MSHF-Net: Multimodal Breast Cancer Molecular Subtype Prediction via Segmentation-Guided Hierarchical Fusion Network

Xinjie Jiang[1], Chengye Zhou[1], Yueyuan Li[2], Yuqian Wang[3], Yue Zhang[3], Qingyu Nian[4], Lingyan Zhang[5], Feiwei Qin[1,✉], and Changmiao Wang[6,✉]

[1] Hangzhou Dianzi University, Hangzhou, China
qinfeiwei@hdu.edu.cn
[2] Shanghai Normal University, Shanghai, China
[3] Guangzhou University of Chinese Medicine, Shenzhen, China
[4] Zhejiang University of Finance & Economics, Hangzhou, China
[5] Longgang Central Hospital of Shenzhen, Shenzhen, China
[6] Shenzhen Research Institute of Big Data, Shenzhen, China
cmwangalbert@gmail.com

Abstract. Breast cancer treatment and prognosis depend strongly on molecular subtype. However, conventional invasive biopsy and postoperative pathology can be affected by tumor heterogeneity and sampling bias, which motivates non-invasive, preoperative subtype prediction. While recent deep learning approaches have enabled imaging-based subtype prediction, relying on a single imaging modality often provides incomplete information. Current multimodal fusion methods also struggle to capture differences in information scale and the complementary relationships among Mammography (MG), Ultrasound (US), and clinical data. In addition, because breast lesions often occupy only a small region of the image, weakly supervised classification can be distracted by surrounding background tissue, reducing reliability. To address these limitations, we propose a Multimodal Segmentation-Guided Hierarchical Fusion Network (MSHF-Net) to classify breast cancer into Luminal and Non-luminal subtypes. The model uses Hierarchical Tri-modal Fusion (H-TMF) to better integrate complementary signals from MG, US, and clinical features. It also introduces a Segmentation-Guided Aggregation (SGA) module to strengthen discrimination by combining global context with localized lesion information. Finally, a Class-Balanced Focal (CBF) Loss improves recognition of minority subtypes and mitigates the Luminal/Non-luminal class imbalance. Experiments on the self-collected LGCH-Breast dataset show that MSHF-Net outperforms existing methods across multiple metrics, indicating strong potential for non-invasive multimodal subtype prediction. Code is available at `https://github.com/knot41/MSHF-Net`.

Keywords: Multimodal Learning · Hierarchical Fusion · Segmentation-Guided · Class-Balanced · Molecular Subtype Prediction.

1 Introduction

Breast cancer is one of the most common malignancies among women worldwide and continues to place a substantial burden on public health systems [22]. Mammography (MG) and Ultrasound (US) are the two imaging modalities most widely used for screening and diagnosis. As cancer care becomes increasingly personalized, treatment decisions rely more on tumor biology and less on anatomical staging alone. Using the expression status of Estrogen Receptor (ER), Progesterone Receptor (PR), Human Epidermal Growth Factor Receptor 2 (HER2), and Ki-67, breast cancer is commonly grouped into Luminal A, Luminal B, HER2-enriched, and Triple-Negative Breast Cancer (TNBC), each associated with distinct prognoses and therapeutic strategies [2,9]. In clinical practice, distinguishing luminal (ER+ and/or PR+) from non-luminal (HER2+ and TNBC) disease is especially important because it guides endocrine therapy, chemotherapy, and targeted treatment selection [13]. Accurate preoperative prediction of molecular subtype is therefore of substantial clinical value.

At present, molecular subtype is primarily determined using invasive core needle biopsy followed by Immunohistochemistry (IHC) [20]. However, biopsy reflects only a limited tumor region and may not capture spatial or temporal heterogeneity, which can lead to incomplete biological assessment and potential subtype misclassification [6, 21]. These limitations have motivated interest in non-invasive alternatives. Prior studies suggest that molecular subtypes exhibit distinct imaging patterns, supporting the feasibility of predicting subtype from routine medical images [17].

