2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Intelligent Diagnosis for Breast Cancer Based on Multi-Modal Hierarchical Fusion of Ultrasound Images and Clinical Semantic Features

Jian Liu[1,3], Jie Ren[1], Guohui Wang[1], Yuqi Yan[2,3], Qunyang Zuo[1], Xinzheng Xue[1], and Dong Xu[2,3,*]

[1] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, 221116, Jiangsu, China.
[2] Department of Diagnostic Ultrasound Imaging & Interventional Therapy, Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, Hangzhou, 310022, Zhejiang, China.
[3] Wenling Institute of Big Data and Artificial Intelligence Institute in Medicine, Taizhou, Zhejiang, China.
[*]Corresponding author: xudong@zjcc.org.cn

Abstract. Ultrasound imaging plays a key role in breast cancer screening and diagnosis due to its advantages of non-invasive, real-time and low cost. This paper proposes an intelligent breast cancer diagnosis method based on Multi-modal Hierarchical Fusion Network (MHFNet), aiming to fully integrate complementary information from B-mode image, Doppler image and clinical semantic features. In MHFNet, a Semantic-augmented ResNet (SAR) was constructed to achieve the deep fusion of image and semantic features. And Hierarchical Semantic Fusion (HSF) module and Semantic Integration Bottleneck (SIB) are designed to enhance the interaction and fusion of Multi-modal information layer by layer. Finally, a unified Multi-modal feature representation was developed for breast cancer diagnosis. The experimental results show that the proposed Multi-modal classification fusion method is superior to other comparison algorithms, which fully verifies the positive role of Multi-modal information complementarity in improving the diagnostic performance of breast cancer.

Keywords: Breast cancer, Ultrasound image, Deep learning, Clinical semantic features, Multi-modal.

1 Introduction

Breast cancer is a highly occurring and lethal cancer disease and is one of the leading causes of cancer death in women [1]. Improving the early detection and diagnosis rate of breast cancer and optimizing the treatment plan have become the key tasks in the medical field. In recent years, the intelligent diagnosis technology based on deep learning and medical image processing has provided a new opportunity to improve the early screening and diagnosis accuracy. Ultrasound is widely used in the preliminary screening, diagnosis and treatment of breast cancer due to its non-invasive, real-time,

economic and non-radiation [2] [3]. Therefore, ultrasound imaging has an indispensable position in the prevention and treatment of breast cancer.

In the early studies of breast ultrasound images, scholars often use manual feature extraction, combined with classifier such as Adaboost [4], random forest [5] or K-nearest neighbor [6] for benign and malignant judgments. However, such methods rely heavily on the quality of manual features. Due to the complex noise and blurred boundary of breast ultrasound images, manual features are difficult to fully capture potential key information, leading to limited diagnostic performance. In addition, with the rapid increase in the amount of image data, the process of relying on manual feature selection is becoming more complicated.

Deep learning methods can automatically learn multi-level features from large-scale data [7]. Classical networks such as AlexNet [8], VGG, ResNet, etc. have achieved excellent results in natural image classification and demonstrated strong feature learning ability when migrating to the task of breast ultrasound diagnosis [9][10][11]. In addition, structures such as Attention mechanism [12] and U-Net [13] have also made positive progress in image segmentation and lesion localization.

It is difficult for a single modal ultrasound image to provide sufficient information, especially in the face of complex lesions, the diagnostic accuracy may be insufficient. Multi-modal ultrasound, such as traditional B-mode and Doppler imaging can show the morphological structure and biological characteristics of tumors from different angles [14]. B-mode and Doppler imaging are common in equipment and procedure and are less expensive. The combination of B-mode images, which show structural features of breast tissue and tumors, and Doppler images, which provide information on blood flow dynamics, can improve breast cancer diagnostic accuracy.

In Breast Image Reporting and Data System (BI-RADS) classification, doctors summarize rationality criteria based on long-term diagnosis and treatment experience, and grade tumor morphology, boundaries, and internal echoes. Combining such semantic features with imaging features can provide a more comprehensive input to machine learning or deep learning models [15] [16]. For example, Hamyoon et al. [17] screened 5 BI-RADS descriptors and then trained a tumor recognition model using support vector machine algorithm, whose diagnostic effect even exceeded the level of some clinicians. Since BI-RADS reflects the doctor's experience judgment of the overall state of the lesion, the introduction of this high-level semantic information can partially make up for the lack of data and the limitations of model learning, so that the network can not only "rely on pixels", but also "understand" key medical attributes.

