Breast Cancer Research

RESEARCH

Open Access

Transfer learning drives automatic HER2 scoring on HE-stained WSIs for breast cancer: a multi-cohort study

Xiaoping Li[1†], Zhiquan Lin[2†], Chaoran Qiu[1], Yiwen Zhang[1], Chuqian Lei[1], Shaofei Shen[3], Weibin Zhang[4], Chan Lai[5], Weiwen Li[1], Hui Huang[6*] and Tian Qiu[7*]

Abstract

Background Streamlining the clinical procedure of human epidermal growth factor receptor 2 (HER2) examination is challenging. Previous studies neglected the intra-class variability within both HER2-positive and -negative groups and lacked multi-cohort validation. To address this deficiency, this study collected data from multiple cohorts to develop a robust model for HER2 scoring utilizing only Hematoxylin&Eosin-stained whole slide images (WSIs).

Methods A total of 578 WSIs were collected from five cohorts, including three public and two private datasets. Each WSI underwent adaptive scale cropping. The transfer-learning-based probabilistic aggregation (TL-PA) model and multi-instance learning (MIL)-based models were compared, both of which were trained on Cohort A and validated on Cohorts B–D. The model demonstrating superior performance was further evaluated in the neoadjuvant therapy (NAT) cohort. Scoring performance was assessed using the area under the receiver operating characteristic curve (AUC). Correlation between the model scores and specific grades (HER2 levels, pathological complete response (pCR) status, residual cancer burden (RCB) grades) were evaluated using Spearman rank correlation and Dunn's test. Patch analysis was performed with manually defined features.

Results For HER2 scoring, the TL-PA significantly outperformed the MIL-based models, achieving robust AUCs in four validation cohorts (Cohort A: 0.75, Cohort B: 0.75, Cohort C: 0.77, Cohort D: 0.77). Correlation analysis confirmed a moderate association between model scores and manual reader-defined HER2-IHC status ($Coefficient_{(Spearman)} = 0.37$, $P_{(Spearman)} = 0.001$) as well as RCB grades ($Coefficient_{(Spearman)} = 0.45$, $P_{(Spearman)} = 0.0006$). In Cohort NAT, with the non-pCR as the positive control, the AUC was 0.77. Patch analysis revealed a core-to-peritumoral probability decrease pattern as malignancy spread outward from the lesion's core.

Conclusion TL-PA shows robust generalization for HER2 scoring with minimal data; however, it still inadequately capture intra-class variability. This indicates that future deep-learning endeavors should incorporate more detailed annotations to better align the model's focus with the reasoning of pathologists.

Keywords Breast cancer, HER2 scoring, Whole slide images, Multi-cohort, Transfer learning

[†]Xiaoping Li and Zhiquan Lin have contributed equally to this work and share first authorship.

*Correspondence:
Hui Huang
feel709394@qq.com
Tian Qiu
timeqiu@hotmail.com
Full list of author information is available at the end of the article

Introduction

Human epidermal growth factor receptor 2 (HER2) is an important treatment target for breast cancer, which is diagnosed by immunohistochemistry (IHC) examination [1]. An IHC score of 3+ or, an equivocal IHC score of 2+ with subsequent amplification by fluorescence in situ hybridization (FISH) is considered to indicate HER2-positive [2]. Almost 15–20% of breast cancer cases are HER2-positive which could be significantly benefit from anti-HER2 targeted therapy [3]. Hence, accurate identification of HER2 status is critical for breast cancer treatment.

Currently, the gold standard for clinical HER2 detection relies on the IHC and FISH [4]. Potential inter-observer variability stemming from pathologists' experience or equipment quality is a crucial clinical issue, particularly within the HER2-low group, that could significantly impact treatment decisions [5, 6]. To address this challenge, machine learning was introduced to enhance the consistency of HER2 identification on IHC-stained whole slide images (WSIs) [7, 8]. However, IHC examinations remained time-consuming and costly, prompting clinicians to seek a more cost-effective method for HER2 testing. In this context, methods based on Hematoxylin and Eosin (HE)-stained WSIs are explored.

Based on the hypothesis that molecular differences often manifest as morphologic phenotypes at the cell-level, HER2 indicators mayt be more readily discernible in a HE-stained WSIs than in other radiology examinations [9, 10]. In related studies, the selection of the region of interest (ROI) and the approach to feature extraction had emerged as two primary research focuses in computational pathology [11–17]. Given the nonspecific labelling for HER2 expression by HE-staining, researchers attempted to reduce the computational cost associated with gigapixel WSIs by exploring methods such as randomization [13], cell density [15], or tumor masking [12] for ROI selection. An ROI was generally composed of a variable number of patches instead of pixel-level segmentation. Then, patches from the ROI were transferred into a feature extractor. According to the approach of extractor training, models can be divided into a multi-instance learning (MIL) as the back-end [11, 13, 14, 16, 17] and a sub-classifier aggregation as the front-end [12, 15]. Previous studies had illustrated that assessing molecular subtype could profit from deep learning and some proper experience implantation.

Tumor heterogeneity in breast cancer, coupled with external factors like variations in slide preparation, scanning, and annotation, may affect the generalizability of computational pathology models [18, 19]. Consequently, multi-cohort validation has become essential. With advancements in precision medicine, HER2-positive breast cancer should receive anti-HER2 treatment, while HER2-low breast cancer also had specific drug (Trastuzumab Deruxtecan) [20]. As a result, classifying HER2 status into only two categories (positive and negative) was unable to support individualized treatment [3]. Addressing the development of a more practical model based on limited resources has become a meaningful topic.

In this study, HE-stained WSIs from five cohorts were collected. Two frameworks were trained and validated. Our work aims to develop a robust model for HER2 scoring and assess its ability to capture the 5-level ordered patterns consistent with pathologist annotations, offering valuable insights for automatic HER2 scoring.

Materials and methods

Data sources

The HE-stained WSIs of breast cancer cohorts were collected for this study from three public cohorts(cohorts A, B, and NAT) and two private cohorts(cohorts C and D). A summary of data collection was provided in Fig. 1 and Supplementary Table S1, including scanning magnification, scanner type, estrogen receptor (ER) and progesterone receptor (PR) status, HER2 status, and HER2-IHC/FISH status. ER/PR positive was defined as $\geq 1\%$. All specimens were invasive breast carcinoma. IHC were performed through paraffin section. HER2 identification strictly adhered to the American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) clinical practice guidelines.

Cohort A from TCGA-BRCA was employed for model training and validation in a ratio 8:2. Cohort B from the Yale Pathology electronic database [12], cohorts C from Jiangmen Central Hospital, and cohorts D from Jiangmen Maternity and Child Health Care Hospital were used for external validation. Neoadjuvant therapy (NAT) cohort [21], comprising 62 HER2-positive patients treated with doxorubicin/cyclophosphamide/taxol together with anti-HER2 targeted therapy, was utilized to investigate the correlation between model scores and NAT efficacy. Pathological complete response (pCR) and residual cancer burden (RCB) were employed as efficacy indicators.

The training set contained 140 cases. A total of 438 cases were assigned to the validation set, including internal validation (35 cases), external validation (341 cases) and NAT cohort (62 cases).

This study was centrally approved by the Ethical Committee of the Jiangmen Central Hospital (No.[2022] 107) and complied with all relevant ethical regulations. All patients provided written informed consent for the samples used for research.

