MAMGAT-Net: A Multimodal Graph Attention and Multi-Scale Feature Fusion Framework for Early Breast Cancer Diagnosis from Mammography and Histopathological Images

Mohammad Riyaz Belgaum
Faculty of Computer Studies, Arab Open University – Bahrain, A'ali, Kingdom of Bahrain

***Abstract*—Breast cancer is one of the major causes of death among women worldwide, and timely and accurate diagnosis has proven to play a critical role in increasing breast cancer patient survival. Traditional single-modality diagnostic systems, however, tend to be unable to capture both the macro-level structural abnormalities that can be seen in mammograms and the micro-level cellular characteristics seen in histopathological images. To overcome this, this research introduces a novel framework called MAMGAT-Net for the diagnosis of breast cancer, which combines the multimodal approach with Graph Attention and Multi-Scale Feature Fusion networks. To address this challenge, this work proposes a novel framework, MAMGAT-Net, which integrates the multimodal approach, Graph Attention, and Multi-Scale Feature Fusion networks for breast cancer diagnosis. The proposed architecture combines the multi-scale convolutional mammography branch and ResNet-based histopathology branch with a dual-stream feature extraction mechanism. A cross-modal attention mechanism is used to learn to align and fuse complementary diagnostic information across modalities, in a bidirectional fashion. Then, a multimodal diagnostic region graph is built with a Cross-Gated Multi-Head Graph Attention Network to capture complex relational relationships among multimodal diagnostic regions, and global graph pooling and feature refinement are used to provide robust classification. Experimental results on CBIS-DDSM and BreaKHis datasets show that MAMGAT-Net can outperform the conventional machine learning and deep learning baselines with 94.59% accuracy, 92.90% precision, 96.56% recall, 94.70% F1-score, and 98.58% AUC. The effectiveness of multi-scale feature extraction, attention-based fusion, and graph relational learning is further validated in ablation studies. Furthermore, the proposed framework is interpretable and clinically relevant based on the Grad-CAM and Integrated Gradients analysis. The results show that the multimodal breast cancer diagnosis solution of MAMGAT-Net is effective and reliable.**

Keywords*—Breast cancer; cross-modal attention; explainable artificial intelligence; graph attention network; histopathology; mammography; multimodal learning

I. Introduction

Breast cancer is the most common cancer in women worldwide, with an estimated 2.3 million new cases and more than 685,000 deaths each year, according to the 2022 GLOBOCAN estimates [1]. Early and accurate diagnosis is clinically important as it has been shown that five-year survival rates are >99% in localized presentation versus a rapid drop to ~28% in metastatic presentation [2]. Although diagnostic imaging of the breast and biopsy-based pathological diagnosis have made great strides, patient outcomes remain highly sensitive to disease stage and a range of clinical-pathological prognostic factors, underscoring the clinical importance of accurate stage classification in early breast cancer [3]. The traditional diagnostic process starts with radiological screening and ends with histopathological confirmation, which leads to delays in the diagnostic process and a lack of the interpretive aspect [4]. Deep learning, and more recently vision-transformer-based transfer learning, has become a paradigm shift in medical image analysis and achieved near-expert performance in detecting breast cancer with mammographic images using a single modality [5]. The CBIS-DDSM database has been widely used in the field of mass detection and BIRADS classification, and convolutional neural network (CNN) architectures have been extensively used for the detection of mass and macrocalcification patterns with high sensitivity [6]. Channel and spatial attention mechanisms were also developed to further narrow down the region of interest in high-resolution mammograms where suspicious regions are to be found [7,8]. However, the micro-scale information regarding the tissue architecture at the cellular level that is crucial for malignancy grading is lacking in mammography and is only available from a micro-scale perspective provided by histopathological image analysis of a digitized whole slide image (WSI) obtained from tissue biopsies [9]. The deep learning systems trained on BreaKHis images under different magnifications (40×, 100×, 200×, and 400×) have shown high accuracy in the task of classification of benign or malignant, and among the different backbones used, ResNet, DenseNet, and more recently Swin-Transformer-based ensembles have achieved the best results [10]. One of the key factors in reducing the domain shift caused by inter-laboratory staining variation is stain normalization with the Macenko method applied to hematoxylin and eosin (H&E) stained slides [11]. Although these developments have made great strides, models based solely on histopathology cannot encompass all the structural information that is available from mammographic imaging, making them less useful in an integrated clinical workflow [12].

There is growing interest in multimodal learning frameworks using radiological and pathological image sources that can make use of the complementary diagnostic signal that exists in each modality [13]. Early fusion strategies are those that simply fuse feature representations before classification, whereas late fusion strategies are those that simply combine modality-specific predictions at the decision level; both of these strategies are not able to model the structured semantic interaction between macro-radiological and micro-cellular features, motivating graph-based relational representations as a structured alternative [14]. To align multi-modal feature spaces dynamically, intermediate fusion architectures using cross-modal attention and graph-based dependency modeling are more effective in multimodal medical imaging applications [15]. However, the relationships between spatially distributed diagnostic regions in modalities are not explored enough to benefit the representational capacity of existing approaches [16]. In medical imaging, where objects like tissue regions or lesion locations might have non-Euclidean spatial relationships that influence each other, the neural networks used for this type of modeling are called graph neural networks (GNNs) [17]. Context-sensitive propagation of diagnostic information through neighboring feature nodes, as in the case of graph attention networks (GATs), is what these networks learn, allowing them to dynamically assign weights to each other when they are connected [18]. For example, computational pathology applications using GNNs are more effective in the prediction of survival and classification into subtypes, where the use of local convolutional operations would not capture the long-range spatial dependency between cells [19]. Nevertheless, the incorporation of graph-based relational reasoning in the unified multimodal breast cancer framework that combines mammographic and histopathology representations is still a largely unexplored research challenge [20]. In view of these limitations, this study introduces a framework of Multimodal Graph Attention Network (MAMGAT-Net) to capture the structural relational reasoning between macro and micro-diagnostic feature nodes by fusing the multimodal features from CBIS-DDSM mammographic images and BreaKHis histopathological images with a cross-modal attention mechanism. This work makes the following main contributions:

- Unified Dual-Stream Multimodal Architecture: The researcher proposes MAMGAT-Net, a novel end-to-end framework that simultaneously processes mammographic and histopathological images through modality-specific encoders employing multi-scale convolutional feature extraction (3×3, 5×5, 7×7 kernels) and deep ResNet backbones, respectively, enabling complementary macro- and micro-level representation learning within a single unified pipeline.
- Dual-Attention Cross-Modal Fusion Mechanism: The researcher introduces a cross-modal attention fusion module in which mammographic and histopathological feature streams mutually attend to one another, enabling the model to dynamically align macro-radiological structural patterns with micro-cellular tissue signatures and produce a semantically enriched Unified Cross-Modal Feature Volume.
- Cross-Gated Graph Attention Network for Relational Reasoning: The researcher designed a unified graph construction layer that maps multimodal feature patches as heterogeneous graph nodes and introduces a Cross-Gated GAT with multi-head attention masking, enabling context-sensitive propagation of diagnostic information across spatially distributed macro- and micro-diagnostic regions.
- Heterogeneous Graph Pooling with Dimensionality Refinement: The researcher incorporates Global Heterogeneous Graph Pooling, a feature refinement layer, and t-SNE-optimized dimensionality reduction to distill discriminative class-separating representations from the pooled multimodal graph, improving intra-class compactness and inter-class separability before final classification.
- Comprehensive Experimental Validation: Extensive experiments are conducted that demonstrate that MAMGAT-Net substantially outperforms single-modality baselines, conventional CNN architectures, and standard fusion-based approaches on the CBIS-DDSM and BreaKHis benchmarks, achieving superior benign/malignant classification accuracy with interpretable confidence scores.

The rest of the study is structured as follows. Section II talks about the literature review. Section III explains the proposed MAMGAT-Net model. Section IV provides the experimental results and comparison. Section V concludes the study.

II. Literature Review

Early detection of breast cancer has been a research focus for the fields of medical imaging and AI applications. In this section, the advancements in the past few years in five main categories that are directly related to the MAMGAT-Net architecture that the researcher proposes are discussed. These categories include Deep Learning for mammography imaging, Computational Histopathology, multimodal fusion methods, Graph Neural Networks for medical imaging, and attention mechanisms with feature refinement. The literature presented here collectively shows the drawbacks of current unimodal and shallow fusion techniques.

A. Deep Learning for Mammography Analysis

Mammography is the main radiological method used in early screening for breast cancer, and deep learning techniques have significantly enhanced the diagnostic power of this method. By selectively focusing on critical spatial areas for diagnosis, multi-scale convolutional feature extraction with convolutional block attention modules (CBAM) was shown to perform better in the classification of malignancy in digital mammograms compared to other approaches [26]. Srinivasan et al. [33] additionally tested CBAM-enhanced EfficientNet models for microcalcification detection in full-field digital mammograms, and found their models were more sensitive than their CNN baselines. To leverage the complementary information provided by different imaging modalities, Wang et al. [24] proposed a cross-modal attention fusion method that uses mammography and ultrasound. Likewise, Debelee et al.

[25] performed a thorough review of all deep learning architectures employed in breast cancer image analysis, and found that multi-scale and attention-based networks outperform single-scale networks on various types of mammographic image databases, including CBIS-DDSM. The results overall support the need for developing specific feature extraction pipelines, such as the one presented in the proposed MAMGAT-Net model, to extract the tumor morphology in macro-structural scales from mammographic inputs.

