Intelligent Oncology 1 (2025) 128–138

Contents lists available at ScienceDirect

Intelligent Oncology

journal homepage: www.keaipublishing.com/en/intonc/

Full Length Article

AWCDL: Automatic weight calibration deep learning for detecting HER2 status in whole-slide breast cancer image

Shuai Wang[a,b,#], Xiaojing Guo[c,#], Jinlu Ma[d,#], Ning Wu[e], Xiaolong Qian[c], Guogui Sun[f], Jia Guo[g], Xiangfei Meng[g,*], Bo Kang[g,*], Bo Xu[b,h]

[a] *Department of Radiation Oncology, Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan Shandong 250117, China*
[b] *Department of Biochemistry and Molecular Biology, Key Laboratory of Breast Cancer Prevention and Therapy, Tianjin Clinical Research Center for Cancer, Ministry of Education, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute & Hospital, Tianjin 300060, China*
[c] *Department of Breast Cancer Pathology and Research Laboratory, Tianjin Medical University Cancer Institute & Hospital, Tianjin 300060, China*
[d] *Department of Radiation Oncology, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an Shaanxi 710061, China*
[e] *State Key Laboratory of Female Fertility Promotion, Center for Reproductive Medicine, Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing 100191, China*
[f] *Department of Hebei Key Laboratory of Medical–Industrial Integration Precision Medicine, Affiliated Hospital, North China University of Science and Technology, Tangshan Hebei 063000, China*
[g] *Department of Research and Development, National Supercomputer Center in Tianjin, Tianjin 300000, China*
[h] *Chongqing Key Laboratory of Intelligent Oncology for Breast Cancer, Chongqing University Cancer Hospital, Chongqing University School of Medicine, Chongqing 400030, China*

ARTICLE INFO

Keywords:
Automatic weight calibration
Artificial intelligence
Breast cancer
HER2 expression
Whole-slide image

ABSTRACT

Defining an *ERBB2* (*HER2/neu*) gene amplification status is critical to guiding human epidermal growth factor receptor 2 (HER2)-targeted therapy in breast cancer. Up to 40% of breast cancer patients are reported as having an immunohistochemistry (IHC) of HER2 2+ and requiring additional testing using fluorescence *in situ* hybridization to confirm the results. This paper aims to establish an automatically weighted calibration deep learning (AWCDL) algorithm to predict *ERBB2* amplification based on IHC images. In this study, we applied IHC HER2 2+ images from 1,073 breast cancer patients at three cancer centers in China and extracted 376,099 tiles. Among these, 269,664 tiles were used for internal and external validation. The designed AWCDL consists of two steps. In Step 1, the internal validation achieved an accuracy of 89%, with a specificity of 0.89 and a sensitivity of 0.89. The external validation in the two other centers showed an average accuracy of 85%, with a specificity of 0.86 and a sensitivity of 0.82. In Step 2, the model achieved higher accuracy for the slides predicted as negative in Step 1 by automatically calibrating the weight. Collectively, these results suggest that this AWCDL model has successfully proved useful as an alternative method to fluorescence *in situ* hybridization for assessing the *ERBB2* amplification status in breast cancer.

* Corresponding authors.
E-mail addresses: mengxf@nscc-tj.cn (X. Meng), kangbo@nscc-tj.cn (B. Kang), xubo@tmu.edu.cn (B. Xu).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.intonc.2025.03.008
Received 31 January 2025; Received in revised form 20 March 2025; Accepted 20 March 2025
Available online 27 March 2025
2950-2616/© 2025 The Author(s). Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Breast cancer is a common malignant tumor among women. A statistical study showed that breast cancer became the leading diagnosed cancer worldwide in 2020.[1] Breast cancer is divided into four inherent molecular subtypes based on the expression of hormone receptors and human epidermal growth factor receptor 2 (HER2) expression: Luminal A, Luminal B, HER2-enriched, and triple-negative breast cancer.[2] The HER2-positive subtype accounts for 15%–20% of breast cancer cases, and its prognosis was poor before anti-HER2 targeted therapy was developed.[3,4] Therefore, accurately identifying the amplification status of the *ERBB2* gene, which encodes HER2, is critical for the clinical management of patients with breast cancer.[5] Immunohistochemistry (IHC) has been used to detect HER2 expression in pathology.[6] According to the guidelines, HER2 expression is defined as negative if the IHC results are zero or 1+ and positive when the result is 3+ .[7] However, when the IHC result is 2+ (weak to moderate complete membrane staining observed in over 10% of tumor cells), an additional fluorescence *in situ* hybridization (FISH) test is recommended to confirm *ERBB2* gene amplification.[8] In general, HER2 IHC is highly subjective, and as many as 40% of these patients could be reported as 2+ .[9] Consequently, a great number of patients require a FISH test. However, the FISH test, itself, requires experienced pathologists and specific techniques that may not be available in many hospitals.[10] Additionally, it imposes an extra financial burden on patients, as the cost may not be covered by health insurance.[11]

Given these challenges, there is a growing need to explore alternative methods for HER2 detection that can overcome the limitations of IHC and FISH. Traditionally, HER2 expression detection in breast cancer has been categorized into positive (IHC 3+/FISH amplification) and negative (IHC 2+/FISH-, IHC 1+, IHC 0) groups.[12] However, recent studies have highlighted the importance of HER2-low and HER2-zero subtypes, which have distinct clinicopathological characteristics and therapeutic responses.[13] The emergence of new anti-HER2 targeted therapies, such as trastuzumab deruxtecan, has shown efficacy in HER2-low breast cancer patients, emphasizing the need for more accurate HER2 detection methods. Traditional methods, such as IHC and FISH, have limited sensitivity and specificity, particularly in the lower range of HER2 expression. Recent advancements in machine learning have enabled the label-free detection of cellular HER2 with high accuracy, providing a promising technique for dynamic therapeutic surveillance.[14] Despite these advancements, existing digital image analysis (DIA) methods for HER2 detection face several limitations. A systematic review by Dunenova et al. highlighted that, while DIA algorithms show high accuracy in detecting HER2 status, there is a notable absence of direct evidence supporting their clinical application.[15] The review emphasized the need for standardized evaluation criteria and clinical validation to enhance the reliability and applicability of these algorithms in real-world clinical settings. Bannier et al. developed a deep-learning model for HER2 detection, achieving high accuracy in identifying HER2-low and HER2-positive tumors. However, the study noted that morphological variables and specific artifacts, such as extensive fibrosis and nonspecific background staining, can cause discrepancies between pathologist and model assessments.[16] These challenges underscore the importance of developing robust and clinically validated tools for HER2 detection.