With recent progress in deep learning, models can learn complex imaging features automatically, and multiple studies have explored image-based subtype prediction [10, 27]. Mota *et al.* [16] developed a MG-only model, but single-modality approaches inherently limit the information available for prediction. Rabah *et al.* [1] incorporated clinical variables into a multimodal framework, yet the absence of US leaves complementary morphological cues underused. Zhang *et al.* [26] proposed the MDL-IIA model to fuse MG and US, but their single-scale fusion excludes clinical data and does not jointly capture global context and fine-grained lesion characteristics. In addition, the marked imbalance between luminal and non-luminal subtypes is often insufficiently addressed, which can degrade performance on less frequent classes.

To address these gaps, we propose the Multimodal Segmentation-Guided Hierarchical Fusion Network (MSHF-Net), a framework that jointly leverages MG, US, and clinical data for breast cancer molecular subtype prediction. Our main contributions are as follows: First, we introduce Hierarchical Tri-Modal Fusion (H-TMF) to integrate the three modalities at multiple levels and promote effective information exchange across them. Second, we propose Segmentation-Guided Aggregation (SGA) to combine whole-image information with lesion-focused information, strengthening subtype discrimination. Third, to mitigate subtype imbalance, we design a Class-Balanced Focal (CBF) Loss that improves recognition of under-represented classes and enhances prediction relia-

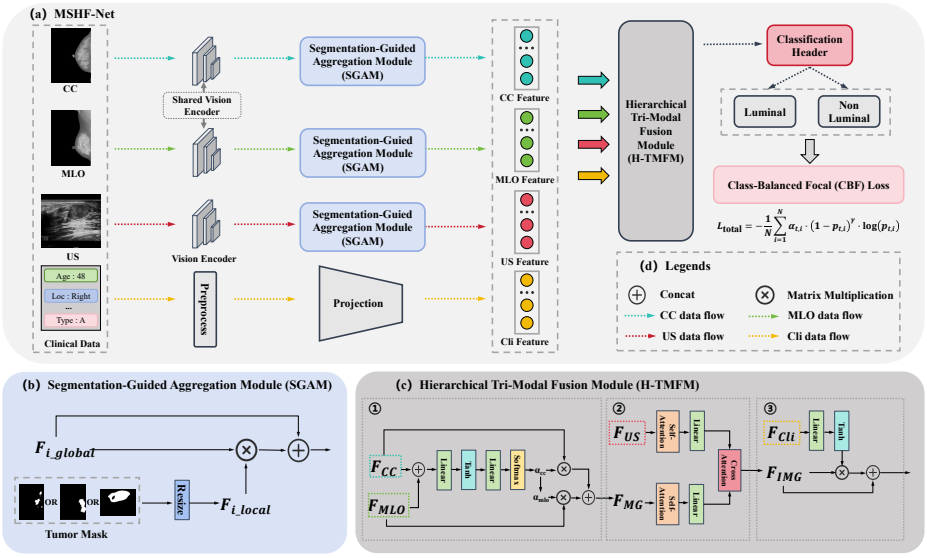

Fig. 1. The overall architecture of the MSHF-Net. (a) Depicting the network flow, including feature extraction, multimodal fusion, and loss function. (b) The SGA module intergrates global and local feature. (c) The H-TMF module facilitates multimodal fusion.

bility. Fourth, extensive experiments on our self-collected LGCH-Breast dataset demonstrate strong performance and computational efficiency of MSHF-Net.

2 Proposed Method

Fig. 1 illustrates the architecture of MSHF-Net for predicting luminal versus non-luminal subtypes by jointly using mammography (MG; Craniocaudal (CC) and Mediolateral Oblique (MLO) views), US, and clinical data. Within the model, the SGA module emphasizes tumor-relevant regions, the H-TMF module combines information across the different data sources, and the CBF Loss guides training while mitigating the effects of unequal class representation in the clinical dataset.