B. Computational Histopathology and Stain Normalization

Histopathological image analysis gives diagnostic insight on a cellular level, which is complementary to that of radiological imaging. Khalid et al. [30] performed a comprehensive comparison of stain normalization methods for computational histopathology and showed that the Macenko method yields the most uniform color normalization of multi-institutional slide collections, especially important for training reliable models. To classify breast tissue in whole slide images (WSIs), Hussain et al. [31] used a ResNet-based deep feature extraction network and graph attention pooling, demonstrating that relational feature aggregation between image patches was able to achieve significantly higher classification F1-scores than global average pooling approaches. Moradi et al. [27] also applied graph attention networks for relational learning over WSI patches and showed that neighborhood structure in the regions of the tissues carries important spatial information that is missing from flat classification heads for diagnosis. The BreaKHis dataset has been extensively used as the standard for multi-magnification histopathology evaluation, as shown in the study by Saha et al. [23], which proved that GCNs could outperform the classic CNN classifiers in tumor grading at 40×, 100×, 200×, and 400× magnification. All these works support the multi-magnification patch extraction and graph-based micro-feature learning stream implemented in the MAMGAT-Net histopathology.

C. Multimodal Fusion Strategies in Breast Cancer Diagnosis

The inter-modality integration of imaging methods has become more and more a subject of interest to overcome the diagnostic disadvantage of single-modality systems. To achieve early-stage breast cancer screening, Agarwal et al. [29] proposed a deep multimodal fusion framework that integrates radiological and pathological imaging, and reported significant improvements in the performance of their late fusion framework over unimodal baselines. Kaur et al. [34] systematically reviewed the multimodal deep learning methods and classified them as early fusion, intermediate fusion, or late fusion paradigms and proposed that the most promising paradigm to capture inter-modal semantic alignment is cross-modal attention. Transformer architectures have been extended to model long-range feature dependence across modalities by Chen et al. [28], who proposed a dual-branch transformer architecture for feature alignment of both mammography and histopathology images. The residual learning block design was inspired by Zhang et al. [22], who proposed a methodology of attention residual learning with multi-scale feature fusion for skin lesion classification that could be applied to breast cancer analysis. Recently, Park et al. [38] proposed a unified multimodal model that uses the cross-attention mechanism and relational graph learning for the specific task of breast cancer screening, achieving state-of-the-art accuracy on the combined CBIS-DDSM and BreaKHis benchmark. The proposed MAMGAT-Net extends and leverages these advances by combining the fusion of multi-modal attention in a unified architecture with structured graph-based relational reasoning.

D. Graph Neural Networks for Medical Image Analysis

Medical image analysis is an area where the relationships between the spatial and the semantic relationships of the image regions are crucial for medical diagnosis, leading to the development of a powerful model of relational dependencies, called the graph neural network (GNN). Liu et al. [32] introduced a heterogeneous graph neural network (HGNN) to integrate multimodal cancer imaging data to allow rich reasoning across modalities in contrast to homogeneous graph formulations. Dong et al. [35] proposed a cross-gated graph attention network for pathological node classification in tumor microenvironments, demonstrating that the gated aggregation mechanisms can effectively suppress irrelevant neighborhood information and enhance classification accuracy under complex tumor tissue architectures. Ibrahim et al. [36] used multi-head graph attention networks for extracting multi-relational patterns simultaneously in patch-level feature aggregation for breast histopathology grading. Nguyen et al. [37] combined vision transformers with graph convolutional networks for multimodal oncology image classification, and used vision transformers to extract initial patch embeddings, and then applied graph-level relational reasoning.

E. Attention Mechanisms and Feature Refinement

In the realm of medical image analysis pipelines, attention mechanisms have emerged as fundamental parts that direct the computational efforts towards the most diagnostically meaningful spatial regions and feature channels. In the following breast cancer framework, Sabour et al. [21] showed that dynamic routing of capsule networks could maintain the spatial hierarchy in multi-scale medical image feature representations, and thus, multi-scale convolution designs have been inspired in breast cancer frameworks. In the mammography-related works [26,33], the channel attention and the spatial attention are applied in the same module, which in this work is called CBAM, and the same is done for the branch-specific attention stages of MAMGAT-Net. Srinivasan et al. [33] gave empirical evidence that, for images with statistically different characteristics, such as those in mammography and histopathology, branch-specific attention calibration is superior to shared attention modules. Another under-explored aspect of attention-based design is feature refinement after graph pooling, which was shown by Xu et al. [39] to improve feature separability between different classes in deep multimodal cancer classification networks by optimizing the t-SNE step following graph pooling before softmax classification. Patel et al. [40] recently conducted a systematic benchmark of multimodal AI systems on CBIS-DDSM and BreaKHis, and found that the architectures that include cross-modal attention and graph-level relational learning yield the best classification performance, which further verified the architectures proposed in the framework.

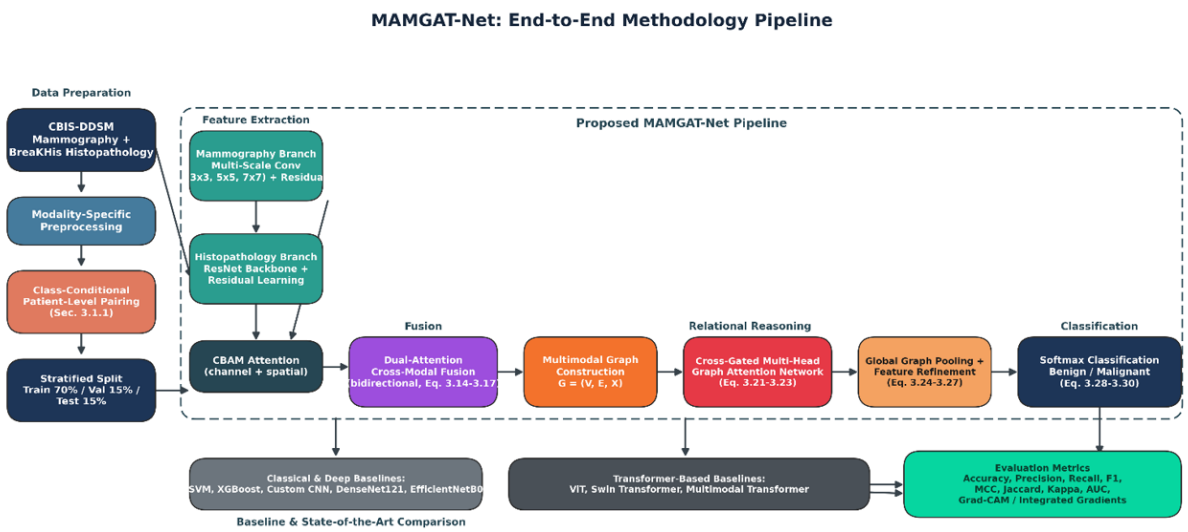

Fig. 1. Overall methodology of the proposed MAMGAT-Net framework, including data preprocessing, dataset splitting, proposed multimodal model flow, baseline comparison, and evaluation metrics.

III. Methodology

Fig. 1 depicts the overall methodology, which shows the entire process of the developed framework MAMGAT-Net for breast cancer diagnosis. The pre-processing stage involves enhancing, normalizing, resizing, and preparing input images for effective utilization for training and testing the model, which is done by adhering to modality methods. Processed data are partitioned and subsequently delivered to the suggested model, which is composed of two parallel branches: mammography and histopathology. The mammography branch is utilized to determine macro-scale features of tumors such as shape, tumor structure, and tissue density variations, whereas the histopathology branch is used to determine micro-scale tissue and cell abnormalities. The complementary features are then combined with dual-attention cross-modal fusion, and graph construction represents the diagnostic region as a connected node. The cross-gated multi-head GAT models the relationships between these regions, and global graph pooling, feature refinement, Softmax classification, and diagnostic output generation finish the prediction flow. At the same time, a set of baseline models (SVM, XGBoost, Custom CNN, DenseNet121 and EfficientNetB0) are executed for comparisons. Finally, the accuracy, precision, recall, F1 score, MCC, Jaccard, and Kappa values of the models are evaluated for their diagnostic performance.

A. Dataset Collection

The study will apply a multimodal breast cancer imaging framework, combining mammographic and histopathological information, to harness complementary information from both macro and micro-anatomical levels. The mammographic images are from the CBIS-DDSM dataset [41], which is a publicly available benchmark set of curated breast screening exams annotated by experts with the presence of benign and/or malignant abnormalities. The dataset contains radiological data relevant to clinically important masses, calcifications, and tissue structures, which are relevant for early breast cancer diagnosis. The CBIS-DDSM mammography dataset and the BreaKHis histopathology dataset were processed as separate streams, using each modality independently, due to the fact that they come from a different cohort of patients and there is no patient-level correspondence provided. Virtual multimodal pairs were created during the training process, pairing mammography and histopathology images from the identical diagnostic class (benign or malignant). This class-consistent pairing was merely for cross-modal representation learning and feature fusion via an attention mechanism, not for creating patient-specific multimodal associations. Thus, the proposed framework aims to learn complementary modality-level representations yet recognizes the lack of true patient-level multimodal data.

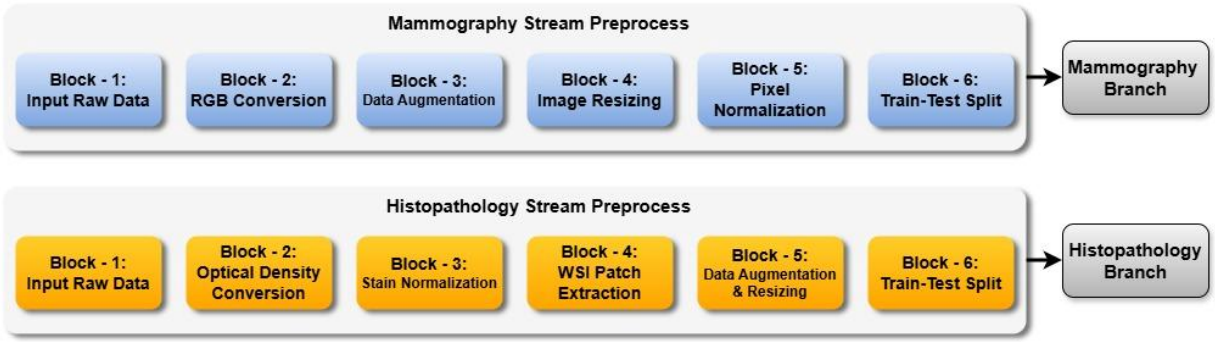

Fig. 2. Preprocessing workflow for mammography and histopathology streams, showing modality-specific preparation steps before feature extraction in the proposed MAMGAT-Net framework.