In this context, the integration of artificial intelligence (AI) and digital pathology has emerged as a promising solution to enhance the accuracy and efficiency of HER2 detection. AI, particularly deep learning algorithms, is gaining extensive attention for its unparalleled excellent performance in digital image information capture.[17] The integration of deep learning and digital pathology has improved diagnostic accuracy and increased work efficiency.[18] In the CAMELYON16 study, AI was used to help pathologists screen for lesions in histopathological slides.[19] Machine learning can also improve microscopy techniques, such as increasing the accuracy of diagnosing prostate cancer and identifying metastatic breast cancer.[20,21] Regarding the same subject matter, features in different models might have connections not visually recognizable by the human eye.[22] The traditional methods that use only color normalization might not always yield reliable results because they overlook other features.[23,24]

Building on these advancements, we developed the automatically weighted calibration deep learning (AWCDL) algorithm to address the specific challenges of HER2 detection in IHC images. The AWCDL algorithm extracts IHC features from different centers and connects them to FISH results, enabling the definition of *ERBB2* gene amplification without a FISH test on digital whole-slide IHC images.

2. Methods

2.1. Datasets

We retrospectively collected whole-slide images (WSI) of HER2 IHC images from 1,073 patients in three hospitals with confirmed FISH results. The hospitals providing the images were Tianjin Medical University Cancer Institute & Hospital (TMUCIH), Xi'an Jiaotong University First Affiliated Hospital (XJTUFAH), and Tangshan People's Hospital (TPH). The enrolled IHC images were confirmed by three pathologists before being sent for AI processing. The HER2 expression was based on "Recommendations for HER2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update."[25] This study complies with the Institutional Review Board of each participating institute (No.: bc2018006). This study was exempt from the requirement to provide informed consent to the Institutional Review Board because of its retrospective nature.

2.2. Training cohort

Between January 1, 2016, and December 31, 2017, HER2 IHC 2+ images from 789 cases were judged as qualified for training and internal validation at the Department of Breast Pathology at TMUCIH, the main research center of this study. Among them, 676 slides with a positive-to-negative ratio of 1:1 were used for training purposes. The slides were delineated using a rectangle frame as regions

of interest (ROIs) and then segmented into tiles with a pixel resolution of 512 × 512 for training (Figure 1). The ROIs were delineated based on the FISH testing regions and representative staining regions. The mean number of ROIs on each slide was four (Figure S1).

2.3. Validation cohort

The validation cohorts consisted of 369 patients from either TMUCIH (internal validation cohort, 113 patients), XJTUFAH (XJTUFAH external validation cohort, 142 patients), or TPH (TPH external validation cohort, 114 patients). In addition, we tested 28 patients from TMUCIH who were diagnosed with IHC 1+ but also had FISH results to observe whether the algorithm could detect false negatives (FNs) of IHC.

2.4. Sample preparation

The thicknesses of all IHC slides were 3 μm, and the slides were stained based on the processing protocols of each hospital. The staining protocols varied slightly but were stained with an IHC Immuno–HistoStainer (Leica, Germany). The slides were then scanned using two high-resolution scanners (×40): Nanozoomer 2.0 (Hamamatsu Photonics, Japan) and Leica Microsystems SCN400 (Leica Microsystems Wetzlar GmbH, Germany). The ROI images were reconfirmed in terms of the 2+ score by three pathologists at TMUCIH, each with ≥10 years of specialized experience in breast cancer pathology.

2.5. Tile selection

We segmented the original ROI pictures into tiles with a pixel resolution of 512 × 512. Representative tiles were predicted to be stained and non-stained regions, the last being excluded for further analysis. The screening process of a qualified tile used the following steps: (1) The staining areas around the nucleus in the original ROI images were extracted by color gamut filtering. (2) The image was binarized to distinguish the stained area from other areas. (3) The pixel proportion “D” of the staining areas in the binary images was calculated corresponding to the original ROI images. (4) The pixel proportion of the stained area of each tile was labeled “d,” and the ROI-stained area as “D.” When d < D× 0.8, it is considered a background interference image and will be removed; otherwise, it is considered a qualified tile.

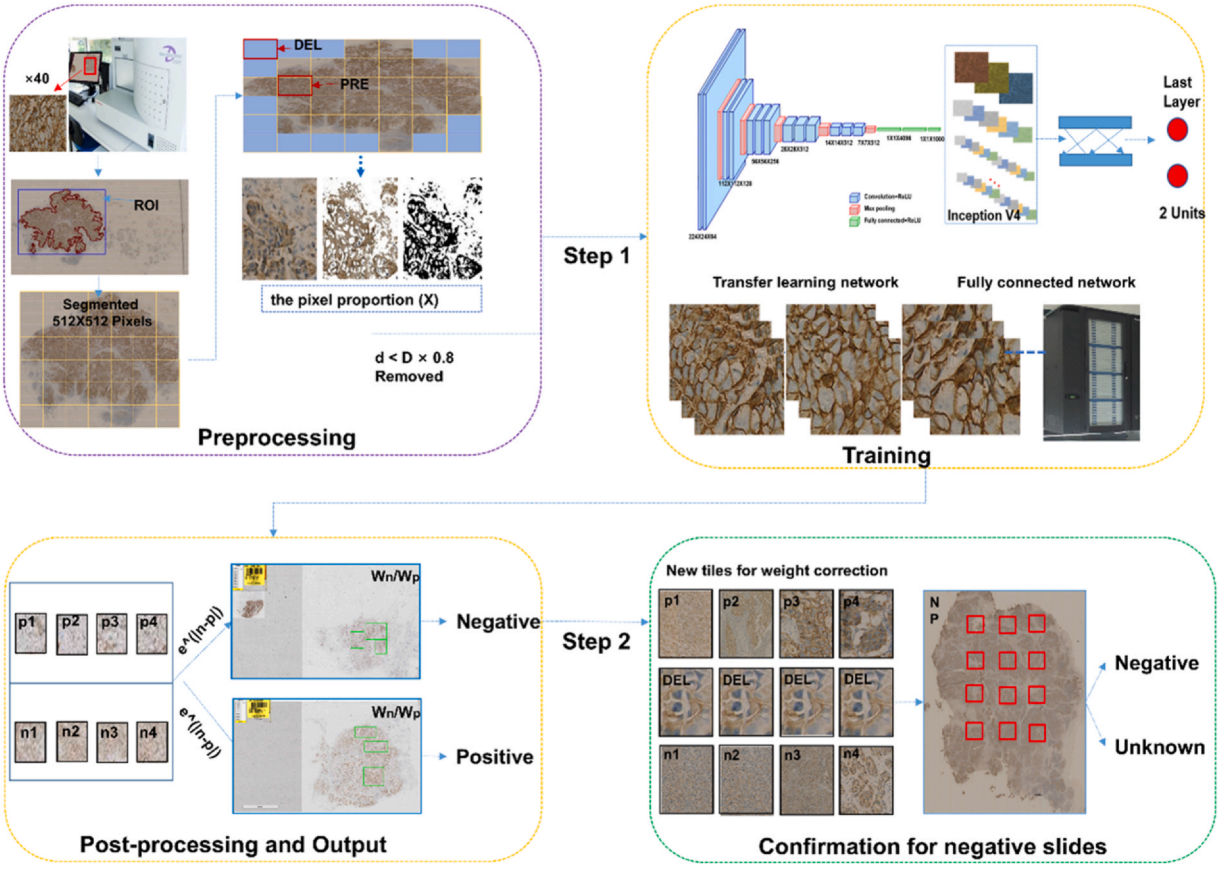


Figure 1. Pipeline of the deep convolutional neural network classification. The WSI was stored in a TIFF format and input into the computer.