2.1 Segmentation-Guided Aggregation Module

To extract lesion-specific features more reliably, we embed the SGA module in each imaging branch (Fig. 1(b)). The module uses the tumor mask to indicate where the lesion is, encouraging the network to emphasize lesion regions while still retaining contextual information from the full image. The CC and MLO mammography views share the same visual encoder to learn consistent representations across views, whereas US uses a separate encoder. For each branch,

the encoder produces a global feature map, which is then reweighted by a spatial map obtained by resizing the tumor mask to the feature resolution. The aggregation is defined as:

$$F_{i\_\text{aggregated}} = F_{i\_\text{global}} + (F_{i\_\text{global}} \otimes F_{i\_\text{local}}), \tag{1}$$

where F_{i_global} is the feature map extracted by the vision encoder, F_{i_local} is the spatial weight map derived from the resized tumor mask, and $\otimes$ denotes element-wise multiplication. This design increases the responses within the lesion region, and the residual connection helps preserve global information that may still be diagnostically relevant.

2.2 Hierarchical Tri-Modal Fusion Module

After lesion-aware feature extraction, we fuse information across modalities using the H-TMF strategy (Fig. 1(c)). The fusion proceeds in three stages to reflect the structure and complementarity of the inputs: (i) fuse the two MG views, (ii) integrate MG with US, and (iii) adjust the fused imaging representation using clinical variables.

MG-View Fusion. We first combine the CC and MLO views because they depict the same anatomy from correlated perspectives. Given their features F_{CC} and F_{MLO}, we estimate a data-dependent weight for each view using a small two-layer perceptron:

$$e_i = \mathbf{W}_2 \cdot \tanh(\mathbf{W}_1 F_i + \mathbf{b}_1) + b_2, \quad i \in \{\text{CC}, \text{MLO}\}. \tag{2}$$

The scores are normalized with softmax to obtain view weights:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j\in\{\text{CC,MLO}\}} \exp(e_j)}. \tag{3}$$

The fused MG representation is then computed as:

$$F_{MG} = \alpha_{\text{CC}} \cdot F_{CC} + \alpha_{\text{MLO}} \cdot F_{MLO}. \tag{4}$$

MG-US Cross-Attention. We next integrate F_{MG} with the US feature F_{US}. MG is effective at depicting microcalcifications, whereas US is often more informative for hypoechoic masses; combining them can therefore provide complementary diagnostic cues, despite differences in their feature distributions.

To model relationships within each modality, we first apply self-attention independently to the MG and US feature sequences, and we reuse the same symbols (F_{MG} and F_{US}) to denote the updated representations for simplicity. We then apply bidirectional cross-attention so that each modality can retrieve complementary information from the other:

$$F_{MG}^{cross} = F_{MG} + \text{MHA}(\text{LN}(F_{MG}), \text{LN}(F_{US}), \text{LN}(F_{US})), \tag{5}$$

$$F_{US}^{cross} = F_{US} + \text{MHA}(\text{LN}(F_{US}), \text{LN}(F_{MG}), \text{LN}(F_{MG})), \tag{6}$$

where $\mathrm{MHA}(Q, K, V)$ denotes multi-head attention with query Q, key K, and value V. $\mathrm{LN}(\cdot)$ is layer normalization for stabilizing training. Finally, we concatenate the `[CLS]` tokens from both cross-attended modalities and compress them with a feed-forward network to obtain the unified image feature F_{IMG}:

$$F_{IMG} = \mathrm{Dropout}\left(\mathrm{GELU}\left(\mathbf{W} \cdot [F_{MG}; F_{US}] + \mathbf{b}\right)\right), \tag{7}$$

where $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters, $[\cdot;\cdot]$ denotes concatenation, $\mathrm{GELU}(\cdot)$ is the activation function, and $\mathrm{Dropout}(\cdot)$ is used to reduce overfitting.

Clinical Gating. Finally, we use five clinical variables, including age, maximal tumor diameter, menopausal state, tumor location, and breast fiber type, to adjust the fused imaging representation. Continuous variables are z-score normalized, and categorical variables are one-hot encoded. Following the idea in Metafusion [19], we transform the clinical feature F_{Cli} with a small network into a per-channel gating vector and apply it to F_{IMG} through a residual form:

$$F_{Total} = F_{IMG} + F_{IMG} \odot \tanh(\mathbf{W}_{meta} F_{Cli} + \mathbf{b}_{meta}), \tag{8}$$

where $\mathbf{W}_{meta}$ and $\mathbf{b}_{meta}$ are learnable parameters and $\odot$ denotes element-wise multiplication. This step calibrates the image-derived features using patient-specific clinical context.