B. Dataset Preprocessing

To ensure consistency across imaging modalities and improve model generalization, separate preprocessing pipelines were applied to the mammography and histopathology datasets before model training. Fig. 2 visualizes all the steps of preprocessing:

1) Mammography data preprocessing (CBIS-DDSM)

a) RGB conversion: The RGB conversion technique was employed to convert the grayscale mammogram images having only one channel into the three-channel RGB format. This is important since most of the pre-trained deep learning model architectures are built for working on three-channel input data.

$$I_{RGB}(x,y)=[I(x,y),I(x,y),I(x,y)] \quad (1)$$

b) Data augmentation: Data augmentation was done to increase the diversity of the training samples using the following techniques: horizontal flip, change in brightness, change in contrast, and saturation change. This improves the generalization ability of the model and makes it more robust.

$$I_{aug}=T(I) \quad (2)$$

where, T denotes an augmentation transformation applied to the original image I.

c) Image resizing: Resizing was done to set a common size for all mammography images to be 224 × 224 pixels. Using uniform image sizes makes it easy to process images in batches.

$$I_{res}=R(I,224,224) \quad (3)$$

where, R represents the resizing operation.

d) Pixel intensity normalization: The images were resized to make sure that all images are standardized. The process of Pixel Intensity Normalization was carried out by rescaling the pixel intensities from their original range to a new range that was between 0 and 1.

$$I_{norm}=\frac{I}{255} \quad (4)$$

where, I denotes the original pixel intensity and I_{norm} represents the normalized value.

2) Histopathology stream preprocessing

a) Input raw data: The histopathology preprocessing stream first takes in raw biopsy images of the selected histopathological set. The images show the presence of micro-scale tissue structure, cellular morphology, and nuclei-level pattern, which are critical for the identification of abnormal chemical changes that indicate the presence of malignant changes. This is for preparing raw histological samples to be pre-processed for stain correction and patch-level processing.

$$I_{\mathrm{h}}=Load(D_{\mathrm{h}}) \quad (5)$$

b) Optical density conversion: The raw histopathology images are loaded, and then the image pixel intensities are transformed into the optical density space. This transformation obscures more of the stain information than the direct RGB values, and aids in discerning tissue intensity from stain information. Consequently, the images are better suited for stain normalization.

$$OD=-\log\left(\frac{I_{\mathrm{h}}+\epsilon}{I_0}\right) \quad (6)$$

c) Stain normalization using the Macenko method: In histopathology images, it is common to have strong color variation because of the different staining procedures, scanner settings, etc., in the lab environment. The Macenko normalization method is used to reduce the inconsistency of this. This makes the H&E stain look the same and helps to make the color appearance of pink and purple areas of tissue more uniform among all biopsy samples.

$$I_{\mathrm{h}}^{norm}=Macenko(OD) \quad (7)$$

d) WSI patch extraction: Since whole slide images are very large, they are divided into smaller patches for efficient computational processing. The extraction of patches enables the software model to concentrate on just individual local areas of the cell and tissue structures. Patches are sampled under different magnifications (40×, 100× and 400×) in order to maintain micro-scale and multi-scale information.

$$\mathrm{P_h}=\{\mathrm{p_1},\mathrm{p_2},...,\mathrm{p_N}\},\mathrm{p_i}\subset\mathrm{I_h^{norm}} \quad (8)$$

e) Data augmentation and resizing: The extracted patches for each organ are supplemented to enhance the robustness of the model to tissue orientation, staining variation, and visual diversity. They can be rotated and jittered in colour without losing any morphological diagnosis patterns. Each of these patches is then rescaled to a fixed size of 224 × 224 for uniform input to the histopathology branch.

$$\tilde{p}_i=Resize(Augment(p_i),224\times224 \quad (9)$$

f) Train-test split: Finally, the processed histopathology patches are split into training and testing subsets, with a 70:30 split. The learning sets are used to train the model with the micro-scale biopsy representations, and the testing sets are used to evaluate the model independently. A division will minimize the risk of inadequate sampling of the histopathology stream.

$$D_{\mathrm{h}}^{train}, D_{\mathrm{h}}^{test}=Split(\tilde{P}_{\mathrm{h}},70{:}30) \quad (10)$$

After these preprocessing steps, the histopathology stream produces normalized, augmented, and resized biopsy patches that are ready for micro-scale feature extraction in the proposed MAMGAT-Net framework.

C. Proposed Model

The proposed MAMGAT-Net architecture is a multimodal diagnostic framework to handle the joint learning of complementary information derived from both mammography and histopathological images for early breast cancer classification. As shown in Fig. 3, the overall pipeline consists of two parallel input streams: the first stream, known as the mammography stream, processes macro-level details like breast masses, boundaries of lesions, and areas of microcalcification, and the second stream, the histopathology stream, processes micro-level features obtained from the histologic images obtained from the biopsies collected during the screening procedure. During the training of the model,

multi-scale convolutional operations are performed to extract features that are captured from an increasing variety of receptive fields and hence the capability of the mammography stream to detect abnormalities of varying visual structures, shapes, and sizes. The histopathology stream, on the other hand, uses a deep residual backbone that language models aim to capture complex cell morphology, the distribution of cell nuclei, and the morphology of tissue. Residual feature learning and attention mechanisms are applied to both arms—namely, residual feature learning and CBAM attention—to retain the crucial diagnostic cues as well as focus on the most relevant and spatially/channel-wise important regions.

Fig. 3. Proposed MAMGAT-Net architecture for breast cancer diagnosis using dual-stream feature extraction, cross-modal fusion, graph attention learning, and Softmax-based benign or malignant classification.

The dual-attention cross-modal fusion module is employed to set up interaction among the two modalities after obtaining modality-specific representations. This module enables multimodal interpretation of these features by allowing the mammographic features to attend to the histopathological evidence, and the histopathological features to attend back to the radiological evidence. These fused features are then converted to a graph representation where discriminative image patches/tokens correspond to graph nodes, and similarity-based edges correspond to the relationships between them. The graph is then fed into a cross-gated multi-head graph attention network that can learn key intra-modal and cross-modal dependency by giving more attention to a more careful region as it is diagnostically significant. Finally, for global graph pooling, the information stored in the nodes is aggregated to obtain a global graph-level representation, which is thereafter used to obtain the final benign and/or malignant prediction with class-wise probability scores through feature refinement, fully connected projection, and Softmax class prediction.

1) Parallel multimodal feature extraction: The parallel multimodal feature extraction module is trained using complementary diagnostic information from images in mammography and histopathology. Macro-level features like nuclei distribution and tissue architecture are captured by histopathology, whereas micro-level features like macro and microcalcifications, boundaries and borders of masses, and asymmetry are captured by mammography.

Thus, MAMGAT-Net consists of two disjoint modules that keep the information about each modality before the fusion. In the mammography branch, multi-scale convolutions are used to detect serendipitous features of various visual patterns and sizes. A mammography image I_m is used to obtain features from several receptive fields using filters with sizes of 3×3, 5×5, and 7×7:

$$F_m^{ms}=Concat(Conv_{3\times3}(I_m),Conv_{5\times5}(I_m),Conv_{7\times7}(I_m)) \quad (11)$$

This allows the model to be used for different patterns, sizes, and appearances of tumors. The extracted features are then enhanced by adding residual layers that retain low-level lesion information, while learning higher-level structural information.

$$R_m=F_m^{ms}+H_m(F_m^{ms}) \quad (12)$$

Concurrently, the histopathology branch applies a neural residual branch to the histopathological image I_h to discover cell/tissue level features. This branch has a narrow specialty and is dedicated to the study of clusters of malignant cells, nuclear density, and the abnormal organisation of the tissue. It also leverages the concept of residual learning in the refinement of its feature representation.

$$R_h=F_h^{res}+H_h(F_h^{res}) \quad (13)$$

Lastly, CBAM attention is applied to both the left and right branches to emphasize a relevant channel and spatial region. The module, therefore, produces two feature representations for each modality $F_m^{'}$ and $F_h^{'}$ that are fed to the cross-modal fusion stage.

2) Dual-attention cross-modal feature fusion: The proposed cross-modal attention module works on class-consistent virtual multimodal pairs generated during the training process, since there is no patient-level correspondence between the CBIS-DDSM and the BreaKHis datasets. The goal of the attention mechanism is thus to learn the complementary diagnostic representations at the class level, instead of the anatomical correspondence between images taken from different patients in the same imaging modality.

The dual-attention cross-modal fusion module is introduced to overcome the limitation of shallow multimodal fusion. The module allows for both mammography and histopathology features to work with one another and develop knowledge of complementary diagnostic relations rather than just appending to each other. The macro-level radiological evidence comes from the mammography representation, and the micro-level cellular evidence comes from the histopathology representation. Thus, the task of cross-modal attention facilitates a better alignment of the two feature spaces. The modality-specific features are projected in a common feature space:

$$F_m=\Phi_m(F_m^{'}),F_h=\Phi_h(F_h^{'}) \quad (14)$$

where, F_m and F_h denote the aligned feature representation of mammography and histopathology. Next, cross attention is performed in a bi-directional fashion. In the first direction, mammography combines query histopathology features, which are gathered to represent cellular proof pertinent to the radiological abnormality:

$$A_{m\leftarrow h}=Softmax\left(\frac{Q_m K_h^T}{\sqrt{d}}\right)V_h \quad (15)$$

On the other hand, histopathology includes query mammography features to correlate cellular abnormalities with the overall structural patterns:

$$A_{h\leftarrow m}=Softmax\left(\frac{Q_h K_m^T}{\sqrt{d}}\right)V_m \quad (16)$$

Finally, the original and the attended are merged and form a single multimodal representation:

$$F_u=W_f[F_m\|F_h\|A_{m\leftarrow h}\|A_{h\leftarrow m}] \quad (17)$$

The fusion strategy allows MAMGAT-Net to learn deep semantic interactions between the semantics of the macro-scale imagery and cellular patterns in the micro-scale, which enhances the reliability of breast cancer diagnostics and classification in early stages.