2.6. Model establishment and classification

A transfer learning neural network based on the Inception V4 architecture was used to train 106,435 tiles (36,846 positive and 69,589 negative). The Inception V4 backbone was pretrained on the ImageNet dataset using standard cross-entropy loss, enabling the extraction of generic visual features. Subsequently, the network was fine-tuned end-to-end on our HER2-specific dataset without freezing the pretrained encoder weights, allowing adaptation to the unique characteristics of IHC images (e.g., membrane staining patterns). The model was iterated for 40,000 epochs, with an initial learning rate of 0.01 for the pretrained layers. The neural network architecture comprised two parts:

- Feature Extraction: The pretrained Inception V4 converted image data into one-dimensional feature vectors.
- Classification Head: A fully connected layer performed binary classification (HER2-positive vs. HER2-negative).

This end-to-end training strategy ensured that both the pretrained encoder and the custom classification head were jointly optimized on the HER2 task.

2.6.1. Automatic weight calibration

Considering the imbalance between the positive and negative samples, we used a weighted cross-entropy loss function. The weight of the loss function for each label was set inversely proportional to the number of samples. Specifically, the weight for the positive class (which had fewer samples) was set higher than that for the negative class. This approach helped the model focus more on positive samples, improving its ability to detect HER2-positive cases. The training loss curve was smooth after 4,000 epochs (Figure S2).

2.6.2. Prediction and weight update

To obtain a prediction score at the individual level, we denoted "X" as the number of tiles per patient. The probability that X tiles would be predicted as negative and positive is "n" and "p," respectively. When $n > p$, the tile was predicted to be negative; otherwise, it was predicted to be positive. The slides of the patients were then classified into two sets of "N" and "P".

We set the weight of each tile as "$w = e\hat{}(|n\text{-}p|)$." The probability of this patient being predicted as negative and positive was then divided into "w_n" and "w_p," respectively, which was the sum of the weights of the tiles in the corresponding negative and positive sets. When $w_n > w_p$, the WSI was predicted as negative; otherwise, it was predicted as positive.

To decrease the FN rate and strengthen the generalization ability of the model, the code was post-processed for Step 2 to yield a better screening tool for negative slides. Negative slides predicted by Step 1 were applied to Step 2 of the model. To promote the generalization of Step 2, at least 100 ROIs were input for weight calibration, and the new weight parameters were generated. To ensure independence between calibration and test sets, the 100 ROIs used for weight calibration were randomly selected from distinct patients. The remaining ROIs from these patients and all ROIs from other patients were allocated to the test set, ensuring no overlap of patient-level data between the calibration and testing phases. The specific procedure is as follows:

(1) **Screening**: Tiles with a negative prediction and a negative gold standard were screened.
(2) **Quartile Calculation**: The quartile of "n" was calculated as $(1/4\,n)$, and three-quarters of "p" was calculated as $(3/4\,p)$.
(3) **Prediction Value Update**: If $n > (1/4\,n)$, $p < (3/4\,p)$, $(n\text{-}p) > 1/4(n\text{-}p)$, and $n > p$, the tile was predicted to be negative. If $n < (3/4\,n)$, $p > 1/4\,p$, $(n\text{-}p) < 3/4(n\text{-}p)$, and $p > n$, the tile was predicted to be positive.
(4) **Counting**: The number of ROIs, the corresponding number of predicted negative tiles (N) and positive tiles (P) on each slide were counted. If N > P, the prediction was negative.
(5) **Final Decision:** Tiles with a negative prediction and a positive gold standard were screened, and three-quarters of N/(N + P) was then calculated as "A." If N/(N + P) > A, the slide was predicted to be negative.

2.6.3. Example to explain the automatic weight calibration process

Consider a dataset for HER2 detection in breast cancercomprising 100 samples, where 10 samples are HER2-positive (positive class) and 90 samples are HER2-negative (negative class). The goal is to train a model to classify these samples accurately.

Weight initialization: The initial weight for each sample is set to 1, meaning all samples have equal importance.

Weighted cross-entropy loss calculation: To address the 1:9 class imbalance (10 positives vs. 90 negatives), we calculated inverse-frequency weights for the loss function: $w_p = 100/10 = 10$, $w_n = 100/90 = 1.11$. These weights ensure the model pays 9 × more attention to the minority (positive) class during training.

Prediction process: For each image tile, the model outputed two probabilities: p: Probability of being HER2-positive; n: Probability of being HER2-negative (where p + n =1). Tiles were initially classified as positive if p > n, otherwise negative.

Tile-level Weight Assignment: Each tile (a smaller image patch from the whole-slide image) is assigned a weight based on the uncertainty of its prediction. The weight is calculated as w= e^(|n-p|), where n is the probability of being negative and p is the probability of being positive. For a tile with n =0.6 and p=0.4, the weight w = e^(|0.6–0.4|) = e^0.2 ≈ 1.2214. For a tile with n = 0.8 and p = 0.2, the weight w = e^(|0.8–0.2|) = e^0.6 ≈ 1.8221. Tiles with higher uncertainty (closer to 0.5) received higher weights.

Prediction Aggregation: The model aggregates the predictions from all tiles in a patient's WSI to determine the final classification. If the sum of weights for negative tiles exceeds the sum of weights for positive tiles, the WSI is classified as negative.

Post-processing Weight Calibration: After the initial prediction, the model applies a post-processing step to recalibrate the weights and reduce the FN rate. For negative slides predicted by Step 1, the model screened tiles with negative predictions and positive ground truth. It recalculated the weights using quartile thresholds. For example, if the first quartile of negative probabilities is 0.3 and the third quartile of positive probabilities is 0.7, any tile with $n > 0.3$, $p < 0.7$, $(n - p) > 0.3 - 0.7$, and $n > p$ is reclassified as negative.

2.7. Visualization of the model

The learning process and explanation of the convolutional neural network model are not intuitive to users. To facilitate users' understanding of the prediction process of the model, we visualized convolutional layers from two perspectives: 1) weight calibration to be identified as the ROI based on the "Tile selection" step and 2) class activation mapping (CAM) techniques to capture potential tumor areas. We used pixel proportion to identify the ROI and reserve the tiles to meet the weights. CAM is a weighted linear sum of the presence of these visual patterns at different spatial locations. By simply upsampling the CAM to the size of the input image, we identified the suspected tumor area most relevant to the particular category. For a breast cancer IHC HER2 slide, a suspicious tumor area on the tiles drew the attention of the model and varied in different tumors.