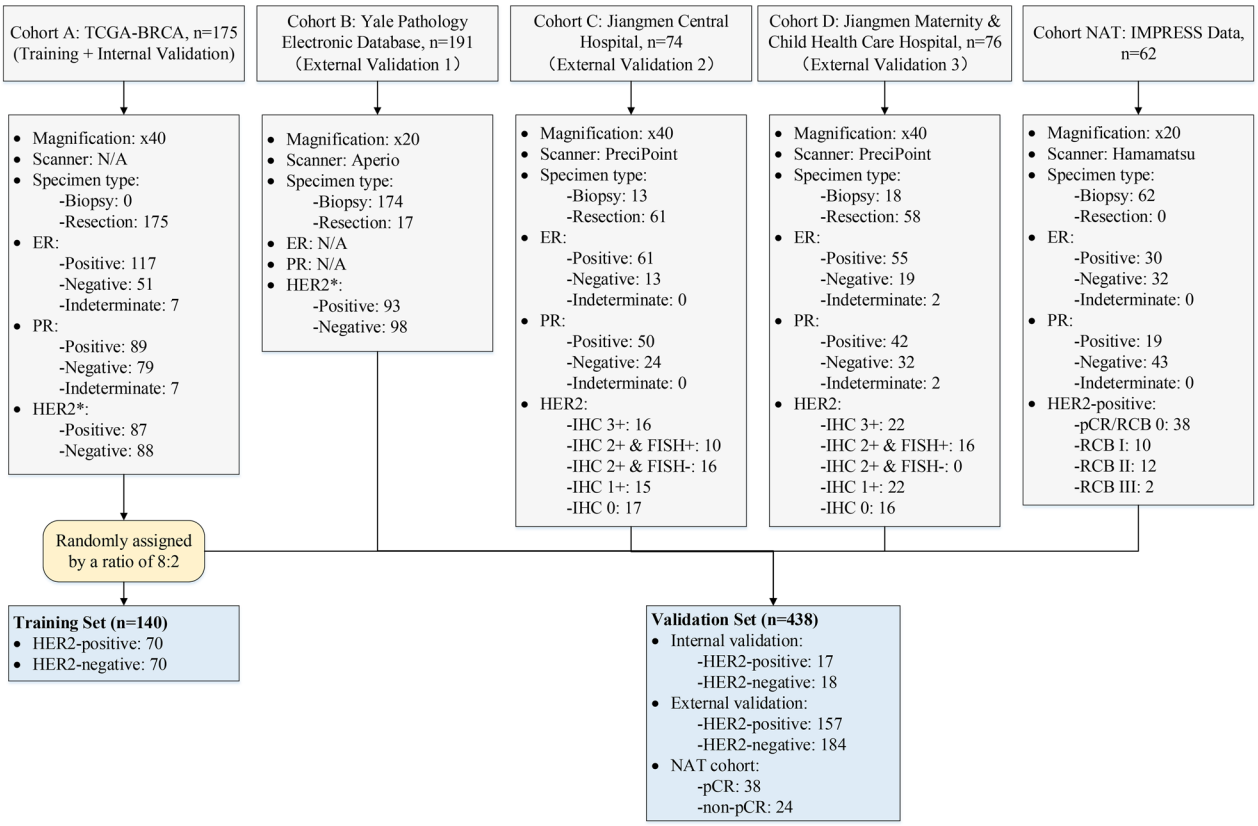

Fig. 1 Data collection. A total of 578 HE-stained WSIs from five BC cohorts were included in this study. Cohort A: 175 cases at a scanning magnification of ×40. Cohort B: 191 cases at a scanning magnification of ×20. Cohort C: 74 cases at a scanning magnification of ×40. Cohort D: 76 cases at a scanning magnification of ×40. Cohort NAT: 62 cases at a scanning magnification of ×20. *The HER2 annotation categorized as positive or negative only. *N/A* Not available

Model frameworks

Two frameworks were evaluated in this study: sub-classifier aggregation and multi-instance learning (MIL). In Framework 1 (sub-classifier aggregation), a sub-classifier was trained to predict the HER2-positive probability for each patch. These probabilities were then aggregated into a WSI-level HER2 score according to a predefined rule. In Framework 2 (MIL), patch features were extracted from a pre-trained backbone model. By sharing a WSI-level label, these features were then combined into a bag-level representation. Subsequently, they were aggregated with an attention-based or graph-based learning approach to predict HER2 status at the WSI level. (Fig. 2).

Sub-classifier aggregation

Transfer learning In Framework 1, the performance of WSI-level prediction largely depended on the patch classifier. To avoid over-fitting occurred within supervised training, transfer learning was introduced [22, 23]. The Swin-T (Swin-transformer-Tiny) was adopted as the carrier of transfer learning for its remarkable generalizability [24, 25]. The training patches were cropped to two sizes (1024 × 1024, 2048 × 2048) for data augmentation. Then, these patches were subjected to stain normalization by Vahadane [26]. The patches in the background were excluded by a variance less than 500. The patches in the tumor mask inherited their original WSI label (HER2±), while the others were labeled Non-tumor. Table 1 showed the numbers of patches included in this transfer learning process.

The Swin-T model was trained using the Adam optimizer with a batch size of 128. The Cross-Entropy loss function was employed. All models in this study were trained on an NVIDIA RTX 3090 Ti 24GB GPU. The Swin-T was fully pre-trained through 200 epochs on the CIFAR-100 dataset. After that, the bottom module was frozen to conduct the transfer learning until the loss plateaued (Supplementary Fig. S1).

Adaptive cropping To minimize the adverse effects introduced by scanning discrepancies, the patch-level view was fixed at (512 × 512) pixels for every gigapixel in