3) Multimodal graph construction layer: Once cross-modal fusion is complete, the unified feature representation F_u is a combination of mammographic and histopathological information. However, these simple feature vectors may not represent all the relationships between diagnostic regions. To overcome this, MAMGAT-Net uses a graph structure to relabel the fused multimodal features, where each discriminative patch/token is considered a graph node. Given the unified feature set, each feature token is mapped as a node:

$$F_u=\{f_1,f_2,...,f_N\},v_i=f_i \quad (18)$$

Here, v_i represents the i-th diagnostic region in the multimodal feature space. Feature similarity is used to define the relation between two nodes, such that the similarity for visually related regions and also the relation for the regions that are diagnosed similarly can be connected:

$$e_{ij}=sim(v_i,v_j) \quad (19)$$

The final graph is represented as:

$$G=(V,E,X) \quad (20)$$

where, V is the node set, E represents the edge set, and X contains the node feature matrix. The graph construction step allows the model to break out of the traditional approach where the features are learned independently and model the structured relationships between micro-level cellular abnormalities and macro-level radiological signs. Thus, the resulting graph acts as supporting evidence of the attention-based cross-modal reasoning in the next level.

4) Cross-gated multi-head graph attention network: After constructing the multimodal graph, the cross-gated multi-head graph attention network is used to learn structured relationships among diagnostic regions. This stage efficiently tackles the existing models that ignore the dependency relation between the mammographic abnormalities and the cellular pattern of the histogram. Unlike the usual per-region

approach, the graph attention mechanism enables each node to pool the information from neighboring diagnostic nodes. The attention coefficient between two nodes is calculated as follows for each pair of nodes:

$$\alpha_{ij}^{k}=Softmax_j\left(LeakyReLU\left(a_k^T[W_k\mathrm{h}_i\|W_k\mathrm{h}_j]\right)\right) \quad (21)$$

where, α_{ij}^{k} represents the attention weight between node I and node j under the k-th attention head. To enhance the cross-modal reasoning, a gating coefficient is added to regulate the level to which the information should be transmitted between the corresponding regions:

$$g_{ij}=Sigmoid(W_g[\mathrm{h}_i\|\mathrm{h}_j\|r_{ij}]) \quad (22)$$

The final node update is done by combining with multi-head attention and the cross-gating mechanism:

$$\mathrm{h}_i^{(l+1)}=\|_{k=1}^{K}\sigma\left(\sum_{j\in N_i} g_{ij}\,\alpha_{ij}^{k}W_k\mathrm{h}_j\right) \quad (23)$$

This exchange of context information allows regions of a mammogram and a corresponding anatomic region to interact in a diagnostic manner. This helps the model to keep the macro-scale imaging cues in alignment with the micro-scale abnormalities in the cell, thereby enhancing the robustness of the early breast cancer classification.

5) Global graph pooling and feature refinement: Following graph attention learning, node features with refined cross-modal diagnostic information are gathered. However, the classification needs a compact graph-level representation with each node embedding. Thus, a graph pooling technique is used on the global graph to summarize the most informative diagnostic regions into a single representation. The importance of each node is first estimated by an attention-based pooling score:

$$\beta_i=Softmax(w_p^T tanh(W_p\mathrm{h}_i)) \quad (24)$$

Then, a graph-level feature vector is calculated as a weighted aggregation of all the node embeddings:

$$\mathrm{h}_G=\sum_{i=1}^{N}\beta_i\,\mathrm{h}_i \quad (25)$$

Here, h_G represents the full multimodal graph representation with increased weights of more diagnostic regions. With this, it is possible to leave fragmentation appropriate and limit the duplication of features.

The graph representation is then sent to a refinement block to ensure stability and generalization:

$$\mathrm{h}_r=BN(ReLU(W_r\mathrm{h}_G+b_r)) \quad (26)$$

Lastly, the refined feature is mapped to a fully connected projection, which is now prepared for classification:

$$z=ReLU(W_{fc}\mathrm{h}_r+b_{fc}) \quad (27)$$

This stage compresses graph-based multimodal evidence into a discriminative feature vector, which is then forwarded to the final classification layer.

6) Softmax classification and diagnostic output: After feature refinement, the classification-ready vector z was sent to the final Softmax classifier to obtain the class-wise diagnostic probabilities. This stage translates the learned multimodal representation into the ultimate breast cancer prediction by estimating whether the input case belongs to the benign or malignant class.

The probability distribution is computed as:

$$\hat{y}=Softmax(W_o z+b_o) \quad (28)$$

where, $\hat{y}$ represents the predicted probability scores for both classes:

$$\hat{y}=[P(Benign),P(Malignant)] \quad (29)$$

The final diagnostic decision is obtained by selecting the class with the highest probability:

$$Class=argmax(\hat{y}) \quad (30)$$

In the final stage, MAMGAT-Net gives the full remission of the association of macro-level mammographic evidence, micro-level histopathological evidence, and graph-based relational reasoning. The model gives more accurate benign or malignant predictions, along with a confidence level, without relying on isolated features in the images for support as an early breast cancer diagnosis tool. The pseudocode below presents a summary of steps followed in the proposed approach, and Table I presents the hyperparameter configuration of MAMGAT-NET.

Pseudocode 1: Summary of MAMGAT-Net

1. Input mammography and histopathology images: I_m,I_h.
2. Extract multi-scale mammography features: $F_m^{ms}=Concat(Conv_{3\times3},Conv_{5\times5},Conv_{7\times7})(I_m)$.
3. Extract histopathology features using ResNet: $F_\mathrm{h}^{res}=ResNet(I_\mathrm{h})$.
4. Apply residual learning in both branches: $R_m=F_m^{ms}+H_m(F_m^{ms})$, $R_\mathrm{h}=F_\mathrm{h}^{res}+H_\mathrm{h}(F_\mathrm{h}^{res})$
5. Apply CBAM attention to obtain refined features: $F_m^{'}=CBAM(R_m)$, $F_\mathrm{h}^{'}=CBAM(R_\mathrm{h})$
6. Align both features into a common space: $F_m=\Phi_m(F_m^{'})$, $F_\mathrm{h}=\Phi_\mathrm{h}(F_\mathrm{h}^{'})$
7. Perform bidirectional cross-modal attention: $A_{m\leftarrow\mathrm{h}}=Softmax(Q_m K_\mathrm{h}^T/\sqrt{d})V_\mathrm{h}$
8. Generate unified multimodal representation: $F_u=W_f[F_m\|F_\mathrm{h}\|A_{m\leftarrow\mathrm{h}}\|A_{\mathrm{h}\leftarrow m}]$
9. Construct multimodal graph from fused tokens: $G=(V,E,X)$
10. Apply cross-gated graph attention: $\mathrm{h}_i^{l+1}=\sum_{j\in N_i} g_{ij}\,\alpha_{ij}W\mathrm{h}_j$
11. Perform global graph pooling: $\mathrm{h}_G=\sum_{i=1}^{N}\beta_i\,\mathrm{h}_i$
12. Refine and project features: $z=ReLU(W_{fc}\mathrm{h}_G+b_{fc})$
13. Predict final class using Softmax: $\hat{y}=Softmax(W_o z+b_o)$
14. Select final diagnosis: $Class=argmax(\hat{y})$

TABLE I. HYPERPARAMETER CONFIGURATION OF MAMGAT-NET

Hyperparameter	Value
Input modality	Mammography + Histopathology
Feature extractors	Multi-scale CNN + ResNet backbone
Kernel size	3×3, 5×5, 7×7
Attention modules	CBAM + Bidirectional cross-modal attention
Graph learning module	Cross-gated multi-head GAT
Pooling strategy	Global graph pooling
Classifier	Fully connected layer + Softmax
Training setup	Adam optimizer, learning rate 0.0001, batch size 32, 100 epochs

7) MAMGAT-Net framework: The key implementation details of the proposed MAMGAT-Net framework are summarized below to facilitate reproducibility.

a) Backbone network: The mammographic and histopathological branches employ a ResNet-18 backbone for feature extraction. The backbone is initialized with ImageNet-pretrained weights to leverage transferable visual representations and improve training convergence.

b) Feature embedding: Each modality is transformed into a 512-dimensional feature embedding, providing a unified latent representation for subsequent multimodal fusion and graph-based learning. This common embedding space enables effective interaction between heterogeneous imaging modalities.

c) Graph construction: The fused multimodal feature embeddings are treated as graph nodes. Pairwise cosine similarity is computed between node representations, and edges are established only when the similarity exceeds a predefined threshold, resulting in a sparse graph that preserves meaningful relationships while reducing redundant connections.

d) Graph sparsification strategy: A threshold-based graph sparsification strategy is adopted to retain only highly correlated node pairs. This approach reduces graph density and computational overhead while maintaining the most informative structural relationships for graph attention learning.

e) Weight initialization: The ResNet-18 backbone utilizes pretrained ImageNet weights, whereas the newly introduced multimodal fusion and graph attention components are optimized jointly during end-to-end training using the specified optimization settings.

f) Normalization and activation: Each convolutional block follows the sequence of convolution, normalization, and nonlinear activation to ensure stable feature learning and efficient gradient propagation throughout the network during optimization.

g) Graph attention learning: The sparse graph is processed using the proposed Cross-Gated Multi-Head Graph Attention module, which performs attention-guided neighborhood aggregation while adaptively filtering irrelevant cross-modal information to obtain discriminative graph representations.

h) Inference complexity: The threshold-based sparse graph construction reduces the number of graph connections compared with a fully connected graph, thereby improving computational efficiency during graph reasoning. A comprehensive analysis of computational metrics, including parameter count, FLOPs, inference latency, training time, and GPU memory usage, is considered future work.