2.8. Manual ROI selection in validation and future clinical applications

To ensure accurate predictions during the validation phase, manual ROI selection was performed by three experienced pathologists (each with ≥ 10 years of specialized experience in breast cancer pathology) following the American Society of Clinical Oncology/College of American Pathologists guidelines. The ROIs were delineated based on the FISH testing regions and representative HER2 IHC 2+ staining areas. In future clinical applications, manual ROI selection will initially be required to calibrate the model for specific hospital staining protocols and scanner equipment. However, we are actively developing an automated tumor segmentation algorithm based on U-Net architecture to eliminate this dependency. Preliminary experiments have shown promising results, achieving > 90% accuracy in semi-automated ROI localization using CAMs and tile-level filtering (Section 2.5). For the initial deployment phase, users will need to manually select ROIs for a small set of cases (e.g., 100 slides) to calibrate the model's weight parameters. Once calibrated, the model can process new cases with minimal manual intervention, as the system automatically filters out background regions ($d < D \times 0.8$) and highlights suspicious tumor areas using CAMs. This approach significantly reduces the manual effort required for ROI selection while maintaining high prediction accuracy.

2.9. Statistical analysis

To test the categorization of the model, we evaluated the receiver operating characteristic (ROC) curve and the area under the curve (AUC). In addition, to evaluate the generalization of the model, we calculated the accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), Youden Index (YI), F1 score, and Kappa value. This formula was based on the confusion matrix recording the number of true positives (TPs), true negatives (TNs), false positives (FPs), and FNs. The accuracy was calculated as the fraction of patients who were correctly identified. The specificity was calculated as the fraction of FISH-negative patients who were correctly identified, and the sensitivity as the fraction of FISH-positive patients who were correctly identified. The PPV was the TP fraction of patients among the predicted positive patients, and the NPV was the TN fraction of patients among the predicted negative patients. The YI was applied as the determining exponent of the optimal bound. The F1 score was a measure of the accuracy of the binary model. Moreover, the performance was evaluated using the F1 score to compare the similarity and diversity of the performance. The agreement between the DL model's prediction and the clinical report was measured using the Kappa value. A simple random sampling was conducted among the three centers of patients, and a one-way ANOVA was used to analyze the different groups.

3. Results

3.1. IHC datasets

To establish the correlation between the IHC and FISH results, we enrolled 1,073 cases from three cohorts for training and multicenter validation. Among these, 676 slides (339 positive and 337 negative) were obtained from TMUCIH for training purposes. A total of 113 slides (57 positive and 56 negative) from TMUCIH, 114 slides (76 positive and 38 negative) from the TPH cohort, and 142 slides (101 positive and 41 negative) from the XJTUFAH cohort were used for validation. A total of 269,664 (19,228 + 153,527 + 92,000 + 4,909) tiles were used for validation. Among them, 19,228 tiles (5,180 positive and 14,048 negative) from TMUCIH were adopted for internal validation and 153,527 tiles (29,302 positive and 124,225 negative) from TPH and 92,000 tiles (14,547 positive and 77,453 negative) from XJTUFAH for external validation. Additionally, 4,909 tiles (2,517 positive and 2,392 negative) reported as IHC 1+ from TMUCIH were employed to test whether the model could detect FN IHC (Figure 2).

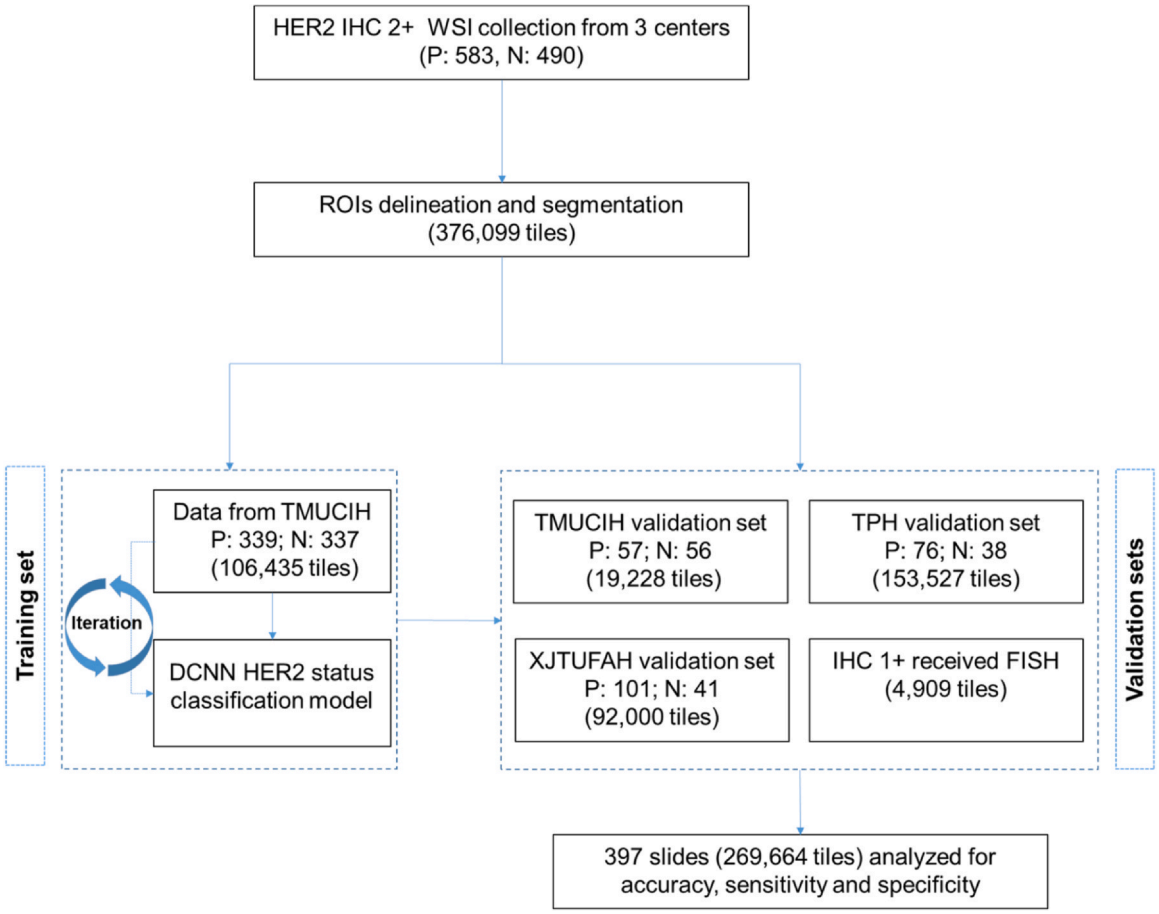


Figure 2. Flow chart of the subjects in the construction and validation of the DCNN model for automated *ERBB2* amplification diagnosis. The controls were FISH-negative HER2 patients based on HER2 IHC 2+ slides. Abbreviations: 1. DCNN: Deep convolutional neural network 2. N: Negative 3. P: Positive 4. ROI: Region of interest 5. WSI: Whole-slide image.

