

HER2 Expression Prediction with Flexible Multi-Modal Inputs via Dynamic Bidirectional Reconstruction

Anonymous Author(s)

ABSTRACT

In the field of HER2 assessment for breast cancer, traditional detection models predominantly accommodate either H&E or IHC imaging in isolation, whereas clinical precision often hinges on synergistic analysis of both modalities. However, acquiring dual-modality imaging for the same patient is frequently constrained by complex clinical workflows and prohibitive costs. To address this challenge, we propose an adaptive bimodal input prediction framework that flexibly supports both single- and dual-modality inputs. This framework overcomes the rigid dependency on input completeness in existing models through a dynamic branch selection mechanism, enabling predictions to be initiated with either H&E or IHC images alone while retaining dual-modality joint inference capabilities to adapt to diverse resource scenarios. The core technical innovation involves a missing-modality branch selector that dynamically activates either a single-modality reconstruction-prediction pipeline or an end-to-end dual-modality joint inference process based on actual inputs, coupled with a bidirectional cross-modal generative adversarial network (CM-GAN) that achieves context-aware reconstruction of missing modalities in feature space. This design elevates single-modality H&E prediction accuracy from 71.44% to 94.25%, substantially mitigating performance degradation caused by incomplete information. Experimental validation demonstrates that the framework achieves 95.09% prediction accuracy under sufficient dual-modality conditions while maintaining 90.28% reliability with single-modality inputs. By adopting this "dual-modality preferred, single-modality compatible" elastic architecture, healthcare institutions can attain near-bimodal analytical precision without mandating synchronized acquisition of both modalities, particularly offering a low-cost clinical solution for regions with limited IHC staining infrastructure, thereby significantly enhancing the accessibility of HER2 detection.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences; Health informatics.**

KEYWORDS

* Article Title Footnote needs to be captured as Title Note

†Author Footnote to be captured as Author Note

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

WOODSTOCK'18, June, 2018, El Paso, Texas USA

© 2018 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06...\$15.00

<https://doi.org/10.1145/1234567890>

Optional dual-modality input; Multi-modal fusion; Dynamic feature reconstruction; HER2 prediction

ACM Reference format:

Anonymous Author(s). 2025. HER2 Expression Prediction with Flexible Multi-Modal Inputs via Dynamic Bidirectional Reconstruction. In *Proceedings of ACM Woodstock conference (WOODSTOCK'25)*. ACM, New York, NY, USA, 7 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Breast cancer is the most common malignancy among women worldwide, making accurate molecular subtype evaluation critical for guiding personalized treatment[1]. Human epidermal growth factor receptor 2 (HER2) is a key biomarker in breast cancer; its expression level directly influences the selection of targeted therapies[2,3]. Precise assessment of HER2 status is thus pivotal. Traditional clinical HER2 testing relies on immunohistochemistry (IHC) and in-situ hybridization (ISH) techniques. While these methods are widely used, they have notable limitations. IHC interpretation is subject to pathologist expertise and can suffer from inter-observer and inter-laboratory variability[5], and even newly introduced scoring categories (e.g. the "HER2-low" subtype) underscore the challenge of consistent interpretation of borderline cases[4,13]. ISH (including FISH/CISH) provides gene amplification information but involves complex protocols and expensive reagents, limiting its use in resource-constrained settings[6]. Furthermore, the standard IHC/ISH workflow can take several days to return results, delaying clinical decision-making. These shortcomings have spurred interest in automated, AI-driven HER2 evaluation to improve objectivity and speed[7,8].

Early efforts toward automation included computer-assisted image analysis for HER2 IHC scoring[7]. More recently, deep learning techniques have demonstrated great potential in pathology. For instance, fully digital HER2 scoring systems have been developed to improve consistency in IHC interpretation[8]. However, existing AI models largely handle each modality in isolation and thus inherit the limitations of single-modality data. H&E-stained histology slides reveal tissue morphology but lack direct protein expression information, whereas IHC highlights HER2 protein distribution but can be confounded by staining variability. Several studies have attempted to predict HER2 status from a single modality: some leveraged only H&E images[11] (to predict HER2 positivity or treatment response), and others used only IHC images with deep learning-based scoring[10]. While these approaches reported encouraging results, they often suffer from insufficient accuracy and robustness due to the inherent information gap of using one modality alone[12]. In principle,

multi-modal fusion of H&E and IHC should offer complementary information – capturing both morphological context and molecular expression – and thus could overcome single-modality bottlenecks.