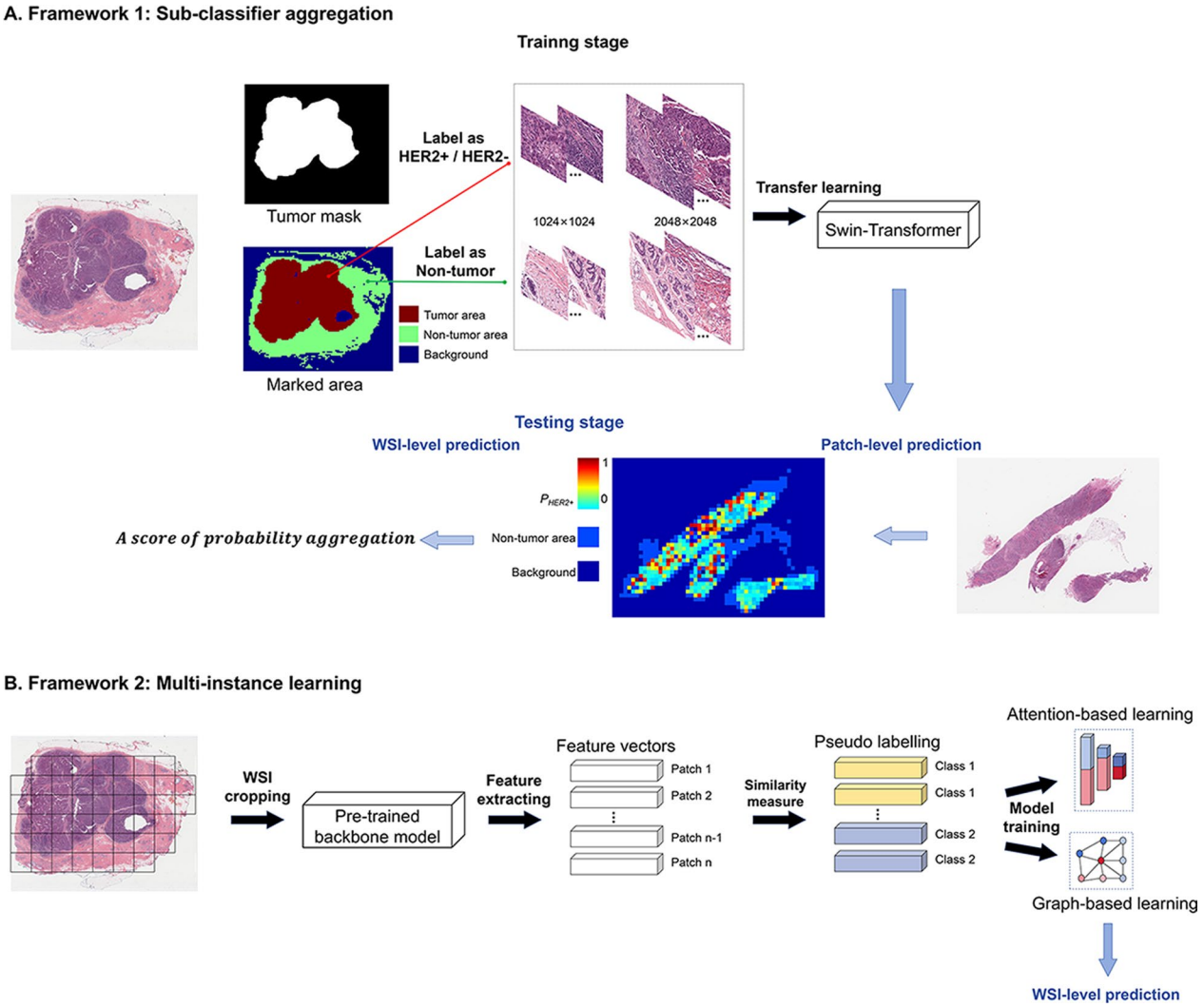

Fig. 2 Model frameworks. **A** Framework 1, Sub-Classifier Aggregation. During training, HE-stained WSIs were cropped into patches of different sizes (1024 × 1024, 2048 × 2048), with background patches (variance < 500) excluded. Tumor patches were selected using manual annotation masks for Swin-T-based transfer training. During testing, patch-level predictions were made using the sub-classifier, and WSI-level predicted scores were obtained through probabilistic aggregation. **B** Framework 2, Multi-Instance Learning (MIL). This general MIL structure included (1) WSI cropping, (2) feature extraction using a pre-trained backbone model, (3) producing pseudo-labels with similarity measure (4) model training based on attention-based or graph-based learning

Table 1 Patches for transfer learning

Patch size	HER2−	HER2+	Non-tumor
1024 × 1024	90,664 (63,464)	97,264 (68,084)	142,336 (69,744)
2048 × 2048	15,441	16,149	21,877 (15,313)

(·) denotes the numbers of patches randomly used for actual training

the validation stage. First, the height (h) and width (w) of each WSI were measured using the Openslide package [27]. Second, the view scale can be calculated by a ratio of total pixels (h × w) to a gigapixel, as shown in Eq. (1).

$$cropping\ size = 512 \times \sqrt{\frac{\mathrm{h} \times \mathrm{w}}{1\mathrm{e}^9}} \tag{1}$$

Patch analysis To further explore the connection between the HE-stained patch and the IHC-staining intensity, 39 patch features were defined in Supplementary Table S2. Then, we divided the $p_{(HER2+)}$ into 5 grades: Grade I, $0.33 \le p_{(HER2+)} < 0.4$; Grade II, $0.4 \le p_{(HER2+)} < 0.5$; Grade III, $0.5 \le p_{(HER2+)} < 0.6$; Grade IV, $0.6 \le p_{(HER2+)} < 0.7$; and Grade V,

$0.7 \leq p_{(HER2+)}$. The remarkable features were revealed through the Dunn's test with a P value < 0.05.

Probabilistic aggregation Let $p_{(HER2+)}$ denoted the IHC-staining intensity of a patch, and set $p_{(HER2+)} > 0.33$ as the cut-off value of the IHC-staining area. The score of probabilistic aggregation can be expressed as Eqs. (2–6).

$$Area_{(tumor)} = N_{(HER2+)} + \mathrm{N}_{(HER2-)} \tag{2}$$

$$Area_{(IHC-staining)} = \frac{N_{[p_{(HER2+)}>0.33]}}{Area_{(tumor)}} \tag{3}$$

$$P = \left\{Sort\left(p_{1(HER2+)}, p_{2(HER2+)}, \cdots, p_{n(HER2+)}\right)\right\},$$
$$n = N_{[p_{(HER2+)}>0.33]} \tag{4}$$

$$Mean_{p_{(HER2+)}} = \frac{\sum_{i=1}^{0.3\times Area_{(tumor)}} (P_i - 0.33)}{0.3 \times Area_{(tumor)}} \tag{5}$$

$$Score = \begin{cases} 0, Area_{(IHC-staining)} \leq 0.3, and, Mean_{P_{(HER2+)}} \leq 0.17 \\ Mean_{p_{(HER2+)}}, Area_{(IHC-staining)} > 0.3 \, or \, Mean_{P_{(HER2+)}} > 0.17 \end{cases} \tag{6}$$

where $N_{(*)}$ denoted the number of classes and $Sort(\cdot)$ denoted the descending sorting function. The proposed mapping between this probabilistic aggregation and the clinical IHC reference was described in Supplementary Table S3 [28]. Consequently, the transfer-learning-based probabilistic aggregation (TL-PA), a method belonging to the sub-classifier aggregation framework, was constructed.

Multi-instance learning

Two MIL-based models, namely, clustering-constrained-attention multiple-instance learning (CLAM) using attention-based learning and SlideGraph$^+$ using graph-based learning, were evaluated in this study [14, 16]. The pre-trained backbone models employed the ResNet50 with weights loaded from PyTorch [https://download.pytorch.org/models/resnet50-0676ba61.pth]. The Swin-T with weights were loaded from our transfer learning. For the objective of instance aggregation shifted from universal visual features to task-specific pathology features, models that incorporate transfer learning were considered a variant of MIL. As a result, they were referred to as TL-CLAM and TL-SlideGraph$^+$. The preprocessing steps, including staining normalization and adaptive cropping, were consistent with those in Framework 1.

Model evaluation

Five validation cohorts were employed for model evaluation. The performance of binary classification (HER2+ vs. HER2−/non-pCR vs. pCR) was evaluated by the area under the receiver operating characteristic curve (AUC). The correlation between the scores distribution and HER2 status, as well as RCB grades, was analyzed via the Spearman rank correlation and Dunn's test. $P < 0.05$ was considered to indicate a significant difference.

Results

Patch-level prediction in TL-PA

After 200 epochs of full training and 20 epochs of transfer learning, the accuracy of the training set plateaued at 0.61. The patches from the four validation cohorts inherited their corresponding WSI-level label (HER2 positive or negative). Subsequently, the ROC curve at the patch-level was illustrated in Fig. 3. The AUCs of the four validation cohorts were approximately 0.6. Meanwhile, the accuracy in validation set from cohort A to D was 0.35, 0.44, 0.23, and 0.38, respectively. These results suggested that our training set included a considerable amount of paradoxically labelled data. This demonstrated that not the whole area of the HE-stained WSI required our attention.

