Breast Cancer Research

RESEARCH

Open Access

Machine learning prediction of HER2-low expression in breast cancers based on hematoxylin–eosin-stained slides

Jun Du[1,2†], Jun Shi[3†], Dongdong Sun[4†], Yifei Wang[4], Guanfeng Liu[5], Jingru Chen[5], Wei Wang[1,2], Wenchao Zhou[2,7*], Yushan Zheng[6*] and Haibo Wu[1,2,7*]

Abstract

Background Treatment with HER2-targeted therapies is recommended for HER2-positive breast cancer patients with *HER2* gene amplification or protein overexpression. Interestingly, recent clinical trials of novel HER2-targeted therapies demonstrated promising efficacy in HER2-low breast cancers, raising the prospect of including a HER2-low category (immunohistochemistry, IHC) score of 1 + or 2 + with non-amplified in-situ hybridization for HER2-targeted treatments, which necessitated the accurate detection and evaluation of HER2 expression in tumors. Traditionally, HER2 protein levels are routinely assessed by IHC in clinical practice, which not only requires significant time consumption and financial investment but is also technically challenging for many basic hospitals in developing countries. Therefore, directly predicting HER2 expression by hematoxylin-eosin (HE) staining should be of significant clinical values, and machine learning may be a potent technology to achieve this goal.

Methods In this study, we developed an artificial intelligence (AI) classification model using whole slide image of HE-stained slides to automatically assess HER2 status.

Results A publicly available TCGA-BRCA dataset and an in-house USTC-BC dataset were applied to evaluate our AI model and the state-of-the-art method SlideGraph + in terms of accuracy (ACC), the area under the receiver operating characteristic curve (AUC), and F1 score. Overall, our AI model achieved the superior performance in HER2 scoring in both datasets with AUC of 0.795 ± 0.028 and 0.688 ± 0.008 on the USCT-BC and TCGA-BRCA datasets, respectively. In addition, we visualized the results generated from our AI model by attention heatmaps, which proved that our AI model had strong interpretability.

Conclusion Our AI model is able to directly predict HER2 expression through HE images with strong interpretability, and has a better ACC particularly in HER2-low breast cancers, which provides a method for AI evaluation of HER2 status and helps to perform HER2 evaluation economically and efficiently. It has the potential to assist pathologists to improve diagnosis and assess biomarkers for companion diagnostics.

Keywords Breast cancer, Machine learning, Prediction, Hematoxylin–eosin, HER2

† Jun Du, Jun Shi and Dongdong Sun contributed equally to this work.

*Correspondence:
Wenchao Zhou
WZAZ@ustc.edu.cn
Yushan Zheng
yszheng@buaa.edu.cn
Haibo Wu
wuhaibo@ustc.edu.cn
Full list of author information is available at the end of the article

Introduction

Breast cancer (BC) is the most commonly diagnosed cancer in women worldwide and the leading cause of cancer deaths globally in the year of 2022 [1]. In China, BC became the most frequent cancer type in women, and Chinese patients accounted for approximately 18% death cases of BC in the world [2]. In the foreseeable future, BC will remain a critical health challenge and a main risk factor for death [3].

Human epidermal growth factor receptor 2 (HER2) is a well-known negative prognostic factor, a predictive biomarker, and a therapeutic target in several kinds of cancers, leading to the development of multiple targeted therapies utilizing the monoclonal antibody trastuzumab as well as other anti-HER2 compounds [4, 5]. The HER2 IHC scoring is a semi-quantitative method that applies four grades, HER2 0, 1+, 2+, and 3+, to denote the expression levels of HER2 protein on cell surface in BC tissues according to the American Society of Clinical Oncology and the American Society of Pathologists (ASCO/CAP) 2018 guidelines. HER2-positive BC is defined as IHC HER2 3+or IHC HER2 2+in combination with *HER2* gene amplification detected by ISH [6]. Traditional viewpoint believed that only patients with HER2-positive would benefit from HER2-targeted drugs, such as trastuzumab, pertuzumab, and most recently, tucatinib and trastuzumab deruxtecan (T-DXd), resulting in improved survival [7]. Recently, there has been a change in the interpretation of HER2 status (negative/positive), which separates the HER2-low expression cases from the HER2-negative category [8]. Such change has been supported by the latest researche that observed the anti-tumour effects of the conjugated antibody T-DXd in the HER2-low subgroup of metastatic or local unresectable BC, represented by the extension of both progression-free and overall survivals [9, 10].

For pathological diagnosis, studies of intra- and inter-observer variability have found a good concordance between different observers for the distinguishment of positive and negative HER2 expression [11, 12]. Now that the clinical importance of the HER2-low subgroup has been recognized, precise identification of this subgroup becomes crucial. However, although the ASCO/CAP guidelines have defined the IHC criteria for HER2 status in BC, identifying HER2-low expression is often indecisive with low reproducibility [8]. Not like determination of HER2-negative and HER2-positive status, distinguishing the HER2 0 and HER2-low, especially HER2 1+subgroups is quite difficult and of low concordance [11, 12]. The HER2 signals might be affected by many pre-analytical and analytical issues such as formalin fixation, staining artifacts, technical diversity, and biological heterogeneity, which will give rise to both false positive and false negative results in the identification of HER2-low expression [13].