Indeed, combining complementary data modalities is increasingly recognized as a key to improving predictive performance in oncology[17]. For example, multi-modal AI systems have achieved notable success in other breast cancer applications: McKinney et al.[18] developed a deep learning system that surpassed radiologist accuracy in screening mammography by integrating multiple views and clinical context, and Joo et al.[19] fused MRI scans with clinical data to accurately predict pathological complete response to neoadjuvant chemotherapy. However, most multi-modal frameworks assume that all modalities are always available, which is often not the case in real clinical workflows where one modality may be missing. A common real-world scenario in pathology is that an H&E slide is readily available, but the corresponding IHC slide might not be obtained due to time, cost, or tissue constraints. Current methods have no mechanism to gracefully handle such missing-modality situations: a dual-modality model would simply fail or resort to naive imputation (e.g. zero-filling or mean values) that can distort feature distributions and degrade performance. Moreover, existing fusion strategies typically use static or fixed weighting of modalities, which cannot adapt to variations in data quality – for instance, an IHC image may be faint or overstained, in which case the model should rely more on the H&E information, but static fusion cannot make such adjustments.

In this work, we propose a novel HER2 prediction model that is both multi-modal and modality-flexible. Our approach leverages the strengths of multi-modal learning while addressing the practical challenges of missing modality and variable data quality. In summary, our contributions are as follows:

- To address the dynamic issue of missing image modalities in practical pathological applications, this study proposes a Missing Modality Branch Selector. It employs a lightweight binary classifier to detect the completeness of the input modalities in real time, thereby dynamically activating the corresponding reconstruction branch for the absent modality.
- An innovative dual-encoder architecture is introduced to disentangle shared and modality-specific features.
- We introduced a Bidirectional Cross-Modal Generative Adversarial Network to perform context-aware reconstruction of the missing modality.
- A Channel Attention Module (CAM) is integrated to adjust feature weights in real time based on the quality of the input modalities.

2 Related Work

Automated HER2 Scoring from Single Modality: Given the limitations of subjective manual scoring, researchers have explored automated HER2 evaluation on individual data modalities. Early work by Masmoudi et al. [7] developed an image analysis algorithm for objective quantification of IHC

HER2 expression, representing one of the first attempts at computer-aided HER2 scoring. In the deep learning era, most efforts have focused on a single image modality at a time. Several studies target IHC slide analysis: Che et al. [10] trained a convolutional neural network to recognize HER2-positive vs. negative tumor cells on IHC whole-slide images, achieving high concordance with human pathologists. Xiong et al. [8] proposed a comprehensive AI system for HER2 scoring on IHC slides, which improved scoring consistency across slides and laboratories. Similarly, Chauhan et al. [9] introduced a deep learning approach that contrasts multi-resolution features from IHC slides to refine HER2 scoring, demonstrating that combining high- and low-magnification features can enhance classification of 0/1+ vs. 2+ vs. 3+ cases. These IHC-focused methods benefit from the specificity of the stain but do not utilize the morphological context available in H&E.

On the other hand, some works have attempted to predict HER2 status from H&E images alone, aiming to infer molecular expression from morphology. Farahmand et al. [11] trained a deep model on H&E tumor regions to predict HER2 positivity and response to anti-HER2 therapy; notably, their H&E-based model could identify some HER2-positive tumors (with an AUC of ~0.88) but was inevitably limited by the absence of direct protein information. Rasmussen et al. [12] focused on the challenging scenario of IHC-equivocal (2+) cases: they developed a model using H&E features to predict final HER2 status for cases that are 2+ by IHC, as a decision support tool. While H&E-driven models are attractive since H&E slides are ubiquitously available, their performance tends to lag behind IHC-based methods due to the intrinsic lack of HER2 expression signals in H&E. In summary, single-modality approaches – whether H&E or IHC – face a performance ceiling because each provides only a partial view of the underlying biology [12]. This motivates combining modalities to leverage complementary information.