3.2. Performance of AWCDL Step 1

The AWCDL Step 1 achieved an AUC of 0.89 (95% CI = 0.885–0.895) on the internal validation. The optimal cutoff for the prediction of HER2 expression (positive or negative) was determined using ROC analysis by selecting the cutoff with the highest YI (Figure 3A). A sensitivity of 0.89, specificity of 0.89, PPV of 0.89, NPV of 0.89, and accuracy of 89% were found (Table 1).

To evaluate the generalization of AWCDL, we assessed its performance in two external cohorts. The external validation at the two centers showed an average accuracy of 84.5% with a specificity of 0.86 and a sensitivity of 0.82, indicating the consistency of the model. More specifically, the AWCDL algorithm yielded an AUC of 0.85 (95% CI = 0.847–0.852) for the TPH (Figure 3B) and 0.79 (95 % CI = 0.786–0.798) for the XJTUFAH (Figure 3C) cohorts. For the TPH and XJTUFAH cohorts, the accuracy was 92% and 77%, the sensitivity was 0.90 and 0.74, the specificity was 0.93 and 0.78, the NPVs were 0.95 and 0.89, and the F1 scores were 0.88 and 0.65, respectively (Table 1). The kappa values were 0.79, 0.82, and 0.49 for the internal and external validation.

3.3. Performance of AWCDL Step 2

To reduce the potential FN rate resulting from Step 1 of the AWCDL, we optimized the model by automatically calibrating the weight parameter automatically to screen for negative slides after the model was run through this step. We conducted simple random sampling three times in the TMUCIH slides and five times in the TPH and XJTUFAH slides. For each center, the calibration set comprised at least 100 ROIs from independent patients, while the test set included ROIs from the remaining patients and unused ROIs from the calibration patients. This separation guarantees patient-level data independence, mitigating potential data leakage. The prediction accuracy of the negative slides reached 94.4%, 97.3%, and 91.86% at the three centers, indicating a significant reduction in the FN rate. A statistical analysis showed no significant difference in the specificity in the three centers (Figure 4), indicating that the model possesses excellent learning and generalization capabilities.

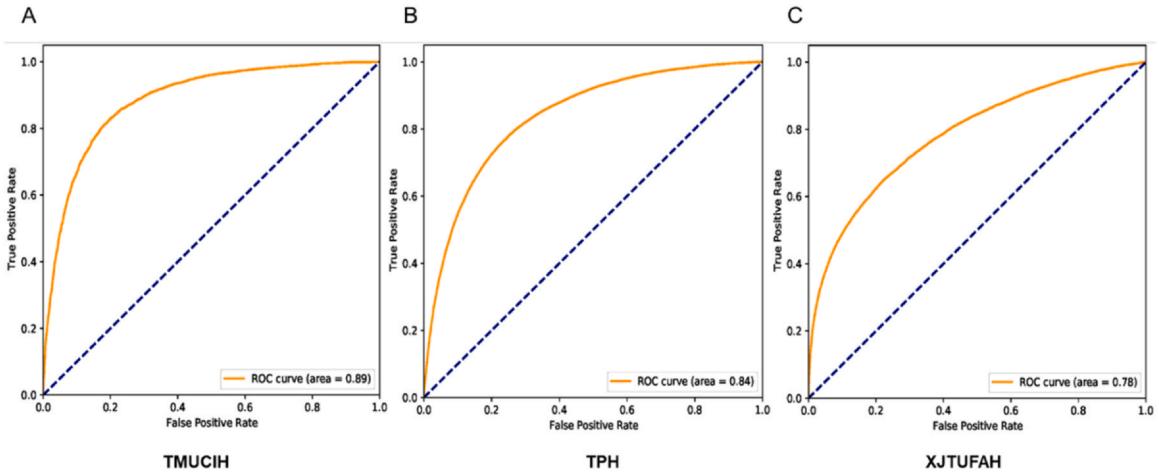


Figure 3. Performance of the ensemble automatically weighted calibration deep learning model in identifying *ERBB2* amplification on IHC 2+ slides in three validation sets. A. ROC of TMUCIH; B. ROC of TPH; C. ROC of XJTUFAH. ROC: Receiver operating characteristic.

Table 1
Performance metrics for the AWCDL model of Step 1.

Performance metric	Internal TMUCIH	External TPH	External XJTUFAH
AUC (95% CI)	0.89 (0.885–0.895)	0.85 (0.847–0.852)	0.79 (0.786–0.798)
Accuracy (%)	89	92	77
Sensitivity	0.89	0.90	0.74
Specificity	0.89	0.93	0.78
PPV	0.89	0.87	0.56
NPV	0.89	0.95	0.89
F1 score	0.89	0.88	0.65
Kappa value	0.79	0.82	0.49
Yoden index	0.79	0.83	0.53

Abbreviations: 1. AUC: Area under the curve 2. PPV: Positive predictive value 3. NPV: Negative predictive value.

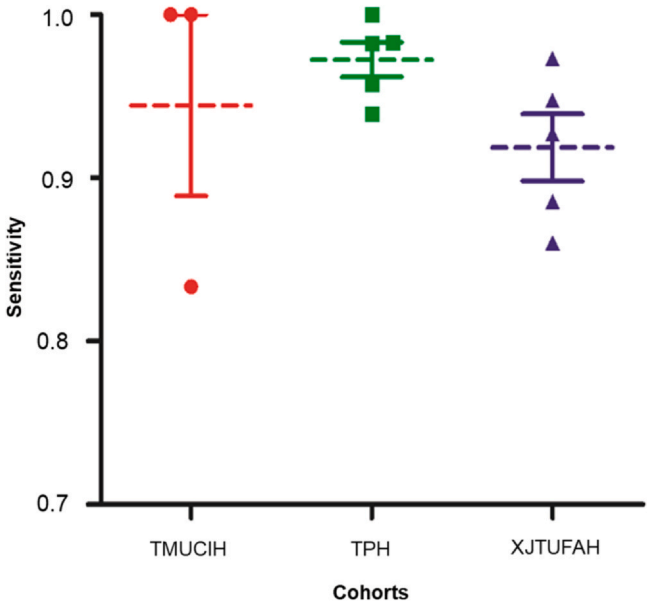


Figure 4. Statistical analysis of prediction for negative slides forecast using Step 1 of the model from three centers. Random sampling was conducted three times in TMUCIH, five times in TPH, and five times in XJTUFAH. The average specificity of the three centers was 0.94, 0.97, and 0.92, respectively. There was no significant difference among the centers.