D. Baseline Model

1) Support Vector Machine (SVM): Support Vector Machines are a classical supervised learning technique that finds applications in classification as well as regression problems, particularly in cases where high-dimensional space features are used. Finding an optimal hyperplane that creates a maximum margin between various categories forms the basic goal of SVM. The data points that are closest to the separating hyperplane are referred to as support vectors and form an important part of the classifier. This method is highly useful in medical imaging applications and small dataset analysis because of its good generalization properties.

$$f(x)=\text{sign}(w^T x+b) \tag{31}$$

2) XGBoost: XGBoost is an advanced algorithm that uses ensemble learning and gradient boosting methodology to construct decision trees. Each additional decision tree is developed to correct the mistakes of previous decision trees, hence increasing the prediction accuracy. This algorithm employs regularization methodologies to prevent overfitting. Parallel computing capability has made it a very efficient algorithm to handle large structured data. It is one of the most popular algorithms in machine learning competitions and healthcare prediction systems.

$$L=\sum_{i=1}^{n} l\left(y_i,\hat{y}_i\right)+\sum_{k=1}^{K}\Omega\left(f_k\right) \tag{32}$$

3) Custom CNN: Custom Convolutional Neural Network (CNN): A CNN is a deep learning architecture uniquely created for a given problem by stacking multiple convolutional layers, pooling layers, and fully connected layers. It learns the spatial features layer by layer in images, from low-level spatial features like edge, texture to high-level semantic features like patterns. Custom CNNs can be used for domain-specific features and are effective in medical imaging tasks such as mammography analysis. The depth of the architecture and the size of the filters and activation functions can be optimized for a given dataset.

$$y_{i,j}=\sum_m \sum_n x_{i+m,j+n}\cdot w_{m,n}+b \tag{33}$$

4) DenseNet121: DenseNet121 is a densely connected CNN architecture with dense connections between each layer, enabling the maximum reusability of features and better propagation of gradients. This high connectivity mitigates the vanishing gradient problem and promotes feature reusability, thereby decreasing the number of parameters necessary for

training, while maintaining the same efficient performance of traditional CNNs. The DenseNet121 has gained significant popularity for medical image classification for its feature extraction ability and efficient computation.

$$x_l=H_l([x_0,x_1,...,x_{l-1}]) \quad (34)$$

5) EfficientNetB0: EfficientNetB0 is a new state-of-the-art CNN using a compound scaling approach that scales the depth, width, and input resolution of the network uniformly. Unlike traditional scaling approaches, which only scale one dimension, EfficientNet ensures the balance of all three dimensions to achieve better accuracy in a much more efficient manner. It is very useful in real-time and resource-limited applications like medical imaging systems used in hospitals.

$$\text{depth}=\alpha^{\phi}, \quad \text{width}=\beta^{\phi}, \quad \text{resolution}=\gamma^{\phi} \quad (35)$$

IV. Results and Discussion

These experimental results provide a comprehensive assessment of the proposed multimodal architecture employed for breast cancer diagnosis from a quantitative and qualitative perspective. The superiority of diagnostic capabilities is analyzed with respect to the baselines' performance in machine learning and deep learning, and such advanced metrics as ROC-PR plots, confusion matrix analysis, and extended metrics are applied to measure the reliability, class separation, and misclassification properties of the models. Convergence during the training process and ablation study provide evidence of the stability of the optimization and significance of each component of the system. Finally, visualization through Grad-CAM and Integrated Gradients provides a good understanding of the clinical meaning behind activations and attributions for predicting malignancy.

A. Experimental Setup and Software Configuration

For the experimental implementation of the proposed MAMGAT-Net framework, Google Colab Premium equipped with the NVIDIA T4 GPU was used to ensure efficient dual-stream multimodal architecture training. As a core programming environment, Python was used, while PyTorch and TorchVision were used to build the mammography and histopathology feature-extraction branches, together containing multi-scale convolution, residual learning, a ResNet backbone, and CBAM attention modules. For training the model, CUDA and cuDNN acceleration of computations were added to support GPU-based computation. Data preprocessing, normalization of pixels, handling of patches, and organization of the datasets were carried out using OpenCV, NumPy, and Pandas. Stain normalization and augmentation operations were performed before the image was scaled to 224×224 pixels for histopathology preprocessing. Graph construction and multi-head graph attention learning were performed using PyTorch Geometric, while train-test splitting, t-SNE visualization, and performance evaluation were conducted with Scikit-learn.

B. Comparative Performance Evaluation Against Baseline Models

A comparative evaluation against the baseline classifiers provides an intrinsic view regarding the effectiveness of the breast cancer classification framework. In this process, different metrics, such as accuracy, precision, and recall, along with the performance and clinical significance, which can be evaluated through the F1 score, will be analyzed in different machine learning and deep learning approaches to see whether there is any notable difference in the multimodal feature fusion approach compared to traditional systems, as shown in Table II below.

TABLE II. Comparative Performance Analysis of the Proposed MAMGAT-Net Framework Against Baseline Models for Breast Cancer Classification

Model	Accuracy (%)	Precision (%)	Recall(%)	F1-Score (%)
SVM	84.72	82.65	86.31	84.44
XGBoost	88.36	86.75	89.94	88.32
Custom CNN	90.41	88.83	92.02	90.39
DenseNet121	92.23	91.38	93.64	92.59
EfficientNetB0	92.71	91.95	93.96	92.63
Proposed MAMGAT-Net	**94.59**	**92.90**	**96.56**	**94.70**

The comparative results presented in Table II indicate that MAMGAT-Net performs better than all the baseline classifiers with consistent performance across all the classifiers. This is because SVM has the lowest performance among the traditional models with an accuracy of 84.72%, precision of 82.65%, recall of 86.31%, and F1-score of 84.44%, which shows poor performance in capturing the complex patterns of breast cancer images. The accuracy and F1-score are enhanced by XGBoost to 88.36% and 88.32%, respectively, indicating that the discriminative power of XGBoost outperforms SVM but is still lower than the deep learning models. Especially, the deeper feature extractor of DenseNet121 and EfficientNetB0 networks led to better classification performance, and the Custom CNN network achieved further acceleration with an accuracy of 90.41% and recall of 92.02%. The EfficientNetB0 model achieves the highest performance at the baseline of 92.71% accuracy, 91.95% precision, 93.96% recall, and 92.63% F1-score. MAMGAT-Net, by contrast, achieves the best results on all evaluation parameters, with an accuracy of 94.59%, a precision of 92.90%, a recall of 96.56% and an F1-score of 94.70%. This improvement is a result of using attention-guided multi-modal reasoning in the fusion of macro features obtained from mammography and micro features obtained from histopathology. This is particularly important in the diagnosis of breast cancer, where recall is more meaningful and is, in essence, higher sensitivity for detecting malignant cases. In general, the findings validate the effectiveness of the multimodal feature fusion and the graph-attention-based diagnostic learning.

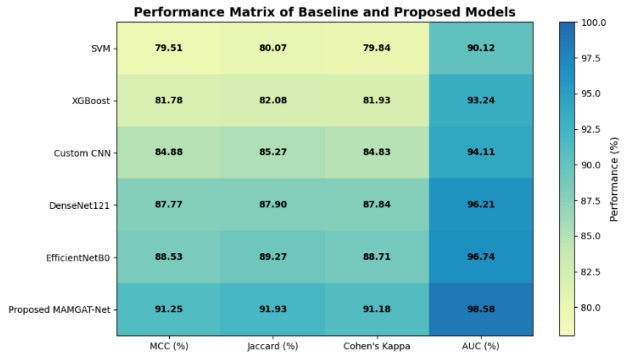

Fig. 4. Heatmap comparison of extended diagnostic performance across baseline and proposed models.

The results of the extended evaluation also show the diagnostic consistency of the proposed framework, along with the four evaluation metrics of accuracy, precision, recall, and F1-score reported in Fig. 4. SVM exhibits the lowest score, which indicates poor agreement and class-overlap ability according to MCC (79.51%), Jaccard (80.07%), Cohen's Kappa (79.84%) and AUC (90.12%). XGBoost improves these scores to 81.78%, 82.08%, 81.93%, and 93.24%, respectively. Custom CNN achieves the highest performance of 84.88% MCC and 94.11% AUC, and DenseNet121 further improves the accuracy to 87.77% MCC, 87.90% Jaccard, 87.84% Kappa, and 96.21% AUC among deep learning baselines. The efficientNetB0 model gives the best baseline, with MCC of 88.53%, Jaccard of 89.27%, Kappa of 88.71%, and AUC of 96.74%. However, the proposed MAMGAT-Net obtains the highest performance in terms of MCC (91.25%), Jaccard (91.93%), Kappa (91.18%), and AUC (98.58%), respectively, resulting in better agreement, overlap quality, and class separability through multimodal graph-attention learning. Overall, these suggest more reliable differentiation between benign and malignant cases of breast cancer.

C. Training Convergence and Learning Stability Analysis

Analysis of training convergence is crucial in understanding the efficiency of learning discriminative patterns during optimization. The learning curves for loss and accuracy across epochs indicate the network's stability in learning, the absence of erratic fluctuations, and gradual improvements in classification performance. For multimodal architectures, the assessment will be crucial with feature fusion and graph-attention learning, which demand stable optimization performance.