Multi-Modal Learning and HER2 Assessment: Multi-modal data fusion has shown promise in various medical imaging tasks, as integrating heterogeneous data can provide a more complete characterization of disease [17]. In the context of breast cancer, beyond pathology, there are examples like McKinney et al. [18] who combined multiple image views and clinical insights for improved cancer detection, and Joo et al. [19] who fused imaging with clinical variables for outcome prediction. However, within pathological HER2 assessment, multi-modal learning is still nascent. One notable work bridging H&E and IHC is by Liu et al. [14], who introduced the BCI (Breast Cancer Immunohistochemical) dataset containing thousands of paired H&E and HER2 IHC image patches. They proposed a pyramid Pix2Pix generative model to translate H&E images to IHC images, demonstrating the feasibility of cross-stain image synthesis for HER2. While their focus was on image generation (augmenting or predicting IHC appearance from H&E) rather than direct HER2 scoring, it established a foundation for cross-modal approaches. To our knowledge, no prior published method performs joint H&E+IHC fusion for HER2 classification in an end-to-end

manner. Traditional pathology workflows simply interpret the two slides side by side, and existing AI models have yet to fully emulate or improve upon this multi-modal interpretation.

Missing Modality and Dynamic Fusion Strategies: A practical obstacle to multi-modal pathology is that one of the modalities may be missing or unavailable. Simplistic solutions like imputing missing inputs with zero or average values are commonly used but often degrade performance by introducing unnatural features. In general machine learning literature, handling missing modalities has been addressed by learning robust representations. Wang et al.

【15】, for example, proposed a shared-specific feature modeling approach: their framework learns latent features that are common to all modalities as well as features specific to each modality, enabling the model to make predictions even when one modality is absent. This approach, applied to vision tasks in their work, inspires our use of separate shared and specific feature extractors. In the medical imaging domain, however, approaches like 【15】 have not been applied to pathology, and challenges like variable stain quality remain. Another line of research involves using generative models to impute missing modality data. Our use of a Pix2Pix-based GAN aligns with this idea: by generating a pseudo-IHC image from H&E (or vice versa), we effectively fill in the missing information with a data-driven guess rather than a constant fill-in. This is akin to data augmentation of modality, and is more sophisticated than linear interpolation methods attempted in some contexts (which fail to capture complex non-linear relationships across modalities).

Equally important in multi-modal fusion is how to combine information from each modality. Many existing multi-modal models rely on simple concatenation or fixed fusion rules, which treat each modality's contribution as static. This is suboptimal when modality relevance varies from case to case. Attention mechanisms have emerged as a powerful way to perform adaptive fusion. The Convolutional Block Attention Module (CBAM) by Woo et al. 【16】 is a representative technique that applies channel-wise and spatial attention to recalibrate feature importance within a CNN. CBAM and similar attention modules can be applied to multi-modal features to dynamically adjust their weights. In our problem, we expect that if an IHC image is faint or noisy, the model should down-weight IHC-specific features and rely more on H&E features, and vice versa. Inspired by CBAM 【16】, we implement a channel attention-based fusion that learns to emphasize more informative features and suppress less reliable ones in a data-driven manner. To the best of our knowledge, our work is the first to integrate a dynamic attention fusion with a missing-modality-tolerant architecture in the medical imaging field. By uniting these components – cross-modal generation, shared-specific encoders, and attention-based fusion – our framework addresses the key shortcomings identified in prior work: it fully exploits multi-modal data when present, and remains robust when one modality is absent or low-quality.

3 Methods

This study proposed a prediction model of HE/IHC incomplete input HER2 expression based on dynamic bidirectional reconstruction. Its core architecture is shown in Figure 1, including the following modules:

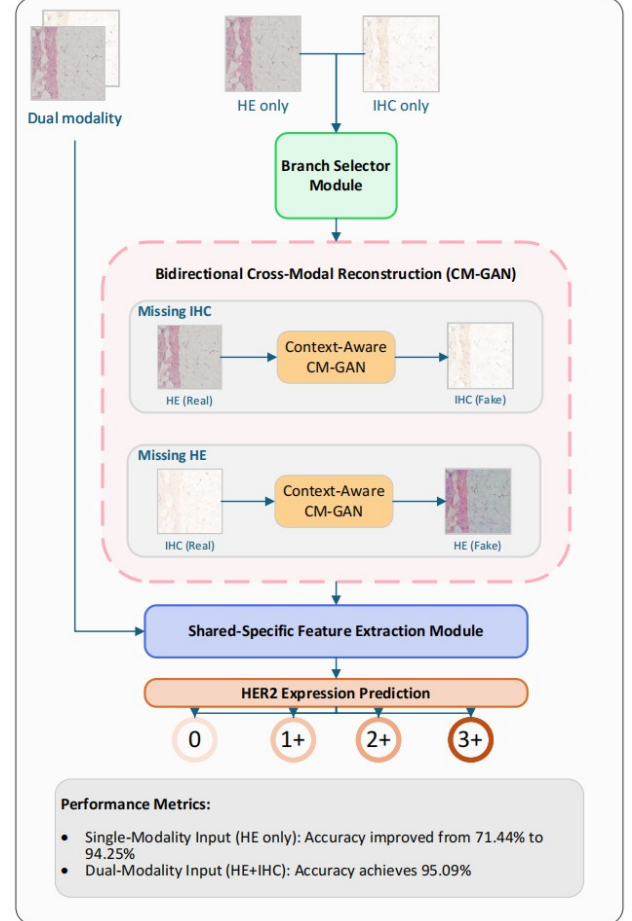


Figure 1: HE/IHC incomplete input HER2 expression prediction model based on dynamic bidirectional reconstruction.

3.1 Missing Modality Branch Selection

The dual-modality optional input module constitutes a key innovation in achieving flexible inference in our framework by employing a multi-branch architecture with a dynamic path selection mechanism to build an adaptive processing workflow tailored to real-world clinical scenarios. This module utilizes a two-stage hierarchical processing strategy for intelligent adaptation of input modalities.

In the first stage, a two-phase classifier is built based on an improved GoogLeNet architecture; it initially employs a global average pooling layer to extract the spatial statistical features of the input image, followed by a fully connected layer to determine whether the input is single-modality (either H&E or IHC) or dual-

modality (H&E + IHC). When a single-modality input is detected, a subsequent shallow convolutional network further analyzes texture features to precisely distinguish between H&E and IHC image types. In the second stage, the corresponding processing pathway is activated based on the classification outcome: if a dual-modality input is identified, an end-to-end joint inference branch is triggered, wherein the H&E and IHC images are concatenated channel-wise and forwarded to subsequent modules; if a single-modality input is detected, a cross-modal reconstruction-prediction cascade is initiated, invoking a pre-trained bidirectional generative adversarial network to replenish the missing modality in the feature space. This "discrimination-reconstruction-alignment" three-stage processing mechanism enables the model to dynamically construct the optimal feature representation based on the available input modalities, thereby preserving the advantages of dual-modality joint inference while significantly enhancing prediction robustness in single-modality scenarios.

3.2 Bidirectional Cross-Modal Reconstruction (CM-GAN)

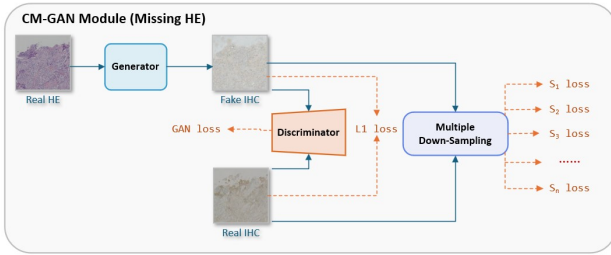


Figure 2: Structure diagram of bidirectional reconstruction module.

This module employs a context-aware generative adversarial network to achieve semantically consistent completion of missing modalities. For example, when only the HE modality is provided, the HE→IHC reconstruction model is invoked; conversely, if only IHC data is available, the IHC→HE reconstruction model is activated. As depicted in Figure 2, when the input consists solely of an HE image, a pyramid Pix2pix architecture is utilized to reconstruct an IHC image from the HE feature space [14]. Specifically, hierarchical features of the HE image are first extracted using multi-scale residual blocks, after which a spatial attention mechanism is employed to locate key regions. Low-level details and high-level semantic information are then fused via skip connections to generate the reconstructed IHC image. Finally, the reconstructed image is combined with the original modality and fed into subsequent classification modules.

3.3 Shared and Specific Feature Encoders

Similarly, this module implements a dual-path encoding architecture to decouple and synergize cross-modal common and specific features[15]. As shown in Figure 3, the module utilizes a pre-trained ResNet50 as a shared encoder to extract the cross-

modal common features (denoted as F_s) from both HE and IHC images, thereby capturing the shared patterns of tissue morphology and molecular expression. Concurrently, to preserve modality-specific information, two independent MobileNetV3 branches serve as dedicated encoders, extracting texture topological features (F_{he}) from HE images and protein distribution features (F_{ihc}) from IHC images.