Even for the most experienced and conscientious pathologists, methodologies in addition to IHC would be of great help in determining the HER2 status of the HER2-low group of BC patients. Fortunately, machine learning-based predictors have emerged as speedy, accurate, and cost-effective approaches in predicting both HER2 status in tumor tissues and patient response to anti-HER2 treatments [14, 15]. In particular, deep learning (DL) algorithms as a set of techniques have the capacity to exploit large and complex real-world datasets for cross-domain and cross-discipline prediction and classification tasks [16, 17]. In the early stages, features of HER2 IHC images were manually obtained and utilized to generate an appropriate classifier for prediction [18]. With the rapid development of deep learning, people started to use convolutional neural network (CNN) for HER2 scoring, which was able to automatically learn high-level semantic features through a hierarchical deep architecture [19, 20]. In recent years, there has been a growing interest in predicting HER2 scores from HE slides that are more cost-effective and readily available. Artificial intelligence (AI) technology has been applied for prediction of HER2 status or scores directly from HE slides [21–23]. Representatively, Lu et al. proposed a graph neural network (GNN)-based weakly supervised method, SlideGraph+, which utilized graph structure to generate slide-level graph representation for prediction under the slide-level annotations [21]. It can capture the medical semantic relationship among different tissue regions within a single whole slide image (WSI) [21]. However, external validation of SlideGraph+in one test set didn't take HER2 2+into consideration, which tend to be a common and controversial typing, while the other external validation datasets taking HER2 2+into consideration but simply separate all cases as negative or positive without FISH results. As for accurate interpretation of HER2 IHC scores to distinguish HER2 0 and HER2 1+staining, AI-assisted technology had been reported to identify patients with HER2 0 tumors more accurately with decreased misinterpretation of HER2 1+tumors [24]. Similarly, this study merely focused on identifying HER2 0 and 1+cases, but did not contain other subtypes.

This study aims to exploit the ability of AI to provide accurate diagnosis of HER2 status from HE images under realistic clinical conditions. We also explored the reliability of HE in comparison with IHC slides as diagnostic basis, and evaluated the potential utility of computer-aided diagnosis in clinical practice.

Materials and methods

Patient cohort

The BC patient cohort with HER2 scores comprising 350 cases, named as USTC-BC, was randomly obtained from the First Affiliated Hospital of University of Science and Technology of China (USTC), from 2019 to 2022. Each case stood for a standard, high-quality HE-stained slide. In the USTC-BC, 59, 81, 105, and 105 cases were annotated as HER2 0, 1+, 2+, and 3+, respectively, according to the 2018 ASCO/CAP guidelines. These cases were allocated into the training and test sets that contained 245 (70%) and 105 (30%) patients, respectively.

The HE-stained whole-slide images with HER2 scores of 502 cases were randomly obtained from the TCGA database and named as the TCGA-BRCA cohort. In the TCGA-BRCA cohort, 57, 211, 146, and 88 cases were scored as HER2 0, 1+, 2+, and 3+, respectively. The TCGA-BRCA cohort was used as an independent test set for our model. Details of the USTC-BC and the TCGA-BRCA datasets were shown in Table 1.

Slide preparation and whole-slide images

Tissue blocks with the minimal proportion of ductal carcinoma in situ were select for hematoxylin-eosin (HE) or immunohistochemistry (IHC) stain. The IHC procedures of all slides were conducted under the same laboratory conditions, using the same equipment and reagents (Ventana Medical Systems, Roche), and the HER2 results were interpreted by two senior pathologists to ensure the consistency.

Whole slide images were obtained for all the HE-stained slides by scanning at×40 magnification with a whole-slide imaging scanner (Slide scanner SQS-120P; Shenzhen Shengqiang Technology Co, Ltd, Shenzhen, Guangdong, China).

Table 1 Composition of training data set and test data set of two data sets

Dataset type	Her2 scores	Training data set	Test data set
USTC-BC	HER2 0	41 slides	18 slides
	HER2 1 +	56 slides	24 slides
	HER2 2 +	74 slides	31 slides
	HER2 3 +	74 slides	32 slides
	Total	245 slides	105 slides
		350 slides	
TCGA-BRCA	HER2 0	40 slides	17 slides
	HER2 1 +	148 slides	63 slides
	HER2 2 +	102 slides	44 slides
	HER2 3 +	62 slides	26 slides
	Total	352 slides	150 slides
		502 slides	