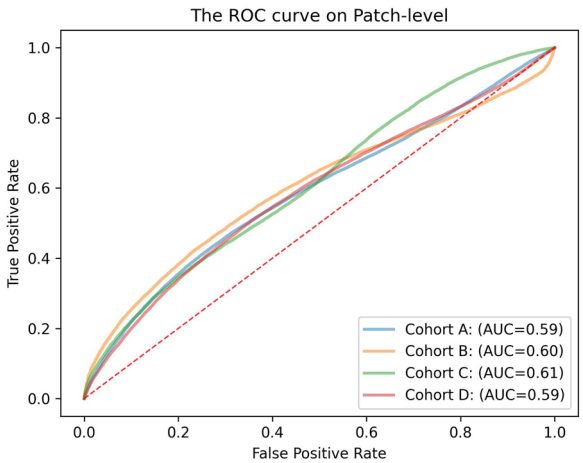

Fig. 3 The ROC curve at the patch-level in the validation sets. The AUC values of Cohort A, Cohort B, Cohort C, and Cohort D were 0.59, 0.60, 0.61, and 0.59, respectively

The patch analysis indicated that higher-grade patches exhibited lower and more dispersed frequency domain energy, along with a reduced cell count, as shown in Fig. S2 (A–C). The Dunn's tests for "Lab-A_dft_1_mean", "Lab-A_dft_1_var", and "cells_num" were described in Fig. S2 (D–F). All *P* values in Fig. S2 (D-F) was found to be less than 0.05, except for the amplitude variance in Lab-A's frequency domain between Grade IV and Grade V.

The feature visualization in Fig. S3 demonstrated a positive correlation between the patch probability and the invasion degree of carcinoma cells. In this context, all adipose tissues, regardless of the presence of cancerous lesions, were categorized as Grade I. Stromal, ductal, and lobular tissues, with no or mild infiltration, were classified as Grades I to III. Grade IV was considered moderate-level stromal infiltration and solid tumors. High-level ductal or lobular infiltration was classified as Grade V.

WSI-level prediction

As shown in Fig. 4, the TL-PA model achieved the highest and most robust AUCs across four validation cohorts, with an AUC of 0.75 in the internal validation set (Cohort A) and 0.76 ± 0.01 in the external validation sets (Cohorts B, C, and D). Under the same training conditions, the MIL-based models exhibited weaker generalization abilities. Specifically, TL-SlideGraph$^+$ achieved an AUC of 0.68 in the internal validation set (Cohort A) and 0.66 ± 0.08 in the external validation sets (Cohorts B, C, and D). TL-CLAM achieved an AUC of 0.57 in the internal validation set and 0.64 ± 0.04 in the external validation sets. The ResNet50-CLAM model, as originally

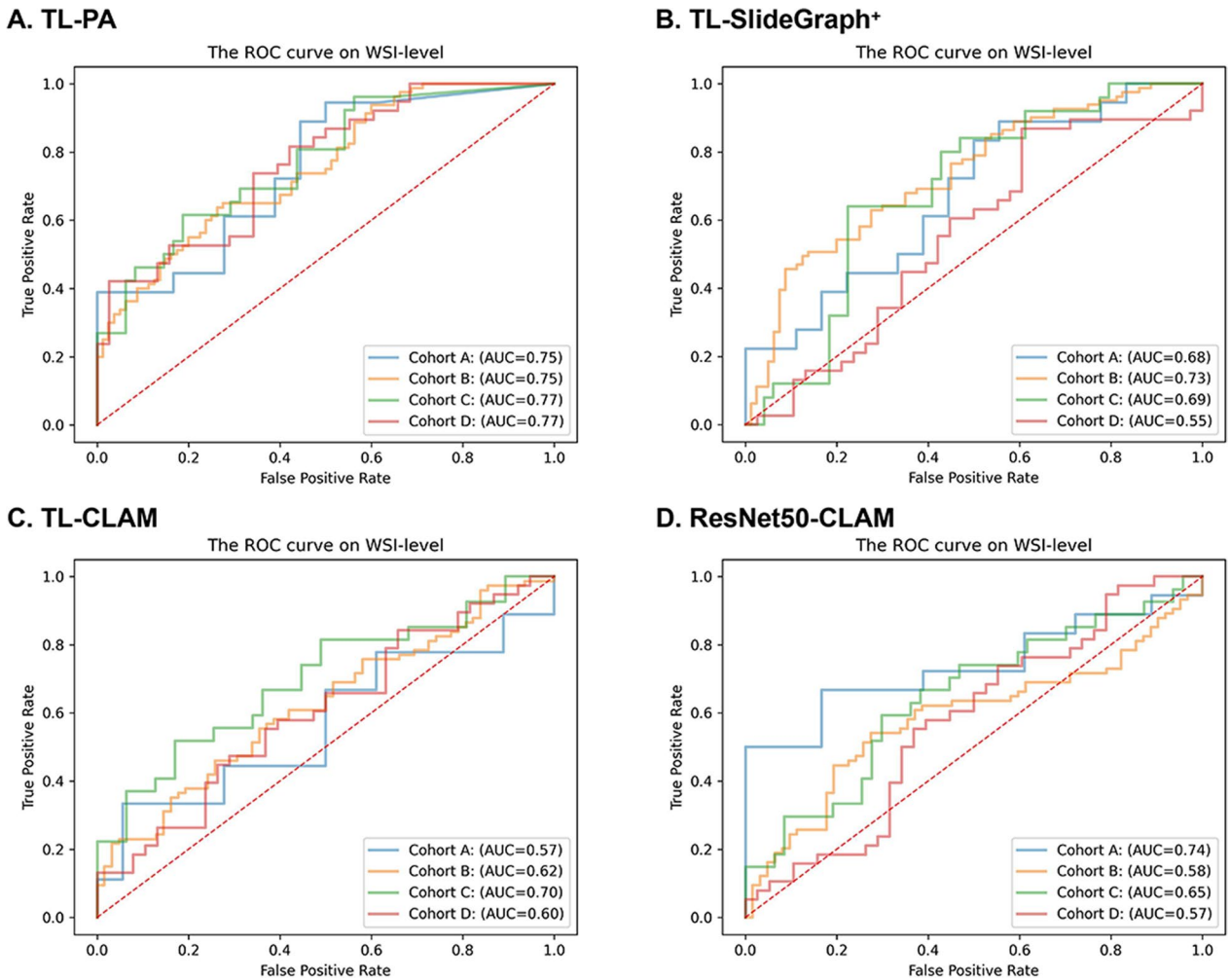

Fig. 4 ROC curves for model evaluation. **A** TL-PA: AUCs in Cohort A to D were 0.75, 0.75, 0.77, and 0.77; **B** TL-SlideGraph$^+$: AUCs in Cohort A to D were 0.68, 0.73, 0.69, and 0.55; **C** TL-CLAM: AUCs in Cohort A to D were 0.57, 0.62, 0.70, and 0.60; **D** ResNet50-CLAM: AUCs in Cohort A to D were 0.74, 0.58, 0.65, and 0.57

implemented, performed well in the internal validation set with an AUC of 0.74, but exhibited the lowest AUC of 0.60 ± 0.04 in the external validation sets.

A more detailed comparison between our work and previous studies was provided in Table 2. This comparison included the required training stages, feature dimensions for instance aggregation, types of instance aggregation, and performance on the shared TCGA-BRCA cohort. The results showed that the sub-classifier framework with rule-based aggregation demonstrated superior performance in HER2 scoring.