Therefore, a Multi-modal Hierarchical Fusion Network (MHFNet) is proposed in this paper to combined Multi-modal ultrasound images with clinical semantic features. MHFNet can not only automatically learn details of images, but also incorporate the "semantic experience" accumulated by clinicians over the years into the model to achieve information from "pixels" to "clinical logic" [18]. The main contributions of this paper are as follows:

1. We designed a Semantic-Augmented ResNet, incorporating hierarchical semantic fusion layers and semantic integration bottlenecks to effectively embed clinical semantic information into ultrasound imaging data.

2. Multi-mode feature fusion of B-mode, Doppler ultrasound image and BI-RADS features is carried out to realize the information complementarity between different modes and improve the discriminating ability of benign and malignant.
3. Our proposed MHSFNet has achieved significant improvement in multiple evaluation indicators such as AUC and accuracy, which verifies the important role of Multi-modal information complementarity in the diagnosis of breast cancer.

2 Materials and Methods

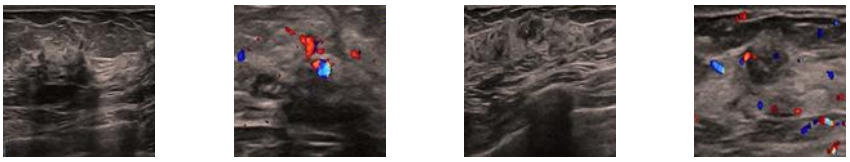
2.1 Data Preprocessing

The pairs of B-mode, Doppler, and BI-RADS data are used as inputs for breast cancer classification. In order to ensure that the data of the two modes of Doppler ultrasound image and B-mode ultrasound image can be effectively input into the deep learning model and enhance the generalization ability of the model, all images were first adjusted to a fixed size of 224×224 pixels. Then, we introduced a variety of data enhancement techniques in the pre-processing stage, mainly including random horizontal flip, random rotation and color jitter.

The BI-RADS features provide many useful pathological information, containing 10 term descriptions [19]. As shown in Table 1, for example, the location descriptors contain five scoring levels, including “outer top”, “outer bottom”, “inner bottom”, “inner top” and “central region”. Figure 1(a)(b) is a representative sample of the benign patient, and Figure 1(c)(d) is a representative sample of the malignant patient. Table 2 lists the scores of 10 BI-RADS descriptors corresponding to the two patients.

Table 1. Classification of BI-RADS feature manifestations.

ID	BI-RADS features	0	1	2	3	4	5
1	Location	-	Outside on	out-under	Inside under	Inside above	Central area
2	Shapes	-	Oval	Round	Irregular	-	-
3	Azimuth	-	Parallel	Non-parallel	-	-	-
4	Borders	-	Clarity	Lack of local clarity	Unsharpness	-	-
5	Internal echoes	-	Evenness	Uneven	-	-	-
6	Echo intensity	-	echoless	Hyperecho	Low echo	Isoecho	-
7	Composition	-	Solid	saccular	Cysticular	-	-
8	Rear echo	-	Unchanged	Enhanced	Attenuation	-	-
9	Calcification	no	Coarse calcification	Microcalcification	-	-	-
10	Calcification site	-	Within the lump	Outside the lump	intraductal	-	-



(a) B-mode of Benign patient (b) Doppler of Benign patient (c) B-mode of Malignant patient (d) Doppler of Malignant patient

Figure 1. The examples of B-mode and Doppler ultrasound images of benign and malignant patients.

Table 2. The scores of 10 BI-RADS descriptors corresponding to the two patients.

BI-RADS Features	1	2	3	4	5	6	7	8	9	10
Benign Patient	1	3	2	3	1	3	1	1	0	0
Malignant Patient	2	3	1	2	2	3	1	1	2	1