Weakly supervised learning

The kernel attention transformer (KAT) method was applied to extract hierarchical context information from local regions of the WSIs and supply various diagnosis information [25] for AI model under weakly supervised learning. Before the training of AI model, the patches were extracted as 512-dimension features by a pathology language-image pre-trained network, namely the Pathology Language and Image Pre-Training (PLIP) [26], which provided the cross-modal, semantic correlation, and multi-perspective feature representations compared to the unimodal (image-based) model. The corresponding coordinates of the patches were clustered to obtain a set of anchors. Each anchor can generate the hierarchical anchor-related masks for all the patches based on spatial distance, which indicated the calculated weights between the anchor and each patch. For modeling the pair-wise dependencies between the anchor and the patch, a set of tokens were defined as the kernels, and each kernel corresponded to an anchor along with its anchor-related masks. To aggregate the information from all the kernels, the class token was used for information exchange with the kernels and achieved the prediction of HER2 score. Overall, the input of the AI model consisted of the patch features, kernels, class token, and anchor-related masks. The structure of AI model comprised multiple KAT modules. Each KAT module included two layer-normalization layers, one kernel-attention layer [7], and one feed-forward layer. During the model training process, we utilized fivefold cross-validation to enhance the stability and generalization capability of our AI model. Specifically, we randomly divide the 245 training cases into five subsets, successively selected one subset as the validation set, and used the remaining four subsets as the training set, conducting five independent training and validation processes. This approach can utilize all the data for training, minimize the bias caused by a single partition, and obtain a more comprehensive performance evaluation result. Moreover, in each fold of training, we employ early stopping to prevent model overfitting. In particular, we continuously monitor the performance on the validation set during training, and when the performance on the validation set no longer improves or even declines over a certain number of iterations, we terminate the AI model training and use the current optimal model to predict the 105 test cases. This method not only effectively avoids overfitting but also helps us determine the best test model, thereby enhancing the generalization capability and reliability of our AI model. In the end, each case was predicted using the trained AI model, and the interpretability analysis was performed and visualized with the attention heatmap.

Like most BC samples processed in clinical practice, the 350 cases in this study had limited annotation information (only available slide-level labels), which made the weakly supervised learning method particularly suitable for the HER2 scoring task. Because the high resolution of WSI hampered direct input for model training, pre-processing was performed for all the WSIs. To this end, each WSI was divided into non-overlapping patches with the fixed-size (256 × 256 pixels) and the corresponding coordinates of these patches were obtained by a window sliding strategy. After then, the background patches were removed and tissue-related patches were acquired by the tissue segmentation. Hierarchical context information from local regions of the WSIs was extracted by KAT to get various diagnosis information for AI model under weakly supervised learning. Information from all the kernels was aggregated to achieve the prediction of HER2 score (Fig. 1).