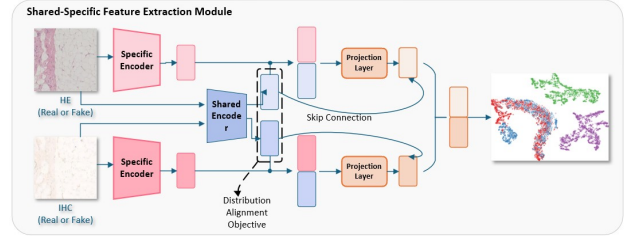


Figure 3: Sharing - specific feature extraction module structure principle.

To further enhance the feature decoupling, a joint optimization strategy incorporating Domain Classification Loss (DCO) and Distribution Alignment Loss (DAO) is employed. The DCO enforces the specific features to differentiate between HE and IHC modalities via binary cross-entropy, whereas the DAO leverages Maximum Mean Discrepancy (MMD) to constrain the shared feature distributions across different modalities.

3.4 Channel Attention-based Dynamic Weight Fusion Module

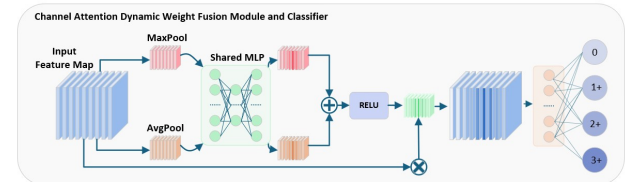


Figure 4: Structure diagram of Channel Attention-based Dynamic Weight Fusion Module.

The Channel Attention Module (CAM) intelligently fuses the aforementioned shared feature (r_{he}, r_{ihc}) with the modality-specific features (s_{he}, s_{ihc}) through adaptive weight assignment[16]. Initially, the module concatenates the shared feature F_s with the specific features F_{he}, F_{ihc} along the channel dimension, forming a multimodal feature sequence. To mitigate the risk of dimensionality explosion, a 1×1 convolutional kernel is applied to compress the number of channels while preserving critical feature information. Subsequently, a channel attention mechanism is employed with a hidden layer dimension configured to 125. This mechanism dynamically computes the weights for each modality's features, facilitating real-time adaptive adjustment based on the input data quality. For instance, if the image quality of one

modality (e.g., IHC) is comparatively low, the model automatically assigns a higher weight to the other modality (e.g., HE) when computing cross-modal feature correlations, thereby enhancing the efficacy of the fused features.

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \\ = \sigma(W_1(W_o(F_{avg}^c)) + W_1(F_{max}^c))$$

After processing through the CAM, the weighted fusion yields the final fused feature F_{fused} , which is then utilized in subsequent classification prediction tasks.

3.5 Classification and Loss Functions

The module achieves precise grading through multi-granularity feature aggregation and adaptive loss weighting. As illustrated in Figure 6, the fused feature F_{fused} is initially processed by a dual-path pooling mechanism comprising global and local branches. In the global branch, global average pooling (GAP) is employed to extract the overall representation, while the local branch utilizes 3×3 max pooling to capture critical regional details. The features from both branches are concatenated and subsequently fed into a classifier composed of fully connected layers, which outputs a four-class probability vector corresponding to HER2 expression levels of 0, 1+, 2+, and 3+.

$$L_{total} = L_{cls} + \alpha L_{dco}$$

To mitigate the issue of class imbalance, the primary classification task is optimized using a weighted cross-entropy loss, denoted as L_{cls} . In addition, the overall training process jointly optimizes an auxiliary task loss that incorporates a distribution alignment loss, L_{dco} , thereby further enhancing the robustness of feature extraction and fusion alongside classification performance.

4 Experiments

In this section, to validate the effectiveness of our proposed HER2 expression prediction model based on dynamic bidirectional reconstruction with incomplete HE/IHC inputs, we first conducted comparative experiments with traditional imputation strategies, evaluating its performance improvement in HER2 expression level prediction. Subsequently, ablation studies were performed to verify the specific contributions of each module to the overall model performance, thereby supporting our assertion that the synergistic effects of these modules significantly enhance prediction accuracy and generalization capability under conditions of missing modalities.

4.1 Datasets and Evaluation Metrics

Datasets. This study employs the BCI Breast Cancer Pathology dataset (refer to Liu et al., 2018), which comprises 4,870 pairs of strictly matched full-slice HE and IHC images. Each image pair is explicitly annotated with HER2 expression levels (0, 1+, 2+, and 3+). The dataset is clinically significant as it covers a spectrum of HER2 expression states, from negative to strongly positive, and the images authentically reflect the morphological characteristics of pathological tissues alongside HER2 protein expression. This

renders the dataset ideal for multimodal fusion research and experiments simulating modality dropout.