2.2 Overview of Multi-modal Hierarchical Fusion Network

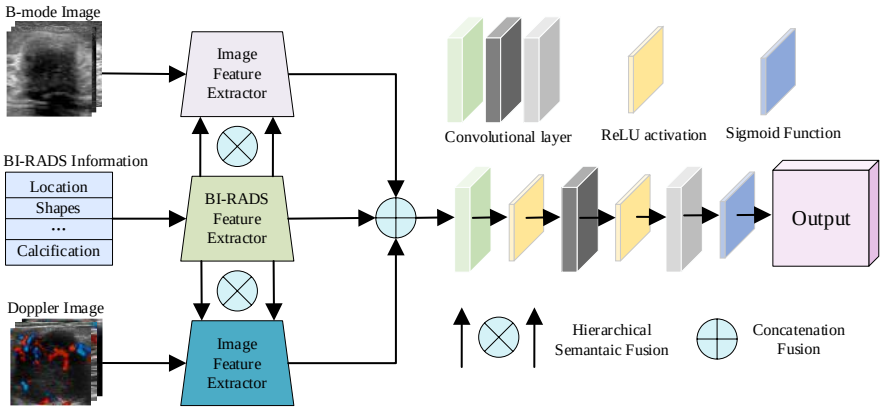


Figure 2. Overview of Multi-modal Hierarchical Fusion Network (MHFNet).

The overview of Multi-modal Hierarchical Fusion Network (MHFNet) is shown in Figure 2. Through the BI-RADS feature extractor, the BI-RADS features are sent to the Image Feature Extractor, and are fused in the extraction process of the Image Feature Extractor. Finally, the features of the three modes are concatenated and classified.

BI-RADS feature extractor. We employ 1D-CNN to extract BI-RADS features [20]. In the Input layer, let the original feature be $\mathcal{T} \in \mathbb{R}^{N \times D_T}$, D_T represent the feature dimension, and N is the batch size. Firstly, the input features $\mathcal{T}$ are adjusted to $\mathcal{T}' \in \mathbb{R}^{N \times 1 \times D_T}$. Then, the local patterns in $\mathcal{T}'$ are extracted through two layers of convolution and pooling successively, and finally the features are flattened and mapped through the fully connected layer to finally get the semantic representation: $\mathcal{T}_{\text{feat}} \in \mathbb{R}^{N \times 2048}$.

Image feature extractor. The specific structure of image feature extractor is described in detail in Section 2.3. In the specific implementation method, we use the Semantic - Augmented ResNet to independently process two image modes: B-mode and Doppler. Image feature extractor for each mode receives semantic representation $\mathcal{T}_{\text{feat}}$ in forward propagation and fuses with image features at all levels of the network.

2.3 Image-Clinical Feature Hierarchical Fusion Network

The idea of our network is to embed the clinical features into each residual module of the image extraction network. The whole design is mainly composed of three key components: Semantic-Augmented ResNet (SAR), Hierarchical Semantic Fusion (HSF) module and Semantic Integration Bottleneck (SIB).

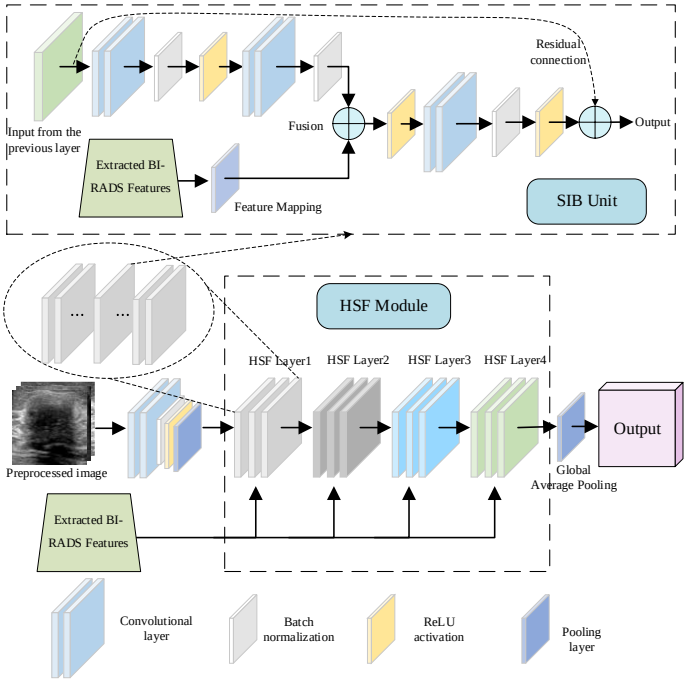


Figure 3. The structure of Semantic-Augmented ResNet (SAR).