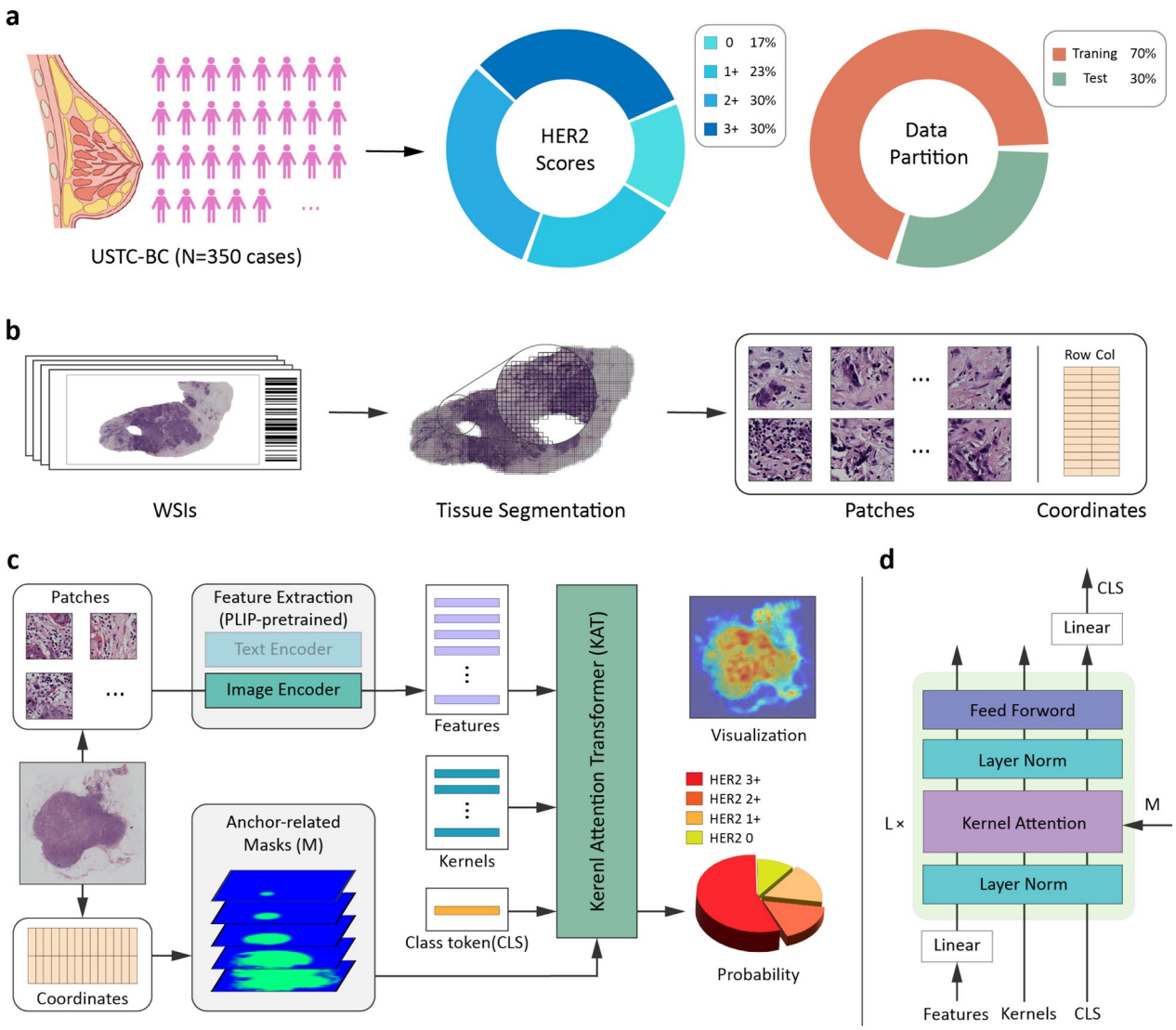

Fig. 1 Overview of the study. **a** The patient cohort of the in-house USTC-BC dataset, which displays the proportions of different HER2 expression levels and the division of data for experiments. **b** The WSI pre-processing. Each WSI is divided into a set of patches and corresponding coordinates. **c** Overview of the AI model. The network structure of the AI model is mainly composed of the Kernel Attention Transformer (KAT) modules. The input of the AI model consists of two parts: the patch features extracted by a pre-trained feature extractor and the anchors generated by the coordinates of the patches. The output includes two parts: the probability corresponding to each predicted category and the attention scores used for interpretative analysis. **d** The structure of the KAT module. Each KAT module has two layer-normalization layers, one kernel-attention layer, and one feed-forward layer

Statistical analysis

In the testing phase, the results generated by the AI model include the predicted probabilities for each category. Among these predicted probabilities, the category with the highest probability is designated as the final predicted category of the AI model. The final prediction results are then compared with the actual outcomes to assess the accuracy of the AI model. Additionally, the F1 score serves as an indicator of the AI model for improving both precision and recall. Importantly, evaluating the area under the ROC curve for each category provides a more significant measure of the generalization ability of the AI model. In this study, we utilize Python (version 3.8) with the 'numpy' (version 1.22.3 for array computation) and 'scikit-learn' (version 1.2.0 for providing some simple tools of data analysis) packages to calculate all evaluation metrics. For Fig. 2, it is conducted and analyzed using the 'matplotlib' package of Python. At last, the false positive and false negative rates on USTC-BC dataset under different threshold settings are calculated.

Results

After the training of the AI model with fivefold cross validation on data of 245 tumor cases, we conducted a comparative experiment and then measured the HER2 scoring power of our AI model on the test set of 105 tumor cases. In the comparative experiment, our AI model and the abovementioned state-of-the-art method SlideGraph + were evaluated on the TCGA-BRCA dataset and our in-house USTC-BC dataset in terms of ACC, AUC, and F1 score (Table 2). In detail, with the in-house USTC-BC dataset, our AI model obtained ACC of 0.556, AUC of 0.795, and F1 score of 0.556, standing for improvement of 6.3% in ACC, 2.7% in AUC, and 6.3% in F1 compared to SlideGraph +. Likewise, our AI model reached ACC of 0.389, AUC of 0.688, and F1 score of 0.389 on the public TCGA-BRCA dataset. Furthermore, we performed an external dataset testing experiment to validate the generalization ability of our AI model. Specifically, the TCGA-BRCA dataset was utilized as an external test dataset for our AI model trained on the in-house USTC-BC dataset, we attained metrics of 0.314, 0.579, and 0.314 for ACC, AUC, and F1, respectively, surpassing the validation results of SlideGraph + (0.301 for ACC, 0.564 for AUC, and 0.301 for F1 score). Overall, our AI model effectively predicted HER2 scores from HE-stained slides on both datasets. In addition, we generated the ROC curves of our AI model and SlideGraph + for each category in these two datasets (Fig. 2). As can be seen, it was more challenging to effectively identify the HER2 1 + compared to the HER2 0, HER2 2 +, and HER2 3 + categories, which was consistent with the frequent misidentification of HER2 1 + as HER2 0 or HER2 2 + by pathologists.