Table 2 Comparison of models between our work and previous studies

Method	Training stage		Feature dimension	Aggregation method	Internal validation	External validation
	Feature extraction	Instance aggregation			AUC	AUC (mean ± std)
TL-PA	√	–	N/A	Rule-based	0.75	**0.76 ± 0.01**
TL-SlideGraph$^+$	√	√	768	Graph-based	0.68	0.66 ± 0.08
TL-CLAM	√	√	768	Attention-based	0.57	0.64 ± 0.04
ResNet50-CLAM [16]	–	√	1024	Attention-based	0.74	0.60 ± 0.04
Farahmand et al. [12]	√	–	N/A	Rule-based	**0.81**[a]	N/A
Rawat et al. [11]	√	√	512	MLP-based	0.71[a]	N/A
DAB-SlideGraph$^+$ [14]	√	√	4	Graph-based	0.75 ± 0.02[a]	N/A
Valieris et al. [17]	–	√	1024	Attention-based	0.61 ± 0.01[b]	N/A

std standard deviation, *N/A* Not available, *DAB* 3,3′-Diaminobenzidine, *MLP* Multi-Layer Perceptron, – No additional training required

[a] The best performance on TCGA-BRCA, as reported in the original paper

[b] The performance of M6 (included 5-level HER2 status data) on TCGA-BRCA, as reported in the original paper

The best performance for each method was highlighted in bold

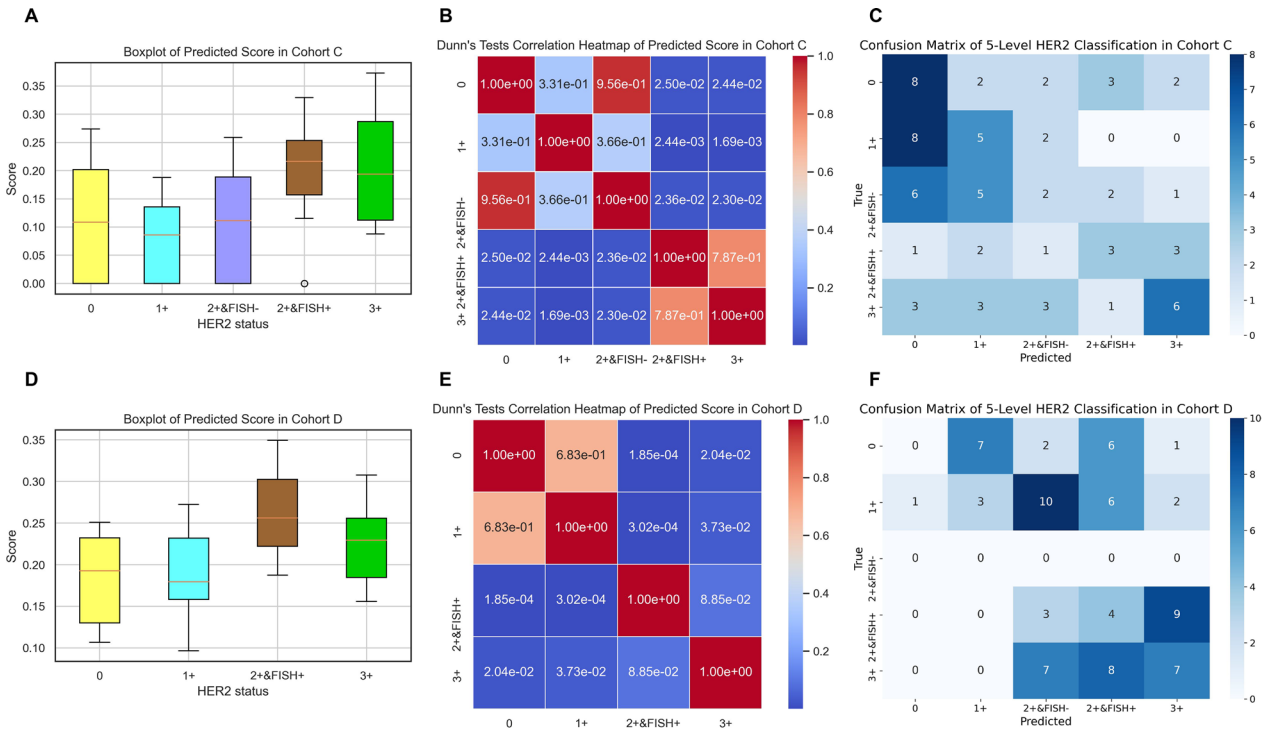

Fig. 5 Correlation analysis of the predicted scores and the true HER2 status in TL-PA. **A**, **D** Boxplots illustrating the predicted scores for Cohorts C and D, featuring Spearman rank correlation coefficients of 0.369 and 0.371, respectively (both with $P_{(Spearman)} = 0.001$). **B**, **E** Dunn's test correlation heatmaps of predicted scores in Cohorts C and D, revealing statistical differences only between HER2-positive and HER2-negative cases ($P < 0.05$). **C**, **F** Confusion matrices of 5-level HER2 classification in Cohorts C and D

Selecting TL-PA as the representative model, we conducted further analysis on the rank correlation between its scores and the 5-level HER2 classification (0, 1+, 2+&FISH−, 2+&FISH+, and 3+), based on data from Cohorts C and D. The correlation analysis revealed that the distribution of predicted scores for intra-class (HER2 positive or negative) did not exhibit a clear monotonic or separable pattern compared to the true HER2 status (Fig. 5A, D). Dunn's tests indicated no significant pairwise differences within the intra-class (Fig. 5B, E). The Spearman rank correlation coefficients for Cohort C and D were 0.371 and 0.369 (both $P = 0.001$), indicating a weak correlation.

Additionally, thresholds set for the 5-level HER2 classification to present the confusion matrix for Cohorts C and D (Fig. 5C, F) were as follow: $0 \leq$ score ≤ 0.1 for HER2 0, $0.1 <$ score ≤ 0.15 for HER2 1+, $0.15 <$ score ≤ 0.2 for HER2 2+&FISH−, $0.2 <$ score ≤ 0.25 for HER2 2+&FISH+, $0.25 <$ score for HER2 3+. A upward shift of approximately 0.5 in predicted scores was observed in Cohort D when compared to Cohort C. TL-PA achieved the optimal AUCs for HER2 binary classification, yet notable misclassification occurred in the more granular classification task.

The WSI-level prediction heatmap for Cohort C was depicted in Fig. 6. In the showcased samples, model scores consistently increased with escalating HER2 status (from 0 to 3+). Our probabilistic aggregation method allowed TL-PA to prioritize the expression intensity of tumor patches over their mere quantity. This was demonstrated in the HER2 2+&FISH+ sample, where a relatively high score was given to fewer highlighted patches. This behavior was consistent across all cohorts, aligning with the intensity-based, limited-scope rule for aggregation. In essence, the model's attention centered on malignant lesions, with a core-to-peritumoral probability decrease pattern as the malignancy spread outwards from the lesion's core.