Semantic-Augmented ResNet (SAR). Based on the modular design, we built SAR, which combines semantic representation and image features in each residual layer to achieve deep integration of Multi-modal information. The structure of SAR can be found in Fig. 3. SAR includes four designed HSF layers, which is encapsulated by SIB units. Let the input image be $\mathcal{V} \in \mathbb{R}^{N\times 3\times H_0\times W_0}$ and the clinical feature be $\mathcal{T}_{\text{feat}}$, and the forward propagation process is as follows:

After convolution (Cov), batch normalization (BN) and ReLU activation, we get

$$\boldsymbol{F}_0' = \text{ReLU}(\text{BN}(\text{Conv}(\mathcal{V}))) \tag{1}$$

Downsampling by maximum pooling layer is applied to $\boldsymbol{F}_0'$

$$\boldsymbol{F}_0 = \text{MaxPool}(\boldsymbol{F}_0') \tag{2}$$

Sequentially through four HSF layers, $\boldsymbol{F}_0$ is fused with clinical features $\mathcal{T}_{\text{feat}}$. The output of each HSF layer can be denoted as

$$\boldsymbol{F}_l = \text{HSF}_l(\boldsymbol{F}_{l-1}, \mathcal{T}_{\text{feat}}), l = 1,2,3,4 \tag{3}$$

Finally, the output $\hat{\boldsymbol{y}}$ of SAR is obtained through global average pooling (GAP)

$$\hat{\boldsymbol{y}} = \boldsymbol{\sigma}(\boldsymbol{W}^*\text{GAP}(\boldsymbol{F_4}) + \boldsymbol{b}) \tag{4}$$

where $\boldsymbol{\sigma}()$ is the sigmoid activation function, $\boldsymbol{W}$ is the learnable weight vector, $\boldsymbol{b}$ is the bias vector.

Hierarchical Semantic Fusion Module (HSF). The HSF module ensures that the semantic representation $\mathcal{T}_{\text{feat}}$ is permeated and fully utilized layer by layer throughout the residual layer by modularistically cascading multiple SIB units. Let the initial input of SIB be $\boldsymbol{x}_0$, for the *i*-th SIB unit, we get its output $\boldsymbol{x}_i$:

$$\boldsymbol{x}_i = \text{SIB}_i(\boldsymbol{x}_{i-1}, \mathcal{T}_{\text{feat}}), i = 1,2,\dots,n \tag{5}$$

In the same HSF layer, the final output of the SIB unit can be defined as $\boldsymbol{x}_n$, which is also the output of the corresponding HSF layer.

Semantic Integration Bottleneck (SIB). The core of SIB unit is to map the semantic features to the same dimension as the convolutional layer output channel number C through the fully connected layer, and then fuse with the convolutional output. Let the original semantic feature be $\mathcal{T}_{\text{feat}}$. Define a full join feature map $f_{\text{fc}}(\mathcal{T}_{\text{feat}}) = \boldsymbol{W}\mathcal{T}_{\text{feat}} + \boldsymbol{b}$. Then, the mapping result is adjusted to a four-dimensional tensor $\boldsymbol{T}_{\text{proj}} \in \mathbb{R}^{N\times C\times 1\times 1}$ with the unsqueeze operation:

$$\boldsymbol{T}_{\text{proj}} = \text{unsqueeze}(f_{\text{fc}}(\mathcal{T}_{\text{feat}})) \tag{6}$$

Let the output of the convolutional layer be $\mathscr{F} \in \mathbb{R}^{N\times C\times H\times W}$ (The upper branch of SIB Unit in Fig.3), and the mapped semantic features are added to the convolutional features by using the addition fusion strategy:

$$\mathscr{F}' = \text{ReLU}\left(\mathscr{F} + \boldsymbol{T}_{\text{proj}}\right) \tag{7}$$

Finally, the original residual connection mechanism is maintained. Let the input from the previous layer be $\boldsymbol{x}$, and the output which incorporates the image feature of the text feature is:

$$\boldsymbol{O} = \text{ReLU}\left(\boldsymbol{x} + \mathscr{F}'\right) \tag{8}$$

2.4 Multi-modal Fusion

Assume that the two image semantic representations obtained separately through the image feature extractor are $\mathcal{V}_1 \in \mathbb{R}^{N\times1024}$ and $\mathcal{V}_2 \in \mathbb{R}^{N\times1024}$, and the semantic representations obtained separately through the BI-RADS feature extractor are $\mathcal{T}_{\text{feat}} \in \mathbb{R}^{N\times2048}$. We use the concatenation operation to integrate these three parts of features into a unified representation $\mathcal{F}_{\text{fusion}}$ in dimension:

$$\mathcal{F}_{\text{fusion}} = \text{Concat}(\mathcal{V}_1, \mathcal{V}_2, \mathcal{T}_{\text{feat}}) \in \mathbb{R}^{N\times4096} \tag{9}$$

The fused Multi-modal features $\mathcal{F}_{\text{fusion}}$ are fed into a subsequent fully connected network for classification.