3.4. Detection of the HER2 amplification status of IHC 1+ slides

The goal of IHC is to detect HER2-positive patients. However, there is a marked FN rate when using IHC. We aimed to develop a model that can significantly reduce the FP rate. To test whether the model could detect FN IHC results, we enrolled an additional 28 patients reported as IHC 1+ from TMUCIH. Although reported as IHC1+ , these patients volunteered to undergo FISH testing. The FISH results showed that 10 out of the 28 had *ERBB2* amplification, suggesting that IHC testing has a FP rate of 0.35. Using the AWCDL model, we achieved an overall accuracy of 82%. More importantly, the model successfully detected five FISH-positive slides, reducing the IHC FP rate by 50%. These results demonstrate the predictive value of the model for IHC 1+ cases regarding the *ERBB2* amplification status.

3.5. Deep learning model visualization and feature analysis

By learning from 106,435 tumor tiles, the model detected features that demonstrated a strong association with *ERBB2* amplification status. We visualized several convolutional filters in the model (Figure 5). The shallow convolutional layer learned low-level

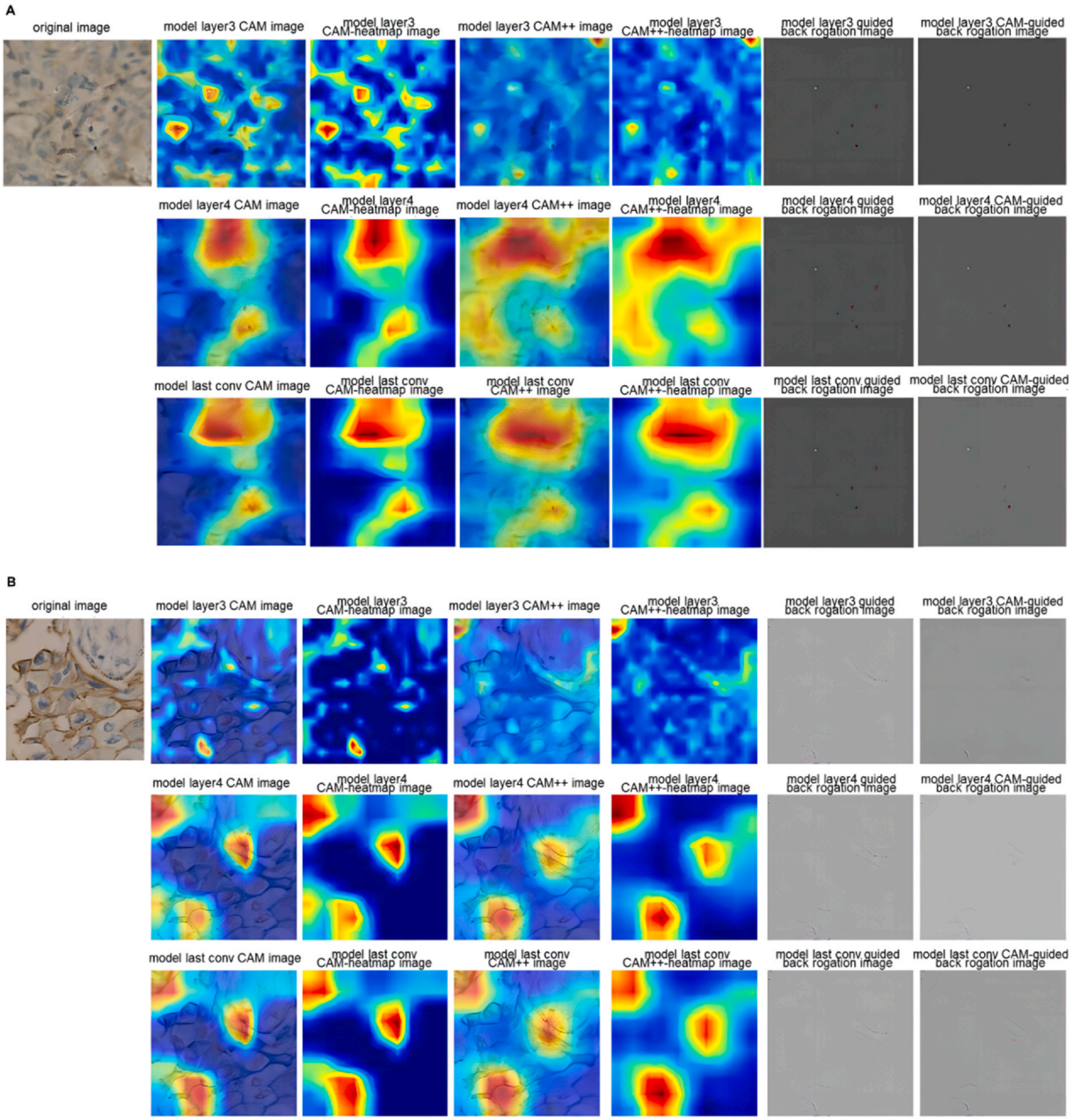


Figure 5. Deep learning model visualization and feature analysis. Figure 5A and 5B show the feature map visualization results for HER2-negative and HER2-positive tiles, respectively.

simple features, while a deeper convolutional layer learned more complex features. The first row is the weighted feature map graph of convolution layer 3 (shallow convolutional layer) learned classification of tumor and stroma region. Transfer learning makes the filters more abstract and specific to HER2 expression when the network layers go deeper. For instance, the second row represents the weighted feature map graph of the global average pooling layer, and the last row corresponds to that of the SoftMax layer, both of which are deeper convolutional layers that learn stricter classification regions. These convolutional layers captured different staining patterns responding to HER2-positive and HER2-negative tiles. The convolutional feature visualization schematic of the deep learning model is shown in Figure 5. For a HER2 IHC 2+ slide, the model generated fusion visualization of the weighted summation heatmap. The feature map by the model varied in different HER2 expression patterns (Figure 5). For example, the hot spot region in Figure 5A focused on lymph cells, the luminal area, and the deeply stained nucleus, ignoring the membrane staining region and dirty spot, and predicted it to be HER2 negative. In contrast, the hot spot area in Figure 5B is the deeply stained membrane. Based on these observations, the model interpreted these two tiles as different HER2 expression patterns. This means that the model tends to focus on the hot spot area on the luminal, stroma area, and lymphocytes of the HER2-negative tiles, while focusing on the deeply stained area of the nucleus or membrane of the HER2-positive tiles. In this way, the *ERBB2* amplification status forms a strong correlation with the area that the model targets. Therefore, it can potentially provide a hotspot region for pathologists to identify ROIs that merit focus.