The interpretability of AI models was crucial for the understanding of the results generated by AI, particularly in predicting HER2 scores from HE slides. In this aspect, attention heatmaps were generated based on model attention scores of different tissue regions to illustrate the visualized results of our AI model (Fig. 3). Specifically, Fig. 3a shows the thumbnails of the HE slides and Fig. 3d displays the randomly magnification of HE images, all showing invasive carcinoma. Figure 3b presents the heatmaps of the HE slides, where regions with higher attention scores indicate areas that the model focuses on, which are strongly correlated with the results of the AI analysis. Whereas HE slides cannot directly reflect HER2 protein levels, the counterpart IHC slides facilitated the revisit of the attention heatmaps generated by the AI model (Fig. 3c). Indeed, the regions of high attention in the attention heatmaps were consistent with the regions of HER2 expression in the IHC images (Fig. 3). Besides, fluorescence in situ hybridization (FISH) testing have been performed on a subset of our cases including 20 cases each of HER2 0, 1 +, 2 +, and 3 + to validate the HER2 status identified by our AI model. All HER2 3 + cases showed HER2 amplification (20/20, 100%), and none of HER2 0 and 1 + cases showed HER2 amplification (0/20, 0%). Four cases of HER2 2 + showed HER2 amplification (4/20, 20%) (Fig. 3e). The FISH results were consistent with the IHC findings, further supporting the reliability of our AI model. Therefore, our AI model is a highly interpretable model with the capacity to identify tumor regions of HER2 protein expression from HE slides.

We have calculated the false positive and false negative rates on USTC-BC dataset under different threshold settings, with results presented in Supplementary Table 1. On the USTC-BC dataset, both false positive rate and false negative rate are low (< 0.25). Overall, through this approach, setting reasonable decision-making strategies can effectively control type 1 and 2 statistic errors that could affect the generalisation of the study.

Discussion

HER2 protein overexpression and/or *HER2* gene amplification occur in ~ 20% of invasive breast cancers, which has been recognized as a sole predictive marker for benefits from HER2-targeted therapy [9]. As a routine practice for newly diagnosed BC, IHC is used to screen and determine the HER2-positive and HER2-negative cases, and ISH is performed as a confirmation test for HER2 IHC equivocal cases. In clinical trials, newly developed conjugated antibodies (Trastuzumab Deruxtecan, DS-8201) demonstrated a significant effect in treating metastatic