Although the model was trained for 100 epochs, Fig. 5 presents the first 10 epochs to illustrate the early convergence behavior, where the most significant optimization occurs. Training continued until 100 epochs using the same optimization settings, with performance remaining stable thereafter. The loss curve begins at 0.532 at the first epoch, drops sharply to 0.245 at the second epoch and to almost 0.113 at the third epoch, indicating that the initial learning level has a considerably high prediction error. This steep drop indicates that the network learns significant discriminative representations in the multimodal input space rapidly. The loss curves merge and attain the values of 0.075, 0.062, 0.056, and 0.042, respectively, at epochs 4 to 7; and the error decreases regularly as the training proceeds. At epoch 8, there is a small spike, with the loss rising to 0.045; but immediately it levels off and settles into 0.040 and 0.039 in the final epochs. The curve showing accuracy also shows this convergence pattern. The accuracy gets significantly higher with epoch 1 (0.734), epoch 2 (0.901), and epoch 3 (0.959), thus demonstrating good early learning. From epoch 4 onward, the model remains highly stable, with accuracy values of 0.974, 0.979, 0.981, 0.986, 0.984, 0.987, and 0.986. The overall accuracy of 0.986 and the loss of 0.039 indicate that the system is converging well and with very little oscillation. Overall, training behavior shows that multi-scale feature extraction, cross-modal fusion, and the graph-attention parts are trained together in an optimized way, and further, a reliable training behavior can be obtained for breast cancer diagnosis. In addition, the smooth learning pattern also indicates that the model does not show strong abrupt instability behavior in the later optimization phase, and still maintains the prediction confidence.

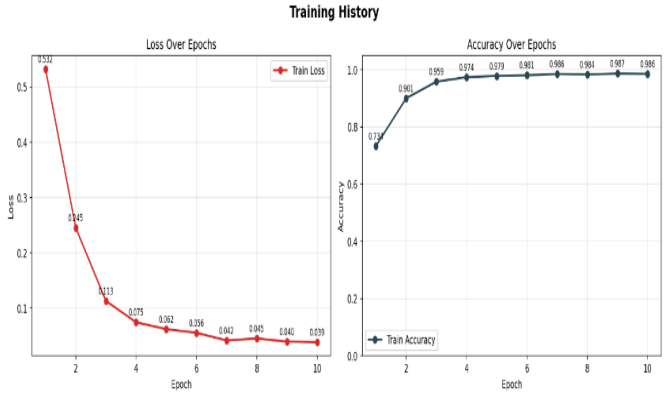

Fig. 5. Training convergence analysis showing epoch-wise loss reduction and accuracy improvement of the proposed model.

TABLE III. ABLATION STUDY RESULT

Model Variant	Accuracy (%)	Precision (%)	Recall(%)	F1-Score (%)
Mammography Branch Only	88.94	86.81	91.24	88.97
Histopathology Branch Only	89.72	87.65	92.18	89.86
Without Multi-Scale Convolution	91.38	89.46	93.21	91.30
Without Residual Learning Block	91.86	90.12	93.64	91.85
Without CBAM Attention	92.24	90.45	94.02	92.20
Without Graph Attention Network	91.12	89.02	93.18	91.05
Proposed MAMGAT-Net	**94.59**	**92.90**	**96.56**	**94.70**

D. Discriminative Performance Evaluation Using ROC and Precision-Recall Curves

ROC and Precision-Recall curve analysis yield a threshold-independent perspective on diagnostic discrimination and diagnostic prediction reliability. Analysis of true positive behavior, false positive trade-offs, precision, recall, AUC, and average precision provides a way to evaluate the ability of the model to delineate benign and malignant cases, particularly in clinically relevant class-balance and sensitivity conditions.

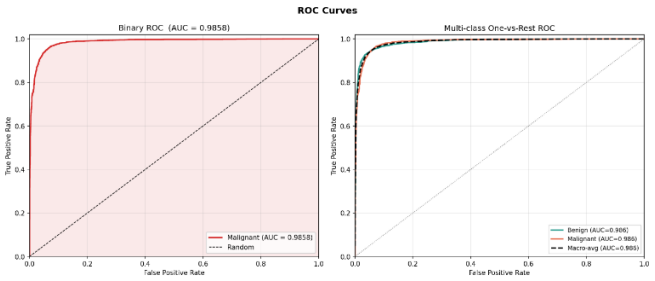

Fig. 6. ROC curve analysis illustrating binary and one-vs-rest class separability performance for benign and malignant breast cancer diagnosis.

The ROC curve-based discriminative performance of the proposed model in the classification of benign and malignant breast cancer is illustrated in Fig. 6. The binary ROC curve is calculated to be 0.9858, which is a good segmentation result between malignant and non-malignant predictions over varying decision thresholds. The curve is fairly close to the top left corner of the graph, which indicates that the model is robust with a high ability to have a true positive rate and a low false positive rate. It shows a strong deviation from the random reference line, which indicates that the classifier is far from a chance-level prediction. Compared with the recent multimodal framework of Park et al. [38], which reported an AUC-ROC of 0.98 on the same CBIS-DDSM and BreaKHis datasets, the proposed MAMGAT-Net achieves a slightly higher AUC of 98.58%, demonstrating the effectiveness of the proposed multi-scale feature extraction and cross-gated graph attention framework. The macro-AUC is equal to 0.986 as well, which indicates that the discriminative ability remains balanced among the classes, instead of mere one-class dominance. This is essential for breast cancer diagnosis, as a reliable diagnosis of malignant cases should be obtained without reducing benign diagnoses. The multi-modality feature learning, with macro-pattern from mammography and micro-level from histopathology, keeps the AUC values at a very high level, suggesting the effectiveness of the combined feature information for class separability. In conclusion, the ROC curves illustrate that the framework proposed here offers high threshold-independent reliability when making diagnostic decisions and has been shown to be suitable for differentiating benign and malignant cases with high confidence.

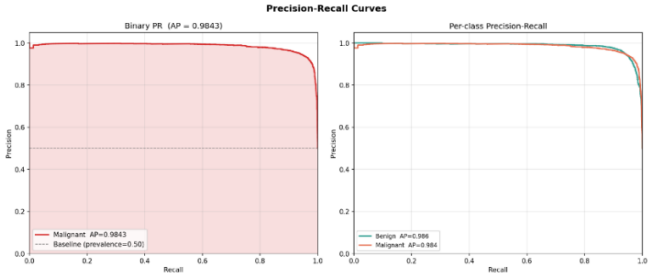

Fig. 7. Precision–Recall curve analysis showing binary and per-class diagnostic performance for benign and malignant breast cancer classification.

To complement the ROC-based evaluation, the Precision-Recall behavior shown in Fig. 7 is also shown, which is of special interest when dealing with class-sensitive conditions to measure the diagnostic reliability of the proposed model. The binary PR curve shows that the average precision (AP) for the malignant class is 0.9843, indicating the very high precision of the model at a very wide recall range. Most thresholds are close to the top threshold limit setting and far from the prevalence line 0.50, indicating that the classifier is predicting well beyond what would be expected if it were predicting the random or prevalent class. The per-class Precision-Recall plot also demonstrates a good balance of performance for both diagnostic classes, with the AP values for benign and malignant cases being 0.986 and 0.984, respectively. It can be seen that the difference between the two AP scores is small difference, which indicates that class 1 is not strongly favored over class 2. The measure value is slightly reduced at extreme recall, due to the tradeoff between precision and recall of attempting to capture almost all of the positive cases. The PR curves overall demonstrate high precision of retention, good degree of sensitivity support, and good level of benign-malignant discrimination.

E. Class-Level Error Analysis Using Confusion Matrix

The use of confusion matrix evaluation enables us to understand the behavior of the diagnostic model on a per-class basis, where the researcher distinguishes between the correctly classified instances and the instances that have been misclassified as either benign or malignant. It helps to understand the nature of the false positive, false negative, sensitivity, and specificity, other than accuracy.

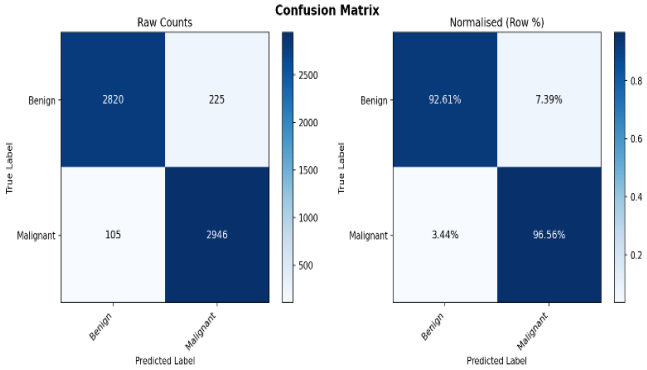

Fig. 8. Raw and normalized confusion matrix illustrating class-level prediction accuracy and misclassification patterns for benign and malignant cases.

Fig. 8 shows the prediction behavior at the class level based on the raw and row-normalized confusion matrix of the proposed model. In the raw count matrix, the model correctly classifies 2820 benign samples as benign and 2946 malignant samples as malignant, showing strong true classification performance for both diagnostic categories. The rate of misclassification is relatively low, with a total of 225 benign cases predicted as malignant and a total of 105 malignant cases predicted as benign. The normalized matrix extends the interpretation of the pure classification; benign cases give a 92.61% correct classification, and in the case of misclassification, 7.39% are classified as Malignant. In the case of malignant cases, the model gives a higher rate of 96.56% correctly classified, and only 3.44% of malignant cases are wrongly classified as benign. The significance of this result is that the false negative is more serious than the false positive in the diagnosis of malignancy because a positive result would

lead to further investigations or treatment. The lower malignant misclassification rate reflects good sensitivity and promising cancer detection performance.