4. Discussion

ERBB2 amplification is frequently found in breast cancer patients.[26] According to TMUCIH statistics, the proportion diagnosed with IHC HER2 2+ has reached 38%. Among the IHC 2+ patients who had received a FISH test, 32% showed *ERBB2* amplification (FISH positive; Figure S3). Correctly identifying these patients is critical for guiding anti-HER2-targeted therapy. FISH is the gold standard for diagnosing *ERBB2* statuses; however, it is technically challenging, labor-intensive, and expensive.[27,28] In the Step 1 prediction, internal validation achieved a total accuracy of 89% with a specificity of 0.89 and a sensitivity of 0.89. The external validation in two other centers showed an average accuracy of 84.5%, with a specificity of 0.86 and a sensitivity of 0.82. In Step 2, the model automatically calibrates weights and identifies slides predicted to be negative in Step 1, achieving an average accuracy of 94.4%, 97.3%, and 91.9% at the three centers, respectively. The successful establishment of this quick model indicates that AI can serve as an alternative tool to FISH to improve the efficiency and accuracy of traditional methods for assessing breast cancer HER2 status.

The AWCDL model was the first AI approach to use deep learning algorithms to assess the correlation of IHC HER2 2+ with the FISH results. It was validated by samples from three centers. When we separated the data from the two external centers, the accuracy of the TPH cohort was similar to the internal validation of 92%, while the accuracy of the XJTUFAH cohort was slightly lower, at 77%. As the ultimate goal for the AWCDL model is to accurately identify HER2-positive patients, reducing the FN rate is critical for the application of the model, while we encountered a relatively high FN rate (26%) in the XJTUFAH external cohort. The poor results of the XJTUFAH center may be due to the inconsistency of IHC caused by the different IHC staining procedures performed, which include tissue preparation, antibody selection, staining techniques, and the experience of the pathologist. In addition, different scanner equipment might also result in differences in digital histological images. Our calibration strategy prioritizes reducing FNs over FPs due to the following critical considerations: In HER2 testing, FNs have severe clinical implications. A missed HER2-positive diagnosis would deprive patients of anti-HER2-targeted therapies, which are essential for improving survival outcomes in HER2-positive breast cancer. Conversely, FPs typically trigger confirmatory FISH testing, aligning with current clinical guidelines. While FPs may incur additional costs, they do not directly compromise patient care, as FISH serves as a reliable diagnostic safeguard. This rationale is reflected in our model design, which prioritizes sensitivity to ensure no true HER2-positive cases are overlooked. While FP reduction remains important, our current focus on FN mitigation addresses the urgent clinical need to minimize undertreatment risks.

In addition, color normalization is an important parameter for an AI model in digital pathology image recognition.[29] We also initially tested the color normalization of IHC images using a cycle-consistent generative adversarial network to convert the original color into a relatively uniform stain style. However, the prediction and generalization capabilities of the model did not achieve a reliable result. These failed attempts indicate that color normalization might not be a vital factor for digital pathology image recognition, especially for large multitudes of slides from different centers. Instead, automatically calibrating each tile's weight parameter after prediction in Step 1 to generate the final prediction in Step 2 on the negative slides seemed more reasonable. We used random sampling at least three times from three centers to avoid bias in Step 2. The results showed that the average accuracy rates of the TMUCIH, TPH, and XJTUFAH centers reached as high as 94.4%, 97.3%, and 91.9%, respectively. This additional step reduced the FN rate by 5% in the internal validation and 4% and 14% in the two external centers. For future applications of the model at any other center, only 100 ROIs from the center will be enough for calibrating the weight. Hence, the model possesses excellent learning and generalization capabilities.

The application of this AWCDL model can benefit many breast patients. As stated before, FISH technology requires technical excellence and experienced pathologists.[30,31] Even in more developed areas such as China's capital city of Beijing, 8.9% of patients have no access to HER2 testing, and almost 10% of patients with uncertain IHC results failed to have a FISH for reasons including a lack of machines and experienced pathologists. Moreover, findings from a nationwide survey regarding the quality of HER2 testing by IHC in 45 Chinese hospitals suggested that only 28 (62%) met acceptable reporting standards.[19,32] Another reason many patients might refuse to take a FISH test is that the test is not covered by national health insurance in China. At TMUCIH, one of the largest breast cancer centers in Asia, 38% of patients have been reported as being diagnosed as IHC 2+. Among them, 12% of the patients did not receive an FISH test for economic or other reasons. Therefore, many patients diagnosed with IHC 2+ will benefit from using the deep continuous convolutional networks (DCCN) model.

Another application of the model is that it can screen IHC 1+ patients and reduce the FP rate. In our cases of 28 IHC HER2, the three pathologists could not reach a consensus and finally made an ambiguous diagnosis of IHC 1. The patients who volunteered received FISH, and 10 out of 28 were finally confirmed as HER2+. Then, we used the DCCN model to predict these slides. The results showed that the model can reduce the FP rate by 50%, reducing it from 35% to 17.5%. This suggests that AWCDL provides broader use for the screening of HER2 IHC images.

In addition, we developed the model as open access, and when more data were collected, the model could be continuously optimized (available at: https://ai.nscc-tj.cn/thai/deploy/public/her2). With the accumulation of the number of samples, we expect the accuracy of the model's negative identification will continue to increase. Optimistically, with a gradual increase in the number of images, the accuracy of the model in determining negative cases can reach close to 100%. Correspondingly, if the prediction result of the model shows that *ERBB2* amplification is negative, a FISH test is no longer required.

5. Conclusions

In this study, we developed the AWCDL model, a two-stage deep learning framework, to predict *ERBB2* amplification status directly from HER2 IHC 2+ WSIs, offering a cost-effective alternative to FISH testing. The model demonstrated robust performance across three independent medical centers, achieving post-calibration accuracies of 94.4%, 97.3%, and 91.9%, with significant reductions in false-negative rates through dynamic weight optimization. By integrating automated tumor segmentation and interpretable CAM, our approach not only enhances diagnostic accuracy but also provides visual guidance for pathologists to identify critical regions of interest. These advancements highlight the model's potential to reduce reliance on FISH in resource-limited settings while maintaining clinical reliability. Future work will extend the model's utility to HER2-low and HER2-zero subtypes, addressing evolving therapeutic needs and further refining its adaptability across diverse clinical workflows.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No.: 81672743 and 81974464).

Ethical statement

The study design was approved by the ethics review committee at Tianjin Medical University Cancer Institute & Hospital.

CRediT authorship contribution statement

Guo Jia: Methodology. **Qian Xiaolong:** Resources. **Sun Guogui:** Resources. **Meng Xiangfei:** Software, Methodology. **Kang Bo:** Supervision, Software. **Xu Bo:** Writing – review & editing, Project administration. **Wang Shuai:** Writing – original draft, Validation, Resources, Project administration, Investigation, Conceptualization. **Guo Xiaojing:** Software, Resources. **Ma Jinlu:** Resources. **Wu Ning:** Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank all the members of Bo Xu's Laboratory for their technical support and discussion.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.intonc.2025.03.008.