Correlation with NAT

With TL-PA, the scoring range of Cohort NAT closely aligned with that of the HER2-positive samples (2+&FISH+ and 3+) in Fig. 5A, B, with scores predominantly exceeding 0.15. This alignment underscored a high degree of consistency in their HER2 expression patterns. Additionally, the predicted scores of the pCR group were generally lower than those of the non-pCR group, as illustrated in Fig. 7A. Taking the non-pCR as the positive control, the AUC was 0.77 (Fig. 7B). Given the previously mentioned probability decrease pattern, the peri-tumoral environment may be more closely associated with NAT outcomes, particularly the stromal region. To further verify this hypothesis, we considered the 4-level RCB grades (0, I, II, and III), which provided a more refined assessment of NAT outcomes compared to the pCR. As expected, the correlation between the predicted scores and RCB grades was moderate (Fig. 7C, D, $Coefficient_{(Spearman)} = 0.45$, $P_{(Spearman)} = 0.0006$). Although the score differences among RCBs I, II, and III were not statistically significant, the noticeable overall rising trend pointed to the significance of peri-tumoral

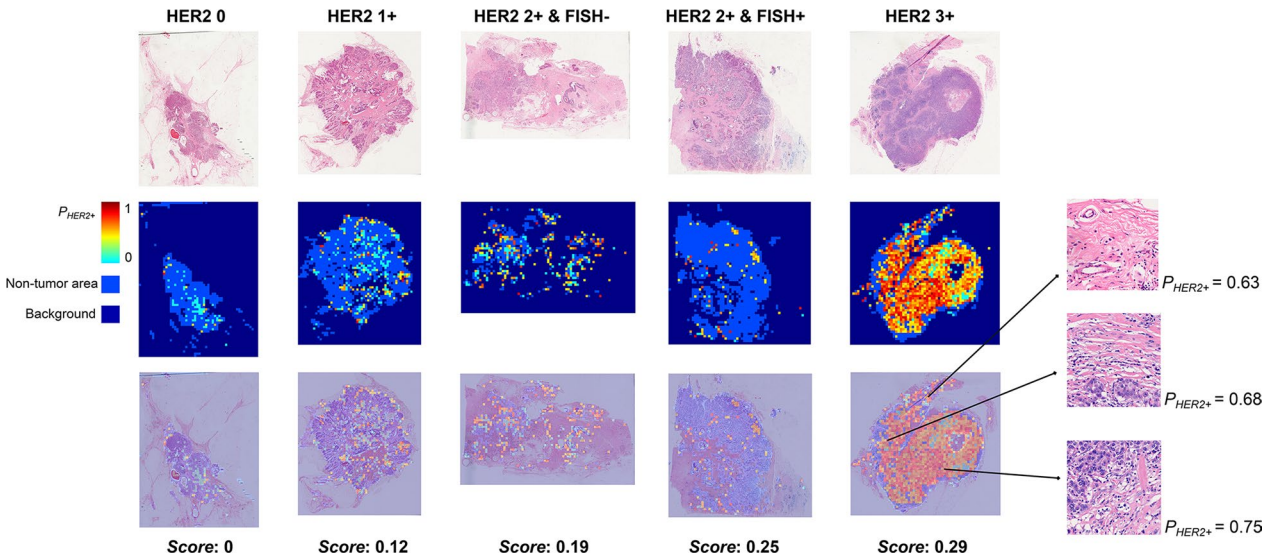

Fig. 6 WSI-level prediction heatmap for Cohort C. From left to right: Clinical HER2 categories 0, 1+, 2+&FISH−, 2+&FISH+ and 3+. Top to bottom: HE-stained WSIs, probability heatmaps, and prediction visualizations. Scores for these 5 samples from TL-PA, 0 for HER2 0, 0.12 for HER2 1+, 0.19 for HER2 2+&FISH−, 0.25 for HER2 2+&FISH+, and 0.29 for HER2 3+

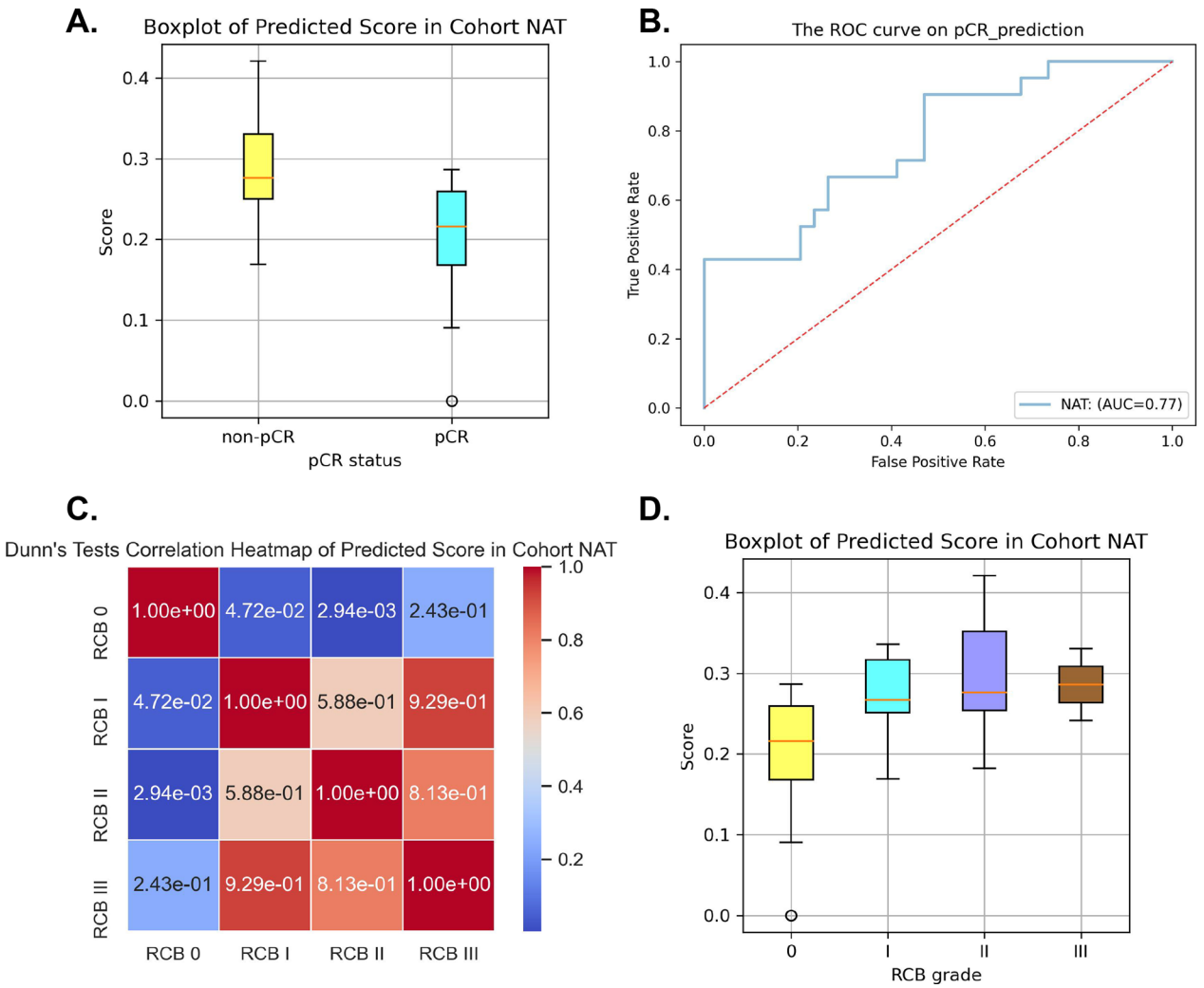

Fig. 7 Correlation analysis of the predicted scores and NAT outcomes. **A** Boxplot of predicted scores in Cohort NAT, $P_{(DunnsTest)} = 0.0009$. **B** The ROC curve for pCR prediction (AUC = 0.77). **C** Dunn's test correlation heatmap of predicted scores in Cohort NAT, indicating statistical differences only between RCB 0 versus I (P=0.0472) and RCB 0 versus II (P=0.0029). **D** Boxplot of predicted scores in Cohort NAT; Compared to RCB grades, $Coefficient_{(Spearman)} = 0.45$, $P_{(Spearman)} = 0.0006$

features in assessing the efficacy of HER2-targeted therapy.