3 Materials and Methods

3.1 Experimental Dataset

This study obtained informed consent from all patients for their information to be used in the study without violating patient privacy. The data set used in this chapter was provided by Zhejiang Cancer Hospital, including 437 benign images and 251 malignant images, as well as clinical semantic features corresponding to all patients. The breast images in the dataset had corresponding diagnostic labels.

3.2 Experimental Setup

In this study, the network was trained and evaluated on a server system with 64GB memory, NVIDIA 3090 24GB GPU, Python3.12, and Pytorch2.3.0 cuda12.1. We applied 5 fold cross validation and each fold trained 40 epochs. Binary cross entropy loss is taken as a loss function and Adam optimizer with an initial learning rate set to 0.001 is used in the experiment. To improve the training process, we also dynamically used the learning rate scheduler, deciding whether to reduce the learning rate by 10% every 5 epochs.

3.3 Experimental Results and Analysis

Comparison of experimental results under different modality conditions. In this section, we compare the results of our MHFNet on single-modality, dual-modality, and three-modality conditions. The experimental results are shown in Table 3. Overall, the experimental results indicate that the results of three-modality condition outperform single-modality and dual-modality across all metrics. Compared to dual-modality, three-modality not only enhances predictive performance but also exhibits a lower standard deviation, reflecting greater stability and consistency, thereby fully demonstrating the superiority of Multi-modal learning.

Table 3. Comparative experimental results under different modal conditions.

Modal	AUC	Accuracy(%)	F1-score(%)	Recall(%)	Precision(%)
Clinical	0.9028±0.0237	84.01±1.34	87.74±1.29	90.35±3.50	85.38±1.26
B-mode	0.8289±0.0297	75.45±5.47	81.10±4.29	83.12±5.98	79.36±4.00
Doppler	0.8650±0.0209	76.90±2.94	82.19±2.78	84.39±4.99	80.28±2.83
Clinical+ B-mode	0.9053±0.0237	84.45±1.08	87.98±1.08	89.87±3.49	86.34±2.56
Clinical+ Doppler	0.9082±0.0214	84.74±2.23	88.17±1.89	89.89±4.25	86.69±2.47
B-mode+ Doppler	0.8597±0.0234	77.91±2.16	82.98±2.05	85.11±4.58	81.22±3.33
Clinical+ B-mode+ Doppler	**0.9163±0.0141**	**86.05±0.54**	**89.25±0.56**	**91.30±1.40**	**87.34±1.52**

Comparison results of different algorithms. We compared our MHFNet with other algorithms. In terms of traditional vision algorithms, we compare ResNet50 and VGG16. About Transformer, we compared Swin Transformer [21] and ViT-B/16[22]. On Multi-modal algorithms, we compare CLIP[23] and ViLT[24]. In addition, due to our use of hierarchical fusion, we also compared the results under the Late Fusion strategy [25]. According to the experimental results in Table 4, our MHFNet has achieved the highest or close to the highest values in the indexes of AUC, Accuracy, F1-score, Recall and Precision.

Table 4. Comparative experimental results under different modal conditions.

Methods	AUC	Accuracy(%)	F1-score(%)	Recall(%)	Precision(%)
Swin Transformer	0.9047±0.0254	84.88±0.53	88.23±0.82	89.41±3.64	87.25±2.23
Resnet50	0.9013±0.0234	83.28±0.65	87.01±0.78	88.28±3.23	85.91±1.92
Vgg16	0.6614±0.1979	58.41±10.67	62.25±31.14	80.00±40.00	50.96±25.52
Vit-B/16	0.9011±0.0258	84.45±2.03	87.89±2.01	89.40±4.17	86.56±1.68
CLIP	0.8151±0.0236	63.52±1.38	77.68±1.03	**100.00±0.00**	63.52±1.38
ViLT	0.8978±0.0145	82.56±2.16	86.65±1.75	89.24±4.08	84.43±3.32
Late Fusion	0.9035±0.0207	81.40±3.25	85.55±3.11	87.54±6.50	83.97±2.81
MHFNet	**0.9163±0.0141**	**86.05±0.54**	**89.25±0.56**	91.30±1.40	**87.34±1.52**