2.3 Class-Balanced Focal Loss

The final representation F_{Total} is passed to a classification head to distinguish luminal from non-luminal subtypes. To mitigate class imbalance and focus learning on harder examples, we adopt CBF loss:

$$L_{\text{total}} = -\frac{1}{N} \sum_{i=1}^{N} \alpha_{t,i} \cdot (1 - p_{t,i})^{\gamma} \cdot \log(p_{t,i}), \tag{9}$$

where N is the number of samples, $p_{t,i}$ is the predicted probability of the ground-truth class, $\alpha_{t,i}$ is the class-balancing weight, and γ controls how strongly the loss emphasizes difficult samples.

3 Experiments

3.1 Data and Implementation

Data Preparation. MSHF-Net is designed to use three inputs, MG, US, and clinical information, so collecting fully paired data is difficult in practice. To address this constraint, we conducted extensive experiments on the large-scale LGCH-Breast dataset collected at our collaborating hospital.

The LGCH-Breast Dataset. LGCH-Breast includes paired MG, US, and clinical data from 401 cases. MG images were acquired using the GIOTTO IMAGE 3D system with standard CC and MLO views. US examinations were

performed on GE LOGIQ E9, Philips ATL HDI-5000, and Toshiba Nemio XG scanners using 7–12 MHz linear probes. We excluded cases with poor image quality, prior breast surgery or treatment, or incomplete pathology and IHC records. Lesion Regions of Interest (ROIs) were independently annotated by two experienced breast radiologists, with disagreements resolved by consensus, and the resulting masks were used in both training and evaluation.

Implementation Details. All experiments were implemented in PyTorch 2.10 and run on a Tesla V100 GPU. We used ResNet-50 as the feature extractor [5], initialized with RadImageNet pretraining [15]. During task adaptation, we applied different learning rates across model components: the final feature block used 1×10^{-6}, while the H-TMF module and classification heads used 1×10^{-4}. Training ran for 50 epochs with the AdamW optimizer (batch size 16). The learning rate was reduced when validation performance plateaued (factor=0.5, patience=15). To mitigate class imbalance, we used weighted sampling based on inverse class frequency during training. Performance was evaluated using Accuracy (Acc), F1-score (F1), Recall (Rec), Precision (Pre), and Area Under the Curve (AUC), with Acc and AUC as the primary metrics. Model evaluation was conducted on all 401 cases using three repeated five-fold cross-validation experiments with different random seeds. All quantitative results in Tables 1–5 are reported as the mean and the half-width of the 95% confidence interval. For each table, paired tests were performed on fold-level AUCs between the best and second-best methods, with the p-values provided in the code repository.

3.2 Experimental Results

Comparisons. Table 1 compares different feature extractor networks for molecular subtype prediction. ResNet-50 achieved the best overall performance, improving AUC and Acc by 1.7% and 1.9%, respectively, over DenseNet-121 [7]. ViT-B [3] performed worse than the convolutional alternatives, which may reflect the Vision Transformer's stronger dependence on large training datasets.

Table 2 compares MSHF-Net with task-specific methods (Luo *et al.* [14], Ben *et al.* [1], MDL-IIA [26]), classical multimodal baselines (MMBT [8], CLIP [18]), and cross-domain methods (HFBSurv [12], TMI-CLNet [24], UniCross [25]). MSHF-Net outperformed all compared methods across every metric, achieving 82.1% Acc and 85.3% AUC. Compared with MDL-IIA, the best-performing bimodal method, MSHF-Net improved F1 and Pre by 3.4% and 4.7%, respectively, highlighting the advantage of effectively fusing three modalities over using only two. When compared with classical multimodal methods such as MMBT and CLIP, which also take M+U+C trimodal input, MSHF-Net consistently achieved higher scores, indicating that simple cross-modal feature stacking is insufficient to fully capture the deep correlations among medical modalities. Furthermore, against strong cross-domain methods like TMI-CLNet, MSHF-Net improved Rec by 4.4%, demonstrating higher sensitivity in identifying potential positive cases, which is crucial for clinical diagnosis. Overall, these results show that MSHF-Net not only achieves superior overall performance and robustness but also generalizes well across heterogeneous multimodal data.