References

1. Trapani D, Ginsburg O, Fadelu T, et al. Global challenges and policy solutions in breast cancer control. *Cancer Treat Rev.* 2022;104:102339.
2. Gao ZJ, Fang H, Sun S, et al. Single-cell analyses reveal evolution mimicry during the specification of breast cancer subtype. *Theranostics.* 2024;14(8):3104–3126.
3. Tapia M, Hernando C, Martinez MT, et al. Clinical impact of new treatment strategies for HER2-positive metastatic breast cancer patients with resistance to classical anti-HER therapies. *Cancers (Basel).* 2023;15(18):4522.
4. Favero A, Segatto I, Capuano A, et al. Loss of the extracellular matrix glycoprotein EMILIN1 accelerates Delta16HER2-driven breast cancer initiation in mice. *NPJ Breast Cancer.* 2024;10(1):5.
5. Ivanova M, Porta FM, D'Ercole M, et al. Standardized pathology report for HER2 testing in compliance with 2023 ASCO/CAP updates and 2023 ESMO consensus statements on HER2-low breast cancer. *Virchows Arch.* 2024;484(1):3–14.
6. Falato C, Schettini F, Pascual T, et al. Clinical implications of the intrinsic molecular subtypes in hormone receptor-positive and HER2-negative metastatic breast cancer. *Cancer Treat Rev.* 2023;112:102496.
7. Denkert C, Seither F, Schneeweiss A, et al. Clinical and molecular characteristics of HER2-low-positive breast cancer: pooled analysis of individual patient data from four prospective, neoadjuvant clinical trials. *Lancet Oncol.* 2021;22(8):1151–1161.

8. Hu Y, Jones D, Zhao W, et al. Incidence, clinicopathologic features, HER2 fluorescence in situ hybridization profile, and oncotype DX results of human epidermal growth factor receptor 2-low breast cancers: experience from a single academic center. *Mod Pathol.* 2023;36(7):100164.
9. Farshid G, Armes J, Dessauvagie B, et al. Development and validation of a HER2-low focused immunohistochemical scoring system with high-interobserver concordance: the Australian HER2-low breast cancer concordance study. *Mod Pathol.* 2024;37(8):100535.
10. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA.* 2017;318(22):2199–2210.
11. Yang X, Jia Y, Xu J, et al. Navigating the implementation gap for Trastuzumab's journey from health insurance to patient access: a preliminary study in a hospital in China. *Glob Health Res Policy.* 2024;9(1):50.
12. Moutafi M, Robbins CJ, Yaghoobi V, et al. Quantitative measurement of HER2 expression to subclassify ERBB2 unamplified breast cancer. *Lab Invest.* 2022;102(10):1101–1108.
13. Jiang M, Liu J, Li Q, et al. The trichotomy of HER2 expression confers new insights into the understanding and managing for breast cancer stratified by HER2 status. *Int J Cancer.* 2023;153(7):1324–1336.
14. Xie Y, Wen Y, Su X, et al. Label-free plasmon-enhanced spectroscopic HER2 detection for dynamic therapeutic surveillance of breast cancer. *Anal Chem.* 2022;94(37):12762–12771.
15. Dunenova G, Kalmataeva Z, Kaidarova D, et al. The performance and clinical applicability of HER2 digital image analysis in breast cancer: a systematic review. *Cancers (Basel).* 2024;16(15).
16. Bannier PA, Broeckx G, Herpin L, et al. Development of a deep-learning model tailored for HER2 detection in breast cancer to aid pathologists in interpreting HER2-low cases. *Histopathology.* 2024;85(3):478–488.
17. Luo J, Fu J, Lu Z, et al. Deep learning in integrating spatial transcriptomics with other modalities. *Brief Bioinform.* 2024;26(1):bbae719.
18. Xu H, Wang M, Shi D, et al. When multiple instance learning meets foundation models: advancing histological whole slide image analysis. *Med Image Anal.* 2025;101:103456.
19. Fan L, Strasser-Weippl K, Li JJ, et al. Breast cancer in China. *Lancet Oncol.* 2014;15(7):e279–289.
20. Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. *Nat Rev Urol.* 2019;16(7):391–403.
21. Zhao J, Sun Z, Yu Y, et al. Radiomic and clinical data integration using machine learning predict the efficacy of anti-PD-1 antibodies-based combinational treatment in advanced breast cancer: a multicentered study. *J Immunother Cancer.* 2023;11(5):e006514.
22. Roseboom W, Fountas Z, Nikiforou K, et al. Activity in perceptual classification networks as a basis for human subjective time perception. *Nat Commun.* 2019;10(1):267.
23. Castaldi PJ, Boueiz A, Yun J, et al. Machine learning characterization of COPD subtypes: insights from the COPD gene study. *Chest.* 2020;157(5):1147–1157.
24. Blasi T, Hennig H, Summers HD, et al. Label-free cell cycle analysis for high-throughput imaging flow cytometry. *Nat Commun.* 2016;7:10256.
25. Wolff AC, Hammond ME, 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. *J Clin Oncol.* 2013;31(31):3997–4013.
26. Gao Y, Wu N, Wang S, et al. Concurrent mutations associated with trastuzumab-resistance revealed by single cell sequencing. *Breast Cancer Res Treat.* 2021;187(3):613–624.
27. Zhong J, Gao B, Wang Q, et al. Exploring the heterogeneity of HER2 gene status and expression in non-positive breast cancer patients: insights from immunohistochemistry and fluorescence in situ hybridization. *Diagn Pathol.* 2025;20(1):4.
28. Beeghly-Fadiel A, Kataoka N, Shu XO, et al. Her-2/neu amplification and breast cancer survival: results from the Shanghai breast cancer study. *Oncol Rep.* 2008;19(5):1347–1354.
29. Zheng Y, Jiang Z, Zhang H, et al. Adaptive color deconvolution for histological WSI normalization. *Comput Methods Prog Biomed.* 2019;170:107–120.
30. Turashvili G, Leung S, Turbin D, et al. Inter-observer reproducibility of HER2 immunohistochemical assessment and concordance with fluorescent in situ hybridization (FISH): pathologist assessment compared to quantitative image analysis. *BMC Cancer.* 2009;9:165.
31. Yao Q, Hou W, Wu K, et al. Using whole slide gray value map to predict HER2 expression and FISH status in breast cancer. *Cancers (Basel).* 2022;14(24):6233.
32. Yuan XM, Wang N, Ouyang T, et al. Current status of diagnosis and treatment of primary breast cancer in Beijing, 2008. *Chin J Cancer Res.* 2011;23(1):38–42.