Discussion

In this study, two frameworks for HER2 scoring were evaluated using limited training data and multi-cohort validation. The TL-PA model achieved optimal performance but struggled to capture intra-class variability, particularly in HER2-negative cases. Compared to previous studies, our work highlights the potential of integrating a transfer learning-based feature extractor with rule-based instance aggregation.

The two key components of computational pathology, feature extraction and instance aggregation, were decoupled to facilitate a more effective comparison. Feature extractors in recent studies could be classified into three categories based on their relevance to pathology tasks, including backbones pre-trained on natural image datasets (such as ImageNet [29] or CIFAR-100 [30]), general-purpose models for pathology [31–33], and transfer learning models based on task-specific datasets [12, 14]. The complexity of instance aggregation methods varied depending on the relevance of patch features to the task. Instance aggregation strategies could be classified into rule-based [12], MLP-based [11], attention-based [16, 17], and graph-based [14] approaches.

Lu et al. [14] demonstrated that model performance improved with an increase in the relevance of features

to the task, with DAB-SlideGraph$^+$ achieving the best results. They developed a feature extractor for DAB density estimates by aligning IHC-stained images with HE-stained images. However, our study indicated that, regardless of whether task-specific knowledge was embedded in the front-end feature extractors, MIL was struggle to achieve good HER2 scoring performance with limited data, as demonstrated by comparison between ResNet50-CLAM and TL-CLAM. Our experiments also revealed that transfer learning improved CLAM's generalization on external validation (TL-CLAM), although it might contribute to underperformance on internal validation when compared to ResNet50-CLAM. It suggested a trade-off between model fitting and generalization when embedding prior knowledge. A similar issue was observed in other CLAM variants. In Valieris et al.'s study [17], the model, based on non-task-specific pathological features extracted from a general-purpose model, was overfitted during external validation. The introduction of transfer learning-based feature extractors held promise in addressing this challenge.

Patch features with strong task relevance could be paired with simpler instance aggregation methods, as demonstrated in Farahmand et al.'s work [12] and our TL-PA, which outperformed MIL-based methods. Farahmand et al. initially developed a sub-classifier based on Inception-v3, utilizing transfer learning to predict HER2 status, and subsequently averaged the probabilities of all patches to aggregate a WSI-level score. However, this global aggregation struggled to capture intensity-dependent patterns that align with the reasoning of pathologists in HER2 scoring. TL-PA integrated an intensity-based limited-scope aggregation method to enhance interpretability and generalization. This rule-based aggregation method also reduced inference costs compared to MIL, as it avoided complex network computations. With the same input data and feature extractor, the inference speed during the instance aggregation stage could improve from minute-level to second-level timescales. A patch analysis of TL-PA indicated that malignant proliferation near ductal or lobular tissue was closely associated with HER2 expression. These areas were often the origin of breast cancer [34]. In the patch-level view, expansive malignant cells proliferated outward in a low-density cluster configuration, resulting in low cell counts and a relatively smooth, non-nuclear texture. These features became the primary characteristics of patches with strong HER2 expression.

However, TL-PA could not fully capture the ordered pattern in the 5-level HER2 status. Intra-class differences were observed primarily within the HER2 positive group, where the predicted score for HER2 3+was generally lower than that for HER2 2+&FISH+. Researches indicated that HER2 3+breast cancer had a greater chance of pCR after receiving anti-HER2 therapy [35, 36]. Similarly, we observed a statistical difference in the model scores between the pCR and non-pCR groups. RCB also exhibited an overall upward trend in scores as the prognosis worsened. All these results implied that higher scores could serve as a significant indicator for poorer NAT. Considering the observed pattern of decreasing probability from the core to the peritumoral region in this study, the significant increase in stromal infiltration may be associated with a stronger immune response. As malignant lesions progress within the stroma, their susceptibility to immune cells and drugs may be simultaneously enhanced [37].

There were several limitations in our work. First, numerous noisy patches during transfer learning significantly might affect the performance of the TL-PA. The tumor-based binary annotation strategy primarily directed the sub-classifier's focus to clustered lesion cores. This tendency resulted in underestimated scores in HER2 3+ cases, particularly those characterized by diffuse stromal invasion. Also it lacked distinct tumor cores and exhibited a more uniform expression intensity across the tissue. Meanwhile, some patches from the positively stained regions of the HER2-negative samples were overlooked due to annotation limitations, making the identification of their highly heterogeneous internal grading features even more challenging. To reduce the blindness driven by noisy data, integrating HER2-IHC images to guide the model's attention presents a promising solution. Additionally, TL-PA exhibited baseline drift during external validation. This drift could make defining standard boundaries for a 5-level classification challenging, a difficulty that may not be reflected by the ROC curve. Federated transfer learning based on foundation models maybe a a potential solution [38].

Conclusions

In this study, a model for HER2 scoring on HE-stained WSIs has established and successfully validated across multiple external cohorts. The feasibility of integrating a transfer learning-based feature extractor with rule-based instance aggregation is demonstrated. However, the model struggles to capture the 5-level ordered patterns consistent with pathologist reasoning. Future work should incorporate more granular annotations to strengthen the training constraints of deep learning models, thereby improving the reliability of automated quantification.

Abbreviations

HER2	Human epidermal growth factor receptor 2
HE	Hematoxylin&Eosin
WSIs	Whole slide images

NAT	Neoadjuvant therapy
AUC	Area under the receiver operating characteristic curve
IHC	Immunohistochemistry
FISH	Fluorescence in situ hybridization
ROI	Region of interest
MIL	Multi-instance learning
TCGA-BRCA	The Cancer Genome Atlas-Breast invasive carcinoma
ASCO/CAP	American Society of Clinical Oncology/College of American Pathologists
pCR	Pathological complete response
RCB	Residual cancer burden
Swin-T	Swin-transformer-tiny
CLAM	Clustering-constrained-attention multiple-instance learning
GNN	Graph neural network
std	Standard deviation
N/A	Not available
DAB	3,3′-Diaminobenzidine
MLP	Multi-layer perceptron

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13058-025-02008-7.

Additional file 1.

Acknowledgements
The authors acknowledge patients, clinicians, and hospital staff who participated in this study, including the staff at Jiangmen Central Hospital, Jiangmen Maternity & Child Health Care Hospital, Hangzhou Dianzi University, and Wuyi University.

Author contributions
Conceptualization: XL, ZL; methodology: ZL, CQ, YZ, CL; data collection: XL, WZ, HH; formal analysis: XL, SS, TQ; resources: XL, TQ; visualization: ZL, WZ; funding acquisition: XL, HH, SS; writing—original draft: all authors; writing—review and editing: all authors.

Funding
This work was supported by Wu Jieping Medical Fund (Funding No. 320.6750.2022-20-3), Jiangmen Science and Technology Planning Project (Funding No. JZ202219), Jiangmen Science and Technology Planning Project (Funding No. 2022YL01011), Guangdong Medical Science and Technology Research Fund Project (Funding No. B2023436) and National Natural Science Foundation of China (No. 82372143).

Availability of data and materials
All private data can be made available upon reasonable request and with the permission of the corresponding author, after an appropriate data access agreement that specifies the terms and conditions of data usage. Cohort A and B were retrieved online (https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=119702524). Cohort NAT: https://tinyurl.com/IMPRESS-DATA.