To assess the proposed H-TMF module, we compared it with early fusion [4], late fusion [4], and attention-based fusion [11]. As shown in Table 3, H-TMF achieved the best overall results in Acc, F1, Rec, and AUC. Compared with attention-based fusion, H-TMF improved AUC and Acc by 2.1% and 1.6%, respectively. Although Rec was slightly lower than the best early fusion result (−0.9%), H-TMF delivered the strongest overall performance, supporting its effectiveness for multimodal feature integration.

Table 1. Comparison of different backbone architectures. M denotes MG, U denotes US, and C denotes clinical data. The best and second-best results are highlighted in **bold** and underline, respectively.

Method	Modality	Acc (%)	F1 (%)	Rec (%)	Pre (%)	AUC (%)
ResNet-50 [5]	M+U+C	**82.1±1.7**	**76.1±2.0**	**71.8±2.6**	**80.9±1.7**	**85.3±3.2**
Inception-v3 [23]	M+U+C	77.4±2.2	65.6±2.6	63.0±2.5	68.5±2.6	79.5±2.8
DenseNet-121 [7]	M+U+C	80.2±1.9	70.5±2.2	68.0±2.4	73.2±2.0	83.6±2.5
ViT-B [3]	M+U+C	78.9±2.0	68.5±2.3	65.5±2.1	71.8±2.4	81.4±2.1

Table 2. Comparison of our MSHF-Net with other multimodal approaches. M denotes MG, U denotes US, and C denotes clinical data. The best and second-best results are highlighted in **bold** and underline, respectively.

Method	Modality	Acc (%)	F1 (%)	Rec (%)	Pre (%)	AUC (%)
Part A: Task-Specific Methods						
Luo *et al.* [14]	M	74.5±2.4	58.6±2.6	55.4±3.1	62.1±2.8	70.2±2.5
Ben *et al.* [1]	M+C	73.2±2.1	61.2±2.5	58.2±2.6	64.5±2.2	73.5±2.3
MDL-IIA [26]	M+U	80.5±1.3	72.7±1.6	69.5±1.8	76.2±1.5	82.8±1.4
Part B: Classical Methods						
MMBT [8]	M+U+C	76.4±1.8	66.5±2.1	63.8±2.2	69.5±1.9	77.8±1.8
CLIP [18]	M+U+C	75.2±2.2	63.8±2.4	61.0±2.5	66.8±2.3	75.5±2.1
Part C: Cross-Domain Methods						
HFBSurv [12]	M+U+C	77.5±1.7	68.2±1.9	65.2±2.0	71.4±1.8	79.2±1.6
TMI-CLNet [24]	M+U+C	78.8±1.5	70.4±1.8	67.4±1.9	73.6±1.7	80.5±1.5
UniCross [25]	M+U+C	75.8±1.9	65.3±2.2	62.5±2.4	68.3±2.0	76.4±1.9
MSHF-Net (Ours)	M+U+C	**82.1±1.7**	**76.1±2.0**	**71.8±2.6**	**80.9±1.7**	**85.3±3.2**

Ablation Study. We conducted ablation experiments on the LGCH-Breast dataset to quantify the contribution of each proposed component. The quantitative results are summarized in Tables 4 and 5. As shown in Table 4, models trained with a single modality exhibit limited performance, whereas the full trimodal setting (MG+US+clinical data) achieves the best results, with an Acc

Table 3. Comparison of our H-TMF module with different fusion strategies. The best and second-best results are highlighted in **bold** and underline, respectively.

Method	Acc (%)	F1 (%)	Rec (%)	Pre (%)	AUC (%)
Early Fusion [4]	78.8±1.5	70.1±1.4	**72.7±2.8**	69.4±2.2	75.6±1.8
Late Fusion [4]	76.1±1.0	71.1±1.2	70.7±1.4	73.4±3.8	76.8±1.4
Attention Fusion [11]	80.0±2.7	73.4±2.1	72.0±1.8	77.8±3.0	83.7±3.1
H-TMF Module (Ours)	**82.1±1.7**	**76.1±2.0**	71.8±2.6	**80.9±1.7**	**85.3±3.2**

of 82.1% and an AUC of 85.3%. This advantage was further supported by the paired AUC test, where the tri-modal setting significantly outperformed MG+US ($p = 0.0253$).