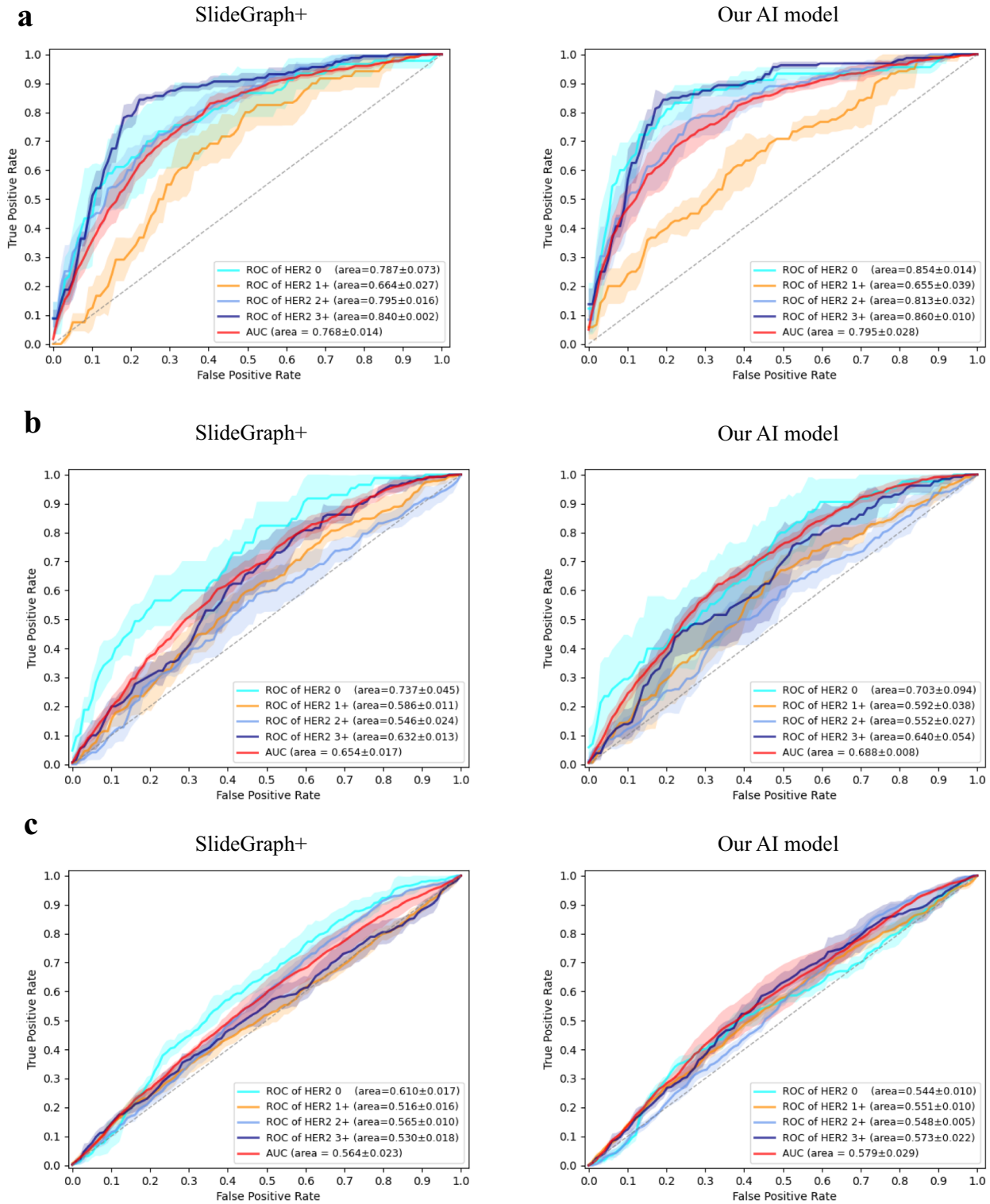

Fig. 2 The receiver operating characteristic (ROC) curves of SlideGraph + (left) and our AI model (right). These ROC curves show the area under the ROC curve (AUC) for each level of HER2 expression: **a** In the USTC-BC dataset. **b** In the TCGA-BRCA dataset. **c** In the TCGA-BRCA as an external test set

Table 2 Results (ACC, AUC, F1) of our AI model and SlideGraph+

Method	USTC-BC (HER2 0, HER2 1+, HER2 2+, HER2 3+)			TCGA-BRCA (HER2 0, HER2 1+, HER2 2+, HER2 3+)			TCGA-BRCA* (HER2 0, HER2 1+, HER2 2+, HER2 3+)		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
SlideGraph+	0.493±0.043	0.768±0.014	0.493±0.043	0.376±0.015	0.654±0.017	0.376±0.015	0.301±0.017	0.564±0.023	0.301±0.017
Our AI Model	0.556±0.050	0.795±0.028	0.556±0.050	0.389±0.020	0.688±0.008	0.389±0.020	0.314±0.020	0.579±0.029	0.314±0.020