4 Conclusion

This paper proposed a novel deep learning method, named as Multi-modal Hierarchical Fusion Network (MHFNet), for breast cancer diagnosis. Specifically, we designed Semantic-Augmented ResNet, Hierarchical Semantic Fusion Layer and Semantic Integration Bottleneck modules to improve classification performance. Experimental results show that the proposed MHSFNet has achieved significant improvement in multiple

evaluation indicators such as AUC, accuracy, F1-score and Recall, which verifies the important role of Multi-modal information complementarity in the diagnosis of breast cancer. The method provides a new technical path for the intelligent diagnosis of the combination of clinical semantics and ultrasound images. Future research work can be carried out from the data scale and modal expansion as well as the further optimization of network structure.

Acknowledgements. This research was supported by the National Natural Science Foundation of China (Grant No. 61906198), the Natural Science Foundation of Jiangsu Province (Grant No. BK20190622), Xuzhou Special Fund for Promoting Science and Technology Innovation-Key R&D Program (Social Development) (Grant No. KC23237).

References

1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Je mal, A.: Global cancer statistics 2022: Globocan estimates of incidence and mor tality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **74**(3), 229–263 (2024)
2. Guo, R., Lu, G., Qin, B., Fei, B.: Ultrasound imaging technologies for breast cancer detection and management: a review. Ultrasound in medicine & biology **44**(1), 37 70 (2018)
3. Ozcan, B.B., Xi, Y., Dogan, B.E., Porembka, J.H.: Cost-effectiveness of an ultrasound-first strategy in the diagnostic evaluation of noncalcified lesions recalled from screening digital breast tomosynthesis. American Journal of Roentgenology **223**(4), e2431422 (2024)
4. Zhou, S., Shi, J., Zhu, J., Cai, Y., Wang, R.: Shearlet-based texture feature ex traction for classification of breast tumor in ultrasound image. Biomedical Signal Processing and Control **8**(6), 688–696 (2013)
5. Pavithra, S., Vanithamani, R., Justin, J.: Computer aided breast cancer detection using ultrasound images. Materials Today: Proceedings **33**, 4802–4807 (2020)
6. González-Luna, F.A., Hernández-López, J., Gomez-Flores, W.: A performance evaluation of machine learning techniques for breast ultrasound classification. In: 2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE). pp. 1–5. IEEE (2019)
7. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep con volutional neural networks. Advances in neural information processing systems **25** (2012)
9. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomed ical image segmentation. In: Medical image computing and computer-assisted intervention–

MICCAI 2015: 18th international conference, Munich, Germany, Oc tober 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
14. Hussain, S., Ali, M., Naseem, U., Avalos, D.B.A., Cardona-Huerta, S., Tamez Peña, J.G.: Multiview multimodal feature fusion for breast cancer classification using deep learning. IEEE Access (2024)
15. Huang, Q., Zhang, F., Li, X.: A new breast tumor ultrasonography cad system based on decision tree and bi-rads features. World Wide Web **21**, 1491–1504 (2018)
16. Baek, J., Lee, D.B., Hwang, S.J.: Learning to extrapolate knowledge: Transductive few-shot out-of-graph link prediction. Advances in neural information processing systems **33**, 546–560 (2020)
17. Zhang, Z., Cai, J., Zhang, Y., Wang, J.: Learning hierarchy-aware knowledge graph embeddings for link prediction. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 3065–3072 (2020)
18. Abhisheka, B., Biswas, S.K., Saha, D., Das, S., Purkayastha, B.: The efficacy of combining deep and handcrafted features for breast cancer classification using ul trasound images. In: 2023 10th IEEE Uttar Pradesh Section International Confer ence on Electrical, Electronics and Computer Engineering (UPCON). vol. 10, pp. 735–740. IEEE (2023)
19. Magny, S.J., Shikhman, R., Keppke, A.L.: Breast imaging reporting and data sys tem. In: StatPearls [Internet]. StatPearls publishing (2023)
20. Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., Inman, D.J.: 1d convolutional neural networks and applications: A survey. Mechanical systems and signal processing **151**, 107398 (2021)
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
24. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convo lution or region supervision. In: International conference on machine learning. pp. 5583–5594. PMLR (2021)
25. Gadzicki, K., Khamsehashari, R., Zetzsche, C.: Early vs late fusion in multimodal convolutional neural networks. In: 2020 IEEE 23rd international conference on information fusion (FUSION). pp. 1–6. IEEE (2020)