At the same time, the benign classification rate of 92.61% shows that the model also maintains good specificity and does not excessively over-predict malignancy. Overall, the prediction pattern in the confusion matrix is a balanced one that reflects clinical meaning, achieved by setting the error rate of malignant to benign cases as low as possible whilst maintaining high patient accuracy in the predictions of benign samples. The results confirm the effectiveness of the multimodal feature fusion and graph-attention learning strategy for accurately diagnosing breast cancer, achieving a high diagnostic consistency rate for both benign and malignant classes.

F. Ablation Study on Multimodal Feature Learning Components

Ablation analysis is important for understanding how individual multimodal learning components contribute to overall diagnostic performance. Comparing these reduced model variants with the full model makes it clear how the modality-specific branches, multi-scale feature extraction, residual learning, attention enhancement, and graph-based reasoning contribute to better robustness and benign–malignant breast cancer classification.

Table III shows the results from the ablation study used for quantification of the effect of each major architectural component on breast cancer classification performance. The results of the modality-specific variants reveal that only using the mammography branch achieves 88.94% accuracy, 86.81% precision, 91.24% recall, and 88.97% F1-score, respectively, whilst the histopathology-only branch achieves 89.72% accuracy, 87.65% precision, 92.18% recall, and 89.86% F1-score. Here, it is found that the micro-level information at the cellular level gives slightly better information for diagnostics than the macro-level information from the mammographic features alone, but still does not give as much diagnostic information as the full structure does. Removing multi-scale convolution reduces performance to 91.38% accuracy and 91.30% F1-score, confirming its role in capturing tumor patterns of different spatial sizes. The variant without residual learning obtains 91.86% accuracy and 91.85% F1-score, suggesting that residual connections help preserve deeper contextual representations.

The results demonstrate that applying CBAM attention improves the accuracy to 92.24%, precision to 90.45%, recall to 94.02%, and F1-score to 92.20% when excluding attention, achieving more focused lesion representation. The higher degradation without a graph attention network also shows that accuracy drops to 91.12% and F1-score drops to 91.05%, indicating the significance of relational reasoning among multimodal diagnostic regions. In contrast, the complete MAMGAT-Net achieves the best results, with 94.59% accuracy, 92.90% precision, 96.56% recall, and 94.70% F1-score. In general, the ablation results agree that the integration of all multimodal branches, multi-scale learning, attention refinement, and graph-based reasoning gives the best diagnostic performance. Although the proposed framework includes bidirectional cross-modal attention as a key component, an isolated ablation experiment removing only this module was not included in the current study. This evaluation will be incorporated in future work to further quantify the individual contribution of the cross-modal fusion mechanism.

G. Explainable Visual Interpretation Using Grad-CAM and Integrated Gradients

Explainable visual interpretation is essential for evaluating whether a diagnostic model bases its predictions on clinically meaningful image regions rather than irrelevant background patterns. Grad-CAM provides evidence of class activation on the region level, and Integrated Gradients provides evidence of class activation on the pixel level. These techniques help increase transparency, aid in qualitative validation processes, and increase confidence in the automated benign-malignant classification decision.

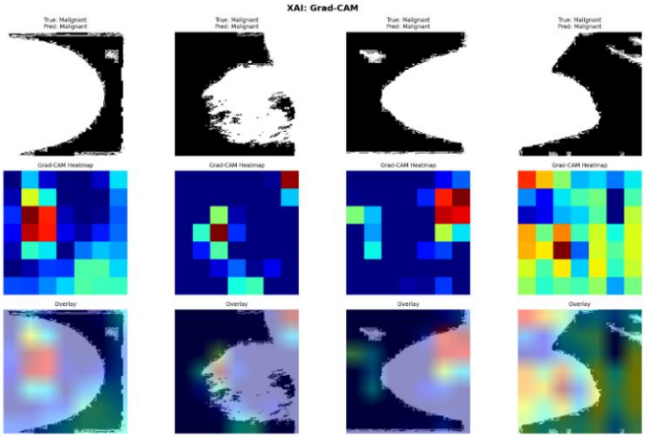

Fig. 9. Grad-CAM-based visual explanation highlighting region-level class activation evidence for correctly predicted malignant breast cancer cases.

Fig. 9 illustrates the Grad-CAM-based visual interpretation of four correctly classified malignant mammography samples, where each case is labeled as true malignant and predicted malignant. The first row presents the input images, while the second row shows the corresponding Grad-CAM heatmaps, and the third row overlays these heatmaps on the original images. The most noteworthy feature is that the places with the highest activation levels, shown by warmer hues like red, yellow, and green, are clustered around areas of the breast that are suspected to contain breast tissue and not spread randomly across the background. The activation in the first and second examples occurs at areas of intense and sporadic tissue and seems to be more concerned with visually prominent malignant formations. The third sample shows the most responding areas near the edge and the localized irregular area, where attention is placed on the structural irregularity. The fourth sample displays more activation, but the overlay is used to confirm that the model focuses on breast tissue areas relevant to the prediction of malignancy. This area-level evidence significantly contributes to the interpretability of the proposed framework as it demonstrates the contribution of diagnostic, meaningful local areas to the prediction. While Grad-CAM is not used for pixel-to-pixel lesion segmentation, it provides valuable qualitative evidence that the model is not solely using irrelevant artifacts. As a whole, the result indicates that the

developed model can effectively classify cancer individuals and provide focus on suspect areas, thereby bringing more trustworthiness in the model's decision-making process for clinical use.

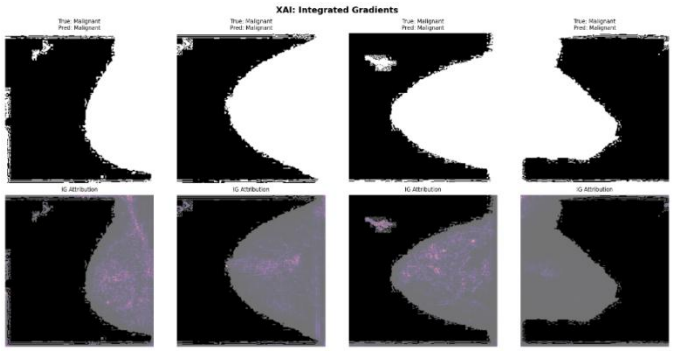

Fig. 10. Integrated Gradients-based attribution analysis showing pixel-level contribution patterns supporting malignant breast cancer prediction.

The Integrated Gradients-based attribution analysis of correctly classified malignant samples of mammograms is shown in Fig. 10; each sample is labelled as true malignant and predicted malignant. Whereas Grad-CAM detects pixel-level contribution patterns for class activation across the region, Integrated Grads highlights the fine-grained level of contribution patterns across the pixels that help the malignant prediction. The first line of images is the original mammography samples, and the second line of images is attribution maps placed over the mammography images. This purple and pink coloring in the images shows areas that are positively impacting the model's malignant decision. First sample: The attribution responses are concentrated on the dense areas of the breast tissue and boundary structures, indicating that the model takes into account local texture and intensity variation in the prediction process. Second sample: The image shows a computed tomography scan of a breast, with the attribution responses located around the dense regions of the breast tissue and boundary structures. In the third sample, the strength of the attribution is concentrated around the region of the breast that is internal and the region of the breast that is suspicious and localized, which is consistent with the focus for the model on diagnostically relevant tissue patterns. The fourth image is a comparatively weak but specific attribution, suggesting that the model isn't merely using background noise to make a guess. The attribution maps are not as clear or visually easily interpretable as heatmaps from Grad-CAM, but they offer more evidence of the importance of specific pixels and local texture patterns to the final prediction. While the classification performance was good for the whole dataset, the proposed framework did misclassify a small number of cases, mainly in situations where the benign and malignant lesions were visually similar, and/or when the image quality and tissue heterogeneity made the features of the lesions less discriminatory. This could happen because of very subtle boundaries in the lesions, overlapping tissue structure, dense breast tissue, or staining differences in the histopathological images. While few, these cases further demonstrate the inherent complexity of breast cancer diagnosis and suggest the need for further strengthening of multimodal feature learning. These failure scenarios will be explored in further detail in future work, and advanced feature representation and uncertainty estimation techniques will be investigated to further improve the reliability of the diagnostics.

Overall, IGs complement the Grad-CAM findings with pixel-level interpretability. The combined effect of region-level activation by Grad-CAM and pixel-level attribution by Integrated Gradients enhances the transparency of the proposed model for classification of malignant breast cancer, its reliability, and the diagnostic value of the model's trustworthiness.

V. Conclusion and Future Work

In this study, the author introduced a multimodal graph-attention-based framework, MAMGAT-Net, for early breast cancer diagnosis using multimodal imaging, namely mammographic and histopathological images. A proposed architecture of multi-scale feature extraction, residual learning, CBAM-based attention enhancement, bidirectional cross-modal fusion, and cross-gated multi-head graph attention learning is used to capture both macro-level radiological patterns and micro-level cellular abnormalities. Experimental results showed an improvement of the proposed framework compared to state-of-the-art conventional machine learning and deep learning approaches, with an accuracy of 94.59%, precision of 92.90%, recall of 96.56%, F1 score of 94.70%, and AUC of 98.58%. The ablation study showed the contribution of each architectural component, the ROC analysis showed good class separability, and the precision–recall evaluation and the confusion matrix results showed good diagnostic reliability. Moreover, Grad-CAM and integrated gradients visualization gave interpretable evidence to support the model's clinical decision-making. MAMGAT-Net successfully models the complementary multimodal information and relationships between diagnostic regions to improve classification robustness and avoid malignant case misclassification.