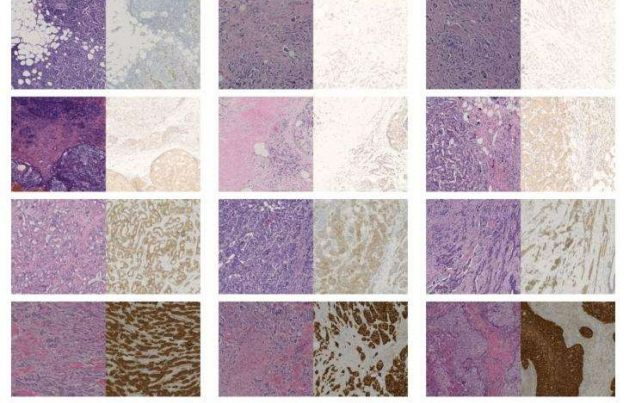


Figure 5: The data set used in the experiment.

Data Preprocessing. To standardize input dimensions and enhance model generalization, all images underwent normalization and were resized to a uniform resolution. Furthermore, based on the HER2 grading information embedded within the image filenames, labels were converted into one-hot encoded vectors. In addition, to simulate potential modality dropout issues encountered in real clinical scenarios during the testing phase, we devised two testing scenarios: fixed modality combinations (HE-only, IHC-only, HE+IHC) and dynamic dropout combinations (where the HE or IHC modality is randomly dropped with a 20% probability). This approach comprehensively evaluates model performance under various modality input conditions.

Performance Metrics. To accurately assess the classification performance of the model, we employed a comprehensive suite of metrics, including Accuracy, Recall, Precision, F1-score, and confusion matrix analysis. The F1-score, in particular, provides an integrated measure of the model's sensitivity to class imbalance and its robustness. Furthermore, t-SNE visualization of the feature space was used to explore the model's interpretability, offering an intuitive understanding of how the model utilizes modality features and allocates weights during the prediction process.

4.2 Implementation Details

To ensure the fairness and reproducibility of our experimental results, all experiments were conducted under a standardized hardware and software environment. The hardware configuration includes a server equipped with an Intel® Xeon® Platinum 8352V CPU @ 2.10GHz and an NVIDIA RTX 4090 GPU (24GB memory). On the software front, the experimental environment was built on Ubuntu 18.04, using Python 3.8 along with PyTorch 2.4 as the core framework. During model training, the AdamW optimizer was employed, with the initial learning rate set to [value] and augmented by a weight decay of [value]. The learning rate

was appropriately adjusted based on validation set performance at different stages to prevent model overfitting.

4.3 Comparative Experiments

This section systematically validates the efficacy of the proposed approach through two sets of controlled experiments. All experiments were conducted under a unified data split (3896:977) and with consistent hyperparameter settings (AdamW optimizer, batch size of 8) to ensure result comparability.

Unimodal vs. Cross-Modal Reconstruction Enhancement. In the unimodal baseline method, only real HE or IHC images were used as input, with MobileNetV2 serving as the backbone network. In contrast, the cross-modal reconstruction enhancement method employs a modified Pix2pix framework to generate cross-modal images (e.g., Fake IHC or Fake HE). The real modality and the reconstructed modality are then combined as input, and feature-level fusion is achieved using a shared-specific feature extraction module. As shown in Table 1, the cross-modal reconstruction method significantly outperforms the unimodal baseline. For example, the Real HE + Fake IHC combination achieved an accuracy of 94.25%, an improvement of 22.81% over the HE unimodal baseline (71.44%), and the F1-score increased to 0.9609. Similarly, the Real IHC + Fake HE combination achieved an accuracy of 90.28%, which is 12.90% higher than the IHC unimodal baseline (77.38%). These results demonstrate that cross-modal generation can effectively compensate for the insufficiency of unimodal information, thereby enhancing the model's generalization capability.

Table 1: Performance Comparison between Unimodal and Cross-Modal Reconstruction Methods.

Method Type	Accuracy	F1-score	Improvement (vs. Unimodal)
HE Unimodal	71.44%	0.6697	-
IHC Unimodal	77.38%	0.7646	-
Real HE + Fake IHC (ours)	94.25%	0.9609	+22.81 (vs. HE Unimodal)
Real IHC + Fake HE (ours)	90.28%	0.9251	+12.90 (vs. IHC Unimodal)

Table 2: Performance Comparison of Dual-Modal Fusion Strategies.