*TCGA-BRCA was used as the external test set

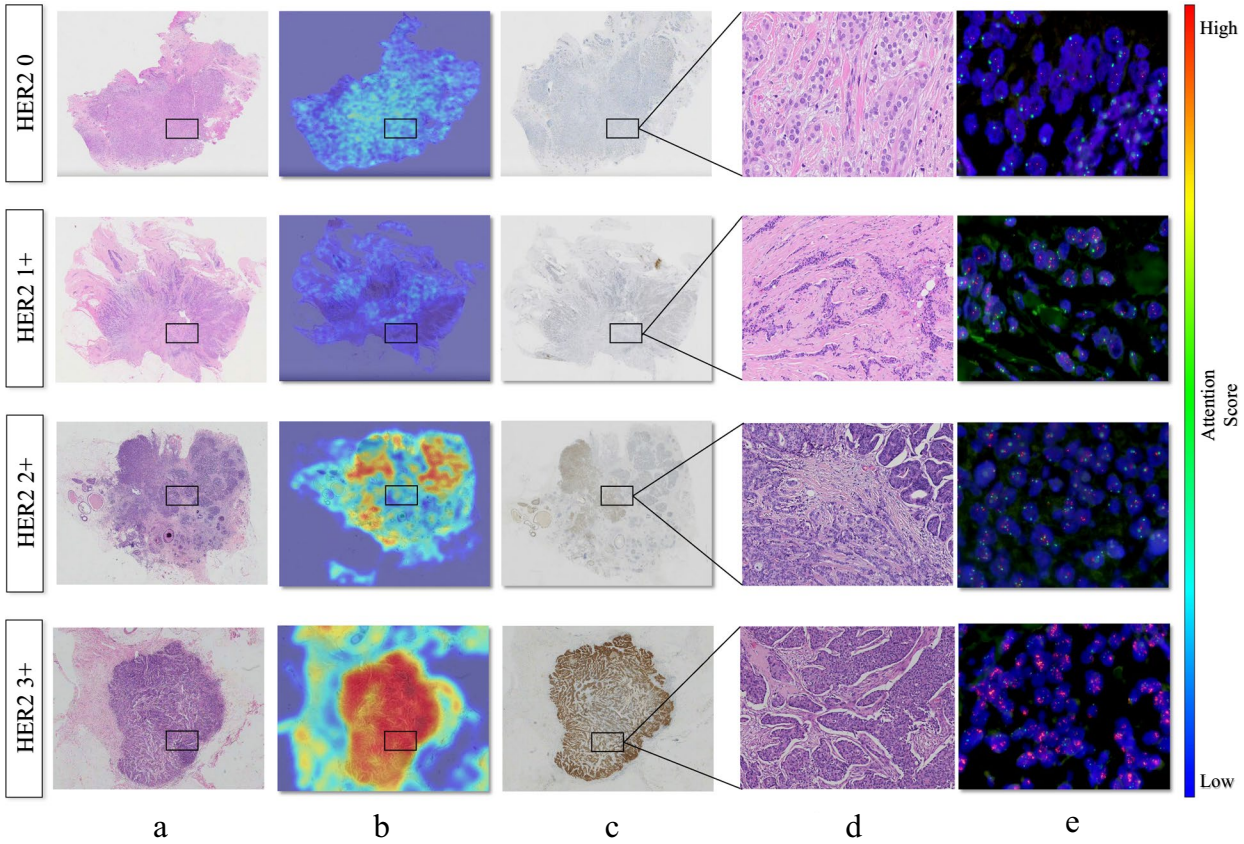

Fig. 3 Visualization based on the attention map in our AI model. **a** The overview thumbnails of HE slides. **b** The attention heatmaps. **c** The overview thumbnails of the corresponding IHC slides. **d** Randomly magnified HE stained images. **e** Fluorescence in situ hybridization (FISH) results corresponding to different HER2 expression statuses

or primary unresectable breast cancer patients, unveiling an advent that HER2-low BC becomes treatable with the new generation of HER2-directed antibody-drug conjugates (ADCs) [10].

As to HER2 interpretation, pitfalls such as intratumoral HER2 heterogeneity and increase in chromosome enumeration probe 17 signals may lead to inaccurate assessment of HER2 status. According to the 2018 ASCO/CAP guidelines, the assessment of IHC-stained slides based on visual observation is affected by the heterogeneous staining patterns and inter-pathologist variability. HER2 heterogeneity was more common in HER2 1+and HER2 2+tumors, making it challenging to distinguish HER2 1+from HER2 0 and HER2 2+. Consequently, diagnostic consistency was significantly lower in determining the difference between HER2 1+and HER2 0 or HER2 2+samples relative to that between HER2 2+and 3+cases (26% vs. 58%) [27]. Pathologist-related reasons, such as fatigue, stress, and other emotional states, also contribute to diagnostic variability. Therefore,

pathologists often diagnose equivocal samples as HER2 2+ in order to request ISH testing for a more conclusive decision, which would be costly and labor-consuming [28].

Digital pathology to a large extent would solve the abovementioned problems. Alternative methods, such as quantitative fluorescence assays or digital image analysis (DIA), have been proposed for the identification of the HER2-low category [29, 30]. With the rise of deep learning, CNN has been employed for automated HER2 scoring of IHC slides [20]. In recent years, a GrayMap + CNN model was developed to predict HER2 expression levels and gene mutations based on WSI of HER2 IHC sections with high accuracy, which demonstrated a promising future of digital pathology for accurate HER2 interpretation [31]. Compared with IHC slides, HE slides are more cost-effective and more readily available in clinical practice. Not surprisingly, there has been a growing interest in predicting HER2 expression from HE images. Based on the method of SlideGraph+, Lu et al. [21] proposed a graph neural network (GNN) to generate the slide-level graph for prediction of HER2 status at the WSI level, which was useful but had its limitations in case selection when doing external validation. In addition, Shovon et al. presented the HE-HER2 Net on the basis of transfer learning to quantify the HER2 expression status from HE images [23]. Despite the utilization of these methods in predicting HER2 scores or status from HE slides, the features of WSI extracted by these methods were not strictly correlated to the information of HER2 protein expression in IHC slides in most studies. Therefore, the interpretability of distinct models requires further research, which constitutes a major task of our AI model.

HER2 status has been evaluated by diaminobenzidine (DAB) density features standing for the areas and intensity of membranous DAB staining on IHC images that represent the level of HER2 protein expression. In HE slides, HER2 status was shown to be associated with the different types of nuclei in BCs. Based on this, we used HE images with corresponding HER2 scores and developed a deep convolutional neural network predictor to evaluate the status of HER2 expression in a given HE image region. In this work, we adopted the KAT method for training in order to extract hierarchical context information from local regions of the WSIs and supply various diagnosis information for predicting HER2 scores in BC. With the in-house USTC-BC dataset, after deep learning of the data from 245 tumor cases with fivefold cross validation, our AI model obtained ACC of 0.556, AUC of 0.795, and F1 score of 0.556 in the test set of 105 tumor cases. The evaluations of our AI model and the state-of-the-art SlideGraph+ method with the public TCGA-BRCA dataset and the in-house USTC-BC dataset for HER2 scoring demonstrated a superior performance of our AI model in terms of ACC, AUC and F1 score. To validate the generalization performance of our AI model, we conducted validation experiments using an external test set. Specifically, our AI model was trained on data from the USTC-BC dataset and then tested on data from the TCGA-BRCA dataset. The metrics of ACC, AUC, and F1 were all superior to the results of SlideGraph+. Furthermore, from the ROC curves of our AI model and SlideGraph+ for each category in the two datasets, it was obviously more challenging to effectively identify HER2 1+ compared to HER2 0, HER2 2+, and HER2 3+ categories, which was consistent with the frequent misidentification of HER2 1+ as HER2 0 or HER2 2+ on conventional microscopic examination. The interpretation consistency of HER2 1+ was worse than that of HER2 0 and HER2 2+, but it was improved significantly with the help of AI [24]. DAB density evaluation with the help of AI gives much higher AUC, indicating a better HER2 prediction performance than the DAB density estimates only.