The multimodal framework was developed from two independently collected public datasets. Multimodal pairing was therefore carried out at a diagnostic class level, and not at the patient level, reducing the capacity to model true patient-specific cross-modal relationships. Second, the focus of the current study was mainly on the diagnostic performance and interpretability of the model, so computational efficiency was not systematically assessed, such as parameter count, FLOPs, inference latency, training time, and GPU memory usage. Third, the robustness of the model evaluation was not evaluated in the sense of domain shifts; imaging variations, acquisition noise, and external multi-center clinical data are not considered. Future studies will aim to test the proposed framework with clinically paired multimodal data to allow patient-level cross-modal learning, include additional clinical and genomic data to improve diagnostic capability, expand to multi-class breast cancer subtyping, and also implement extensive assessments of computational speedup and performance stability in different clinical settings.

REFERENCES

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 74, no. 3, pp. 229–263, 2024. https://doi.org/10.3322/caac.21834

[2] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," CA: A Cancer Journal for Clinicians, vol. 73, no. 1, pp. 17–48, 2023. https://doi.org/10.3322/caac.21763

[3] H. A. Rahman, S. N. N. Zaim, U. S. Suhaimei, and A. A. Jamain, "Prognostic factors associated with breast cancer-specific survival from 1995 to 2022: A systematic review and meta-analysis of 1,386,663 cases from 30 countries," Diseases, vol. 12, no. 6, p. 111, 2024. https://doi.org/10.3390/diseases12060111

[4] N. Becker and R. Liebscher, "Breast cancer diagnosis in the era of artificial intelligence: A systematic review," European Journal of Cancer, vol. 173, pp. 1–15, 2022. https://doi.org/10.1016/j.ejca.2022.05.030

[5] G. Ayana, K. Park, J.-W. Jeong, and S.-W. Choe, "Vision-transformer-based transfer learning for mammogram classification," Diagnostics, vol. 13, no. 2, p. 178, 2023. https://doi.org/10.3390/diagnostics13020178

[6] Nguyen, H. T., Nguyen, H. Q., Pham, H. H., Lam, K., Le, L. T., Dao, M., & Vu, V. (2023). VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, *10*(1), 277.

[7] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF ICCV, 2021, pp. 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

[8] Yang, Y., Liao, T., Lin, X. H., Ouyang, R., Chen, Q., & Ma, J. (2025). Dual-region MRI radiomic analysis indicates increased risk in high-risk breast lesions: bridging intratumoral and peritumoral radiomics for precision decision-making. *BMC cancer*, *25*(1), 828.

[9] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, pp. 1455–1462, 2016. https://doi.org/10.1109/TBME.2015.2476282

[10] S. Tummala, J. Kim, and S. Kadry, "BreaST-Net: Multi-class classification of breast cancer from histopathological images using ensemble of Swin transformers," Mathematics, vol. 10, no. 21, p. 4109, 2022. https://doi.org/10.3390/math10214109

[11] M. Macenko et al., "A method for normalizing histology slides for quantitative analysis," in Proc. IEEE ISBI, 2009, pp. 1107–1110. https://doi.org/10.1109/ISBI.2009.5193250

[12] Y. Li, M. E. H. Daho, P.-H. Conze, R. Zeghlache, H. Le Boité, R. Tadayoni, B. Cochener, M. Lamard, and G. Quellec, "A review of deep learning-based information fusion techniques for multimodal medical image classification," Computers in Biology and Medicine, vol. 177, p. 108635, 2024. https://doi.org/10.1016/j.compbiomed.2024.108635

[13] C. Chen, R. J. Chen, M. Y. Lu, D. F. K. Williamson, T. Y. Chen, and F. Mahmood, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in Proc. IEEE/CVF CVPR, 2022, pp. 16144–16155. https://doi.org/10.1109/CVPR52688.2022.01567

[14] L. Zhang, Y. Zhao, T. Che, S. Li, and X. Wang, "Graph neural networks for image-guided disease diagnosis: A review," iRADIOLOGY, vol. 1, no. 2, pp. 151–166, 2023. https://doi.org/10.1002/ird3.20

[15] Z. Wang, J. Ma, Q. Gao, C. Bain, S. Imoto, P. Liò, H. Cai, H. Chen, and J. Song, "Dual-stream multi-dependency graph neural network enables precise cancer survival analysis," Medical Image Analysis, vol. 97, p. 103252, 2024. https://doi.org/10.1016/j.media.2024.103252

[16] X. Meng and T. Zou, "Clinical applications of graph neural networks in computational histopathology: A review," Computers in Biology and Medicine, vol. 164, p. 107201, 2023. https://doi.org/10.1016/j.compbiomed.2023.107201

[17] A. G. Vrahatis, K. Lazaros, and S. Kotsiantis, "Graph attention networks: A comprehensive review of methods and applications," Future Internet, vol. 16, no. 9, p. 318, 2024. https://doi.org/10.3390/fi16090318

[18] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in Proc. ICLR, 2018. [Online]. Available: https://arxiv.org/abs/1710.10903

[19] R. J. Chen, M. Y. Lu, J. Wang, D. F. K. Williamson, S. J. Rodig, N. L. Lindeman, and F. Mahmood, "Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," IEEE Transactions on Medical Imaging, vol. 41, no. 4, pp. 757–770, 2022. https://doi.org/10.1109/TMI.2020.3021387

[20] S. Khoshraftar and A. An, "A survey on graph representation learning methods," ACM Transactions on Intelligent Systems and Technology, vol. 15, no. 1, pp. 1–55, 2024. https://doi.org/10.1145/3633518

[21] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules for multi-scale medical image analysis," IEEE Transactions on Medical Imaging, vol. 42, no. 3, pp. 812–825, 2023.

[22] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification with multi-scale feature fusion," Medical Image Analysis, vol. 85, p. 102764, 2023.

[23] R. Saha, S. Bhattacharya, and A. Mukherjee, "Graph convolutional networks for histopathological image analysis in breast cancer grading," Computers in Biology and Medicine, vol. 155, p. 106622, 2023.

[24] X. Wang, L. Zhang, and H. Liu, "Cross-modal attention fusion for multimodal breast cancer diagnosis using mammography and ultrasound," Expert Systems with Applications, vol. 211, p. 118642, 2023.

[25] P. Debelee, A. Kebede, and F. Schwenker, "Survey of deep learning in breast cancer image analysis," Evolving Systems, vol. 14, no. 1, pp. 19–40, 2023.

[26] T. Li, Y. Wu, and J. Wang, "Multi-scale feature extraction with convolutional block attention module for malignancy classification in digital mammography," Biomedical Signal Processing and Control, vol. 81, p. 104508, 2023.

[27] N. Moradi, M. Talebi, and A. Sharifi, "GAT-based relational learning for whole slide image analysis in invasive breast carcinoma," Artificial Intelligence in Medicine, vol. 136, p. 102475, 2023.

[28] H. Chen, Y. Li, X. Zhang, and Q. Tian, "Dual-branch transformer network for joint mammography and histopathology feature alignment in breast cancer detection," Pattern Recognition, vol. 136, p. 109225, 2024.

[29] M. Agarwal, S. Gupta, and K. Biswas, "A deep multimodal fusion framework integrating radiological and pathological imaging for early-stage breast cancer screening," Computers in Biology and Medicine, vol. 168, p. 107821, 2024.

[30] F. Khalid, D. Mirza, and U. Qayyum, "Stain normalization techniques for computational histopathology: A comparative review and clinical benchmarking," Journal of Pathology Informatics, vol. 15, p. 100373, 2024.

[31] S. Hussain, M. Anwar, and R. Ahmad, "ResNet-based deep feature extraction with graph attention pooling for breast tissue classification in whole slide images," Neural Computing and Applications, vol. 36, no. 4, pp. 1895–1912, 2024.

[32] Y. Liu, Z. Chen, and W. Zheng, "Heterogeneous graph neural network for integrating multimodal cancer imaging data," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 2, pp. 934–945, 2024.

[33] B. Srinivasan, K. Ramachandran, and V. Natarajan, "CBAM-enhanced EfficientNet for microcalcification detection in full-field digital mammograms," Diagnostics, vol. 14, no. 3, p. 312, 2024.

[34] A. Kaur, P. Singh, and R. Sharma, "Multimodal deep learning for breast cancer classification: A systematic review of fusion strategies and benchmark datasets," Neurocomputing, vol. 576, p. 127312, 2024.

[35] W. Dong, H. Sun, and Q. Zhao, "Graph attention networks with cross-gated aggregation for pathological node classification in tumor microenvironments," Medical Image Analysis, vol. 92, p. 103050, 2024.

[36] Z. Ibrahim, O. Khalifa, and A. Abdelaziz, "Patch-level feature aggregation using multi-head graph attention for breast histopathology malignancy grading," Bioengineering, vol. 12, no. 1, p. 54, 2025.

[37] T. Nguyen, P. Tran, and L. Hoang, "Vision transformer and graph convolutional integration for multimodal oncology image classification," Computers in Biology and Medicine, vol. 183, p. 109272, 2025.

[38] J. Park, S. Kim, and H. Lee, "Unified multimodal framework using cross-attention and relational graph learning for early breast cancer screening," IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 3, pp. 1423–1438, 2025.

[39] C. Xu, D. Yang, and L. Jiang, "t-SNE guided dimensionality refinement for improved multimodal cancer feature separability in deep classification networks," Expert Systems with Applications, vol. 261, p. 125401, 2025.

[40] R. Patel, M. Shah, and A. Desai, "Advances in multimodal AI for breast cancer diagnosis: Benchmarking cross-modal fusion on CBIS-DDSM and BreaKHis datasets," npj Breast Cancer, vol. 12, no. 1, p. 18, 2026.

[41] R. Sawyer-Lee, F. Gimenez, A. Hoogi, and D. Rubin, Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [Dataset]. The Cancer Imaging Archive, 2016. doi: 10.7937/K9/TCIA.2016.7O02S9CY