense breasts: a comparative study with supplemental screening US, Radiology 312 (1) (2024) e233391.

[12] K. Atrey, B.K. Singh, N.K. Bodhey, R.B. Pachori, Mammography and ultrasound based dual modality classification of breast cancer using a hybrid deep learning approach, Biomed. Signal Process. Control. 86 (2023) 104919.

[13] G. Habib, N. Kiryati, M. Sklair-Levy, A. Shalmon, O. Halshtok Neiman, R. Faermann Weidenfeld, Y. Yagil, E. Konen, A. Mayer, Automatic breast lesion classification by joint neural analysis of mammography and ultrasound, in: Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures: 10th International Workshop, ML-CDS 2020, and 9th International Workshop, CLIP 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 9, Springer, 2020, pp. 125–135.

[14] J. Cong, B. Wei, Y. He, Y. Yin, Y. Zheng, A selective ensemble classification method combining mammography images with ultrasound images for breast cancer diagnosis, Comput. Math. Methods Med. 2017 (1) (2017) 4896386.

[15] K. Atrey, B.K. Singh, N.K. Bodhey, Integration of ultrasound and mammogram for multimodal classification of breast cancer using hybrid residual neural network and machine learning, Image Vis. Comput. 145 (2024) 104987.

[16] J. Chen, T. Pan, Z. Zhu, L. Liu, N. Zhao, X. Feng, W. Zhang, Y. Wu, C. Cai, X. Luo, et al., A deep learning-based multimodal medical imaging model for breast cancer screening, Sci. Rep. 15 (1) (2025) 14696.

[17] M.T. Mandelson, N. Oestreicher, P.L. Porter, D. White, C.A. Finder, S.H. Taplin, E. White, Breast density as a predictor of mammographic detection: comparison of interval-and screen-detected cancers, J. Natl. Cancer Inst. 92 (13) (2000) 1081–1087.

[18] E. Avera, L. Valentic, L. Bui, Current understanding and distinct features of multifocal and multicentric breast cancers, Cancer Rep. 6 (9) (2023) e1851.

[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[20] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.

[21] J. Zheng, H. Liu, Y. Feng, J. Xu, L. Zhao, CASF-Net: Cross-attention and cross-scale fusion network for medical image segmentation, Comput. Methods Programs Biomed. 229 (2023) 107307.

[22] X. Li, M. Li, P. Yan, G. Li, Y. Jiang, H. Luo, S. Yin, Deep learning attention mechanism in medical image analysis: Basics and beyonds, Int. J. Netw. Dyn. Intell. (2023) 93–116.

[23] L.R. Lamb, C.D. Lehman, A. Gastounioti, E.F. Conant, M. Bahl, Artificial intelligence (AI) for screening mammography, from the AJR special series on AI applications, Am. J. Roentgenol. 219 (3) (2022) 369–380.

[24] Z. Hussain, F. Gimenez, D. Yi, D. Rubin, Differential data augmentation techniques for medical imaging classification tasks, in: AMIA Annual Symposium Proceedings, vol. 2017, American Medical Informatics Association, 2017, p. 979.

[25] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.