Declarations

Ethics approval and consent to participate
Approval for this study was obtained from the Ethics Committees of both Jiangmen Central Hospital and Jiangmen Maternity & Child Health Care Hospital.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Breast Department, Jiangmen Central Hospital, Jiangmen, China. [2]Hangzhou Dianzi University, Hangzhou, China. [3]Shanxi Key Lab for Modernization of TCVM, College of Life Science, Shanxi Agricultural University, Taiyuan 030000, Shanxi, China. [4]Department of Pathology, Jiangmen Central Hospital, Jiangmen, China. [5]Radiology Department, Jiangmen Central Hospital, Jiangmen, China. [6]Department of Breast Surgery, Jiangmen Maternity and Child Health Care Hospital, Jiangmen 529000, Guangdong, China. [7]Wuyi University, 99 Yinbin Avenue, Jiangmen 529000, Guangdong, China.

Received: 7 March 2024 Accepted: 24 March 2025
Published online: 23 April 2025

References

1. Wolff AC, Hammond MEH, Hicks DG, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. Arch Pathol Lab Med. 2014;138(2):241–56.
2. Katayama A, Miligy IM, Shiino S, et al. Predictors of pathological complete response to neoadjuvant treatment and changes to post-neoadjuvant HER2 status in HER2-positive invasive breast cancer. Mod Pathol. 2021;34(7):1271–81.
3. Goutsouliak K, Veeraraghavan J, Sethunath V, et al. Towards personalized treatment for early stage HER2-positive breast cancer. Nat Rev Clin Oncol. 2020;17(4):233–50.
4. Wolff AC, Somerfield MR, Dowsett M, et al. Human epidermal growth factor receptor 2 testing in breast cancer: ASCO–College of American Pathologists Guideline Update. J Clin Oncol. 2023;41(22):3867–72.
5. Baez-Navarro X, van Bockstal MR, Nawawi D, et al. Interobserver variation in the assessment of immunohistochemistry expression levels in HER2-negative breast cancer: can we improve the identification of low levels of HER2 expression by adjusting the criteria? An international interobserver study. Mod Pathol. 2023;36(1): 100009.
6. Nielsen K, Sode M, Jensen MB, et al. High inter-laboratory variability in the assessment of HER2-low breast cancer: a national registry study on 50,714 Danish patients. Breast Cancer Res. 2023;25(1):139.
7. Ibrahim A, Gamble P, Jaroensri R, et al. Artificial intelligence in digital breast pathology: techniques and applications. The Breast. 2020;49:267–73.
8. Qaiser T, Rajpoot NM. Learning where to see: a novel attention model for automated immunohistochemical scoring. IEEE Trans Med Imaging. 2019;38(11):2620–31.
9. Alizadeh E, Castle J, Quirk A, et al. Cellular morphological features are predictive markers of cancer cell state. Comput Biol Med. 2020;126: 104044.
10. Palla G, Fischer DS, Regev A, et al. Spatial components of molecular tissue biology. Nat Biotechnol. 2022;40(3):308–18.
11. Rawat RR, Ortega I, Roy P, et al. Deep learned tissue "fingerprints" classify breast cancers by ER/PR/Her2 status from H&E images. Sci Rep. 2020;10(1):7275.
12. Farahmand S, Fernandez AI, Ahmed FS, et al. Deep learning trained on hematoxylin and eosin tumor region of Interest predicts HER2 status and trastuzumab treatment response in HER2+ breast cancer. Mod Pathol. 2022;35(1):44–51.
13. Naik N, Madani A, Esteva A, et al. Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. Nat Commun. 2020;11(1):5727.
14. Lu W, Toss M, Dawood M, et al. SlideGraph+: whole slide image level graphs to predict HER2 status in breast cancer. Med Image Anal. 2022;80: 102486.
15. Kulkarni PM, Robinson EJ, Sarin Pradhan J, et al. Deep learning based on standard H&E images of primary melanoma tumors identifies patients at risk for visceral recurrence and death. Clin Cancer Res. 2020;26(5):1126–34.
16. Lu MY, Williamson DFK, Chen TY, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 2021;5(6):555–70.
17. Valieris R, Martins L, Defelicibus A, et al. Weakly-supervised deep learning models enable HER2-low prediction from H & E stained slides. Breast Cancer Res. 2024;26(1):124.
18. Morales S, Engan K, Naranjo V. Artificial intelligence in computational pathology—challenges and future directions. Digit Signal Process. 2021;119: 103196.
19. Song AH, Jaume G, Williamson DFK, et al. Artificial intelligence for digital and computational pathology. Nat Rev Bioeng. 2023;1(12):930–49.

20. Modi S, Jacot W, Yamashita T, et al. Trastuzumab deruxtecan in previously treated HER2-low advanced breast cancer. N Engl J Med. 2022;387(1):9–20.
21. Huang Z, Shao W, Han Z, et al. Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images. NPJ Precis Oncol. 2023;7(1):14.
22. Zhuang F, Qi Z, Duan K, et al. A comprehensive survey on transfer learning. Proc IEEE. 2020;109(1):43–76.
23. Liang G, Zheng L. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. Comput Methods Programs Biomed. 2020;187: 104964.
24. Kim D, Wang K, Sclaroff S, et al. A broad study of pre-training for domain generalization and adaptation. In: European conference on computer vision. Cham: Springer, 2022: 621–638.
25. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012–10022.
26. Vahadane A, Peng T, Sethi A, et al. Structure-preserving color normalization and sparse stain separation for histological images. IEEE Trans Med Imaging. 2016;35(8):1962–71.
27. Goode A, Gilbert B, Harkes J, et al. OpenSlide: a vendor-neutral software foundation for digital pathology. J Pathol Inform. 2013;4(1):27.
28. Wulfing P, Borchard J, Buerger H, et al. HER2-positive circulating tumor cells indicate poor clinical outcome in stage I to III breast cancer patients. Clin Cancer Res. 2006;12(6):1715–20.
29. Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009: 248–255.
30. Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. 2009.
31. Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. Nat Med. 2024;30(3):850–62.
32. Juyal D, Padigela H, Shah C, et al. PLUTO: pathology-universal transformer. arxiv preprint arXiv:2405.07905, 2024.
33. Xu H, Usuyama N, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. Nature, 2024: 1–8.
34. Inoue M, Nakagomi H, Nakada H, et al. Specific sites of metastases in invasive lobular carcinoma: a retrospective cohort study of metastatic breast cancer. Breast Cancer. 2017;24:667–72.
35. Solinas C, Ceppi M, Lambertini M, et al. Tumor-infiltrating lymphocytes in patients with HER2-positive breast cancer treated with neoadjuvant chemotherapy plus trastuzumab, lapatinib or their combination: a meta-analysis of randomized controlled trials. Cancer Treat Rev. 2017;57:8–15.
36. Lien HC, Lo C, Lee YH, et al. In situ HER2 RNA expression as a predictor of pathologic complete response of HER2-positive breast cancer patients receiving neoadjuvant chemotherapy and anti-HER2 targeted treatment. Breast Cancer Res. 2024;26(1):100.
37. Ingold Heppner B, Untch M, Denkert C, et al. Tumor-infiltrating lymphocytes: a predictive and prognostic biomarker in neoadjuvant-treated HER2-positive breast cancer. Clin Cancer Res. 2016;22(23):5747–54.
38. Li B, Liu Z, Shao L, et al. Point transformer with federated learning for predicting breast cancer HER2 status from hematoxylin and eosin-stained whole slide images. In: Proceedings of the AAAI conference on artificial intelligence. 2024, 38(4): 3000–3008.

Publisher's Note