Table 5 reports the component-level results. Adding the SGA module increases Acc, F1, Rec, Pre, and AUC by 1.7%, 5.7%, 8.7%, 2.2%, and 3.1%, respectively. Incorporating the CBF loss further improves Acc, F1, and Pre by 5.8%, 5.8%, and 9.3%, indicating that it effectively mitigates class imbalance. The complete model delivers the strongest overall performance across all metrics (Acc: 82.1%, F1: 76.1%, Rec: 71.8%, Pre: 80.9%, AUC: 85.3%), demonstrating the benefit of combining the SGA module with the CBF loss. This component-level gain was further supported by the paired AUC test, with the complete SGA+CBF configuration significantly outperforming the SGA-only variant ($p = 0.0064$).

Table 4. Ablation Study A: Input Modality Ablation. The best and second-best results are highlighted in **bold** and underline, respectively.

Modality			**Metrics**				
MG	US	Clinical	Acc (%)	F1 (%)	Rec (%)	Pre (%)	AUC (%)
✓			69.5±2.4	54.2±2.8	52.5±2.6	61.2±2.7	68.5±2.5
	✓		70.8±2.1	53.5±2.5	52.0±2.4	60.5±2.2	67.4±2.3
✓	✓		80.2±1.8	74.8±2.0	**73.2±1.8**	75.5±1.9	81.8±1.7
	✓	✓	78.5±1.6	67.2±1.8	66.5±1.7	74.2±1.6	80.5±1.5
✓	✓	✓	**82.1±1.7**	**76.1±2.0**	71.8±2.6	**80.9±1.7**	**85.3±3.2**

4 Conclusion

In this study, we propose the MSHF-Net for non-invasive, preoperative prediction of breast cancer molecular subtypes. The proposed H-TMF integrates complementary information from MG, US, and clinical data. In addition, the SGA module and the CBF loss improve lesion representation while reducing the impact of subtype imbalance. Experiments on the self-collected LGCH-Breast dataset show that MSHF-Net achieves strong performance and consistently outperforms competitive baseline methods. Rather than aiming to build a complete

Table 5. Ablation Study B: Component Ablation. The best and second-best results are highlighted in **bold** and underline, respectively.

Component		Metrics				
SGA module	CBF Loss	Acc (%)	F1 (%)	Rec (%)	Pre (%)	AUC (%)
✗	✗	75.8±2.1	63.6±2.9	59.5±3.1	68.2±2.8	77.4±2.5
✓	✗	77.5±1.9	69.3±2.3	<u>68.2±2.5</u>	70.4±2.4	<u>80.5±2.1</u>
✗	✓	<u>81.6±1.8</u>	<u>69.4±2.2</u>	62.8±2.7	<u>77.5±2.1</u>	78.1±2.3
✓	✓	**82.1±1.7**	**76.1±2.0**	**71.8±2.6**	**80.9±1.7**	**85.3±3.2**

segmentation-classification pipeline, this work investigates whether high-quality lesion masks can enhance molecular subtype prediction. Future work will evaluate the robustness of MSHF-Net under automatic or noisy masks and validate its generalizability using multi-center data.

Acknowledgments. This work was funded by the Guangxi Science and Technology Program (No. FN2504240022), Guangxi Key R&D Project (No. AB24010167), the Fundamental Research Funds for the Provincial Universities of Zhejiang (No. GK259909299001-006), and Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011617).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ben Rabah, C., Sattar, A., Ibrahim, A., Serag, A.: A multimodal deep learning model for the classification of breast cancer subtypes. Diagnostics **15**(8), 995 (2025)
2. Curigliano, G., Burstein, H., Gnant, M., Loibl, S., Cameron, D., Regan, M., Denkert, C., Poortmans, P., Weber, W., Thürlimann, B., et al.: Understanding breast cancer complexity to improve patient outcomes: the st gallen international consensus conference for the primary therapy of individuals with early breast cancer 2023. Annals of Oncology **34**(11), 970–986 (2023)
3. 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)
4. 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)
5. 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)
6. Holmes, D.R.: Reducing the risk of needle tract seeding or tumor cell dissemination during needle biopsy procedures. Cancers **16**(2), 317 (2024)

7. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
8. Kiela, D., Bhooshan, S., Firooz, H., Perez, E., Testuggine, D.: Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950 (2019)
9. Korde, L.A., Somerfield, M.R., Carey, L.A., Crews, J.R., Denduluri, N., Hwang, E.S., Khan, S.A., Loibl, S., Morris, E.A., Perez, A., et al.: Neoadjuvant chemotherapy, endocrine therapy, and targeted therapy for breast cancer: Asco guideline. Journal of Clinical Oncology **39**(13), 1485–1505 (2021)
10. Li, C., Huang, H., Chen, Y., Shao, S., Chen, J., Wu, R., Zhang, Q.: Preoperative non-invasive prediction of breast cancer molecular subtypes with a deep convolutional neural network on ultrasound images. Frontiers in Oncology **12**, 848790 (2022)
11. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
12. Li, R., Wu, X., Li, A., Wang, M.: Hfbsurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. Bioinformatics **38**(9), 2587–2594 (2022)
13. Loibl, S., Poortmans, P., Morrow, M., Denkert, C., Curigliano, G.: Epidemiology and risk factors. Breast Cancer. Lancet **397**, 1750–1769 (2021)
14. Luo, Y., Wei, J., Gu, Y., Zhu, C., Xu, F.: Predicting molecular subtype in breast cancer using deep learning on mammography images. Frontiers in Oncology **15**, 1638212 (2025)
15. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., et al.: Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), e210315 (2022)
16. Mota, A.M., Mendes, J., Matela, N.: Breast cancer molecular subtype prediction: a mammography-based ai approach. Biomedicines **12**(6), 1371 (2024)
17. Qi, Y.J., Su, G.H., You, C., Zhang, X., Xiao, Y., Jiang, Y.Z., Shao, Z.M.: Radiomics in breast cancer: current advances and future directions. Cell Reports Medicine **5**(9) (2024)
18. 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)
19. Raghu, A., Raghu, A.: Metafusion: a novel method for integrating clinical metadata with imaging modalities for medical applications. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
20. Schwarze, K., Buchanan, J., Taylor, J.C., Wordsworth, S.: Are whole-exome and whole-genome sequencing approaches cost-effective? a systematic review of the literature. Genetics in Medicine **20**(10), 1122–1130 (2018)
21. Shen, H., Zheng, Q., Pan, J., Jin, Y., Zheng, X., Yuan, Q., Tan, D., Zhou, Q., Wang, J., Sun, T.: Intra-tumor heterogeneity-resistant gene signature predicts prognosis and immune infiltration in breast cancer. Frontiers in Immunology **16**, 1598858 (2025)
22. Siegel, R.L., Kratzer, T.B., Wagle, N.S., Sung, H., Jemal, A.: Cancer statistics, 2026. Ca **76**(1), e70043 (2026)

23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)
24. Wu, L., Shan, X., Ge, R., Liang, R., Zhang, C., Li, Y., Elazab, A., Luo, H., Liu, Y., Wang, C.: Tmi-clnet: triple-modal interaction network for chronic liver disease prognosis from imaging, clinical, and radiomic data fusion. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
25. Yin, L., Ye, C., Liu, T., Wu, J., Yan, T.: Unicross: Balanced multimodal learning for alzheimer's disease diagnosis by uni-modal separation and metadata-guided cross-modal interaction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 638–648. Springer (2025)
26. Zhang, T., Tan, T., Han, L., Appelman, L., Veltman, J., Wessels, R., Duvivier, K.M., Loo, C., Gao, Y., Wang, X., et al.: Predicting breast cancer types on and beyond molecular level in a multi-modal fashion. NPJ Breast Cancer **9**(1), 16 (2023)
27. Zhang, X., Li, H., Wang, C., Cheng, W., Zhu, Y., Li, D., Jing, H., Li, S., Hou, J., Li, J., et al.: Evaluating the accuracy of breast cancer and molecular subtype diagnosis by ultrasound image deep learning model. Frontiers in Oncology **11**, 623506 (2021)