Input Type	Accuracy	F1-score	Improvement (vs. Concatenation Baseline)
simple concat	92.00%	0.9103	-
Real HE + Real IHC (ours)	95.09%	0.9532	+4.29 (vs. simple concat)

Dual-Modal Concatenation vs. Shared-Specific Feature Fusion. The dual-modal concatenation baseline fuses Real HE

and Real IHC images using a simple concatenation operation, employing the same backbone network as in the unimodal scenario. In contrast, the shared-specific feature fusion method jointly inputs Real HE and Real IHC and achieves refined feature interaction through the decoupling of modality-shared and modality-specific features. As reported in Table 2, the shared-specific feature fusion method reached an accuracy of 95.09%, an improvement of 4.29% over the concatenation baseline (92.00%), with the F1-score rising to 0.9532. This indicates that simple concatenation only facilitates shallow feature interaction, whereas the shared-specific architecture effectively captures cross-modal complementary information and suppresses noise interference between modalities.

4.4 Ablation Studies

The ablation experiments comprehensively validated the contributions of key modules to the overall model performance. In these experiments, a baseline model that omitted the channel attention module (CAM) was compared against the full model that incorporated this dynamic weighting module. In the baseline, shared and specific features were directly concatenated and subsequently reduced in dimension using a 1×1 convolution, while the full model utilized a CAM module, which leverages a multi-head self-attention mechanism to dynamically adjust modality weights in real time. The results demonstrated that the full model achieved an accuracy of 95.32% and an F1-score of 0.9532, marking a 4.91% improvement in the weighted F1-score compared to the baseline model's accuracy of 92.41% and F1-score of 0.9041. These findings underscore that the CAM module significantly enhances feature fusion effectiveness, thereby boosting the model's overall predictive performance under conditions of missing modalities.

Table 3: Ablation Experiment Results for CAM Removal.

CAM Inclusion	Accuracy	F1-score	Improvement
Yes	95.32%	0.9532	+4.91
No	92.41%	0.9041	-

5 Conclusion

The dual-modal optional input model proposed in this study leverages dynamic branch selection and bidirectional cross-modal reconstruction mechanisms to achieve flexible HER2 prediction under both unimodal and dual-modal conditions, effectively overcoming the rigid dependency of traditional methods on complete modality inputs. Notably, the model attains an accuracy of 95.09% with dual-modal input and maintains a high performance of 94.45% even with solely HE input, thereby significantly reducing clinical reliance on IHC staining equipment.

A key limitation of this study is that all data were sourced from a single institution. Although data augmentation was employed to

simulate variability, the model has not yet been evaluated on external datasets collected from diverse institutions or scanners. In future work, we plan to conduct multicenter evaluations to verify the model's generalizability across different staining protocols and imaging devices. In such cases, maintaining high performance may require integrating domain adaptation strategies or fine-tuning the model on new data. Furthermore, we aim to extend this dynamic multimodal approach to other pathological tasks. Promising directions include combining H&E images with genomic data or radiological images for outcome prediction, or integrating multiple immunohistochemical markers within a single model. The core concept of "on-demand activation" is anticipated to be highly beneficial in applications characterized by inconsistent data availability.

REFERENCES

- [1] Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., et al. (2017). Risk Factors and Preventions of Breast Cancer. *International Journal of Biological Sciences*, 13(11), 1387–1397. DOI: 10.7150/ijbs.21635
- [2] Waks, A. G., & Winer, E. P. (2019). *Breast Cancer Treatment: A Review*. JAMA, 321(3), 288–300. DOI: 10.1001/jama.2018.19323
- [3] Ahn, S., Woo, J. W., Lee, K., & Park, S. Y. (2020). HER2 Status in Breast Cancer: Changes in Guidelines and Complicating Factors for Interpretation. *Journal of Pathology and Translational Medicine*, 54(1), 34–44. DOI: 10.4132/jptm.2019.09.28
- [4] Viale, G., & Bardia, A. (2023). HER2-low breast cancer – Diagnostic challenges and opportunities. *Targeted Oncology*, 18, 1–9. DOI: 10.1007/s11523-022-00918-3
- [5] Press, M. F., Hung, G., Godolphin, W., & Slamon, D. J. (1994). Sensitivity of HER-2/neu Antibodies in Archival Tissue Samples: Potential Source of Error in Immunohistochemical Studies of Oncogene Expression. *Cancer Research*, 54(10), 2771 – 2777.
- [6] Furrer, D., Sanschagrin, F., Jacob, S., & Diorio, C. (2015). Advantages and Disadvantages of Technologies for HER2 Testing in Breast Cancer Specimens. *American Journal of Clinical Pathology*, 144(5), 686 – 703. DOI: 10.1309/AJCP1L5X1HEZPHJF
- [7] Masmoudi, H., Hewitt, S. M., Petrick, N., Myers, K. J., & Gavrielides, M. A. (2009). Automated Quantitative Assessment of HER-2/neu Immunohistochemical Expression in Breast Cancer. *IEEE Transactions on Medical Imaging*, 28(6), 916 – 925. DOI: 10.1109/TMI.2008.2010406
- [8] Xiong, Z., Liu, K., Liu, S., Feng, J., Wang, J., Feng, Z., et al. (2024). Precision HER2: A Comprehensive AI System for Accurate and Consistent Evaluation of HER2 Expression in Invasive Breast Cancer. *BMC Cancer*, 24(1), 1204. DOI: 10.1186/s12885-024-10485-8
- [9] Chauhan, E., Sharma, A., Sharma, A., Nishadham, V., Ghughyatl, A., Kumar, A., et al. (2025). Contrasting Low and High-Resolution Features for HER2 Scoring using Deep Learning. *arXiv preprint arXiv:2503.22069*.
- [10] Che, Y., Ren, F., Zhang, X., Cui, L., Wu, H., & Zhao, Z. (2023). Immunohistochemical HER2 Recognition and Analysis of Breast Cancer Based on Deep Learning. *Diagnostics*, 13(2), 263. DOI: 10.3390/diagnostics13020263
- [11] Farahmand, S., Fernandez, A. I., Ahmed, F. S., Rimm, D. L., Chuang, J. H., Reisenbichler, E., & Zarrinhalam, K. (2022). Deep Learning Trained on Hematoxylin and Eosin Tumor Region-of-Interest Predicts HER2 Status and Trastuzumab Treatment Response in HER2+ Breast Cancer. *Modern Pathology*, 35(1), 44 – 51. DOI: 10.1038/s41379-021-00872-5
- [12] Rasmussen, S. A., Taylor, V. J., Surette, A. P., Barnes, P. J., & Bethune, G. C. (2022). Using Deep Learning to Predict Final HER2 Status in Invasive Breast Cancers That Are Equivocal (2+) by Immunohistochemistry. *Applied Immunohistochemistry & Molecular Morphology*, 30(10), 668 – 673. DOI: 10.1097/PAI.0000000000000981
- [13] Nicolò, E., Boscolo Bielo, L., et al. (2023). The HER2-low Revolution in Breast Oncology: Steps Forward and Emerging Challenges. *Therapeutic Advances in Medical Oncology*, 15, 17588359231152842. DOI: 10.1177/17588359231152842
- [14] Liu, S., Zhu, C., Xu, F., Jia, X., Shi, Z., & Jin, M. (2022). BCI: Breast Cancer Immunohistochemical Image Generation through Pyramid Pix2Pix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1815 – 1824). DOI: 10.1109/CVPR52688.2022.00186
- [15] Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., & Carneiro, G. (2023). Multi-Modal Learning with Missing Modality via Shared-Specific Feature Modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 15878 – 15887). DOI: 10.1109/CVPRW53098.2023.00310
- [16] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3 – 19). DOI: 10.1007/978-3-030-01234-2_1
- [17] Steyaert, S., Pizurica, M., Nagaraj, D., Khandelwal, P., Hernandez-Boussard, T., Gentles, A. J., & Gevaert, O. (2023). Multimodal Data Fusion for Cancer Biomarker Discovery with Deep Learning. *Nature Machine Intelligence*, 5(4), 351 – 362. DOI: 10.1038/s42256-023-00633-5
- [18] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafi, H., et al. (2020). International Evaluation of an AI System for Breast Cancer Screening. *Nature*, 577(7788), 89 – 94. DOI: 10.1038/s41586-019-1799-6
- [19] Joo, S., Ko, E. S., Kwon, S., Jeon, E., Jung, H., Kim, J. Y., et al. (2021). Multimodal Deep Learning Models for the Prediction of Pathologic Response to Neoadjuvant Chemotherapy in Breast Cancer. *Scientific Reports*, 11(1), 18800. DOI: 10.1038/s41598-021-98408-8