To assess the interpretability and capability of our AI model, the attention heatmaps generated by our AI model for the HE slides were compared with the corresponding the staining patterns of the corresponding IHC slides due to HE slides cannot directly reflect HER2 protein levels. In the HER2-positive cases, large areas were estimated as of high DAB density that were displayed in orange and red in the heatmaps. In contrast, in HER2-negative cases, the majority of the tissue region was estimated as of low DAB density lacking HER2 protein expression. The highlighted activation areas in the heatmaps are consistent with the DAB density in the corresponding IHC images. This supports the idea of using DAB density as a potential feature for HER2 status prediction. It could be observed that only few areas in the HER2-negative samples contribute to the HER2 positivity, whereas the majority of tissue regions in the positive cases have high HER2 prediction scores. It should also be noted that regions with high HER2 prediction scores are consistent with high DAB intensity areas in the corresponding IHC images. Visualization of HER2 scores demonstrated the capacity of our AI model in identifying regions with HER2 protein expression from HE slides. Although the aforementioned methods can predict the HER2 scores or status directly from HE slides, the features learned from the HE slides may not fully reflect the expression level of HER2 protein hidden in the IHC slide, and the model interpretability may need further exploration [21–23, 30]. In conclusion, compared with previous methods, our AI model is more powerful in prediction of HER2 status with HE instead of IHC sections, including HER2 0, 1+,

2 + and 3 + cases, while the study of Wu et al. [24] focused only on differentiating HER2 0 and HER2 1 + tumors, and the method of SlideGraph + skipped HER2 2 + cases [21]. Up to now, few studies have evaluated the role of AI in differentiating all HER2 status by HE-staining, except for our AI model.

Despite the promising results demonstrated in this study, several limitations and challenges should be addressed, which highlight the need for further refinement and validation. One critical limitation lies in the dataset, whose sample size remains insufficient to ensure robust model generalizability, thereby necessitating expansion to include a larger and more diverse cohorts. Furthermore, the dataset suffers from class imbalance, particularly in the HER2 0 and 1 + categories, which, despite the application of resampling strategies, continues to undermine model performance and may lead to biased predictions. The predictive performance of the AI model, especially in distinguishing between HER2 0 and 1 + categories, requires significant improvement, as these categories exhibit subtle pathological differences that complicate accurate classification.

In addition to these limitations, the study faces several challenges that hinder the clinical applicability of the model. Variations in slide preparation, staining protocols, and digitization techniques across different medical centers introduce substantial heterogeneity into the data, which limits the generalizability of the AI model and complicates its deployment in diverse clinical settings. Another challenge arises from the subjective nature of pathological annotations, which rely heavily on the judgment of individual pathologists and may lead to inconsistencies that affect both model training and validation. Given these challenges, the current performance of AI model remains insufficient for clinical implementation due to inadequate accuracy and reliability.

To address these limitations and challenges, future research will focus on several critical directions. Multi-center data will be incorporated to enhance the ability of AI model to handle data heterogeneity, which is critical for improving generalizability across different institutions. Additional training data will be collected to address class imbalance and further refine model performance, particularly for underrepresented categories. Moreover, multi-modal data integration will be explored to provide more comprehensive information, thereby enhancing the predictive capabilities of the model. These efforts, which aim to bridge the gap between research and clinical practice, are essential for ensuring the model's utility and reliability in real-world applications. Of course, from the perspective of treatment and clinical outcome, in future research, we will continue the follow-up to evaluate the relationship between HER2 interpretation results by our AI model with long-term clinical outcomes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13058-025-01998-8.

Additional file1 (DOC 30 KB)

Author contributions

Study conception and design contributed by JD, JS, YZ, HW. Acquisition of data con-tributed by JD, GL, JC, WW, HW. Statistical analysis contributed by JD, JS, DS, YZ, WW. Machine learning analysis contributed by JS, DS, YW, YZ. Writing, drafting, and reviewing manuscript contributed by JD, JS, DS, GL, JC, WZ, YZ, HW. All authors reviewed the manuscript.

Funding

This work was partly supported by Joint Fund for Medical Artificial Intelligence (No. MAI2023C014), partly supported by Research Funds of Centre for Leading Medicine and Advanced Technologies of IHM (No. 2023IHM01043), partly supported by the Anhui Provincial Natural Science Foundation (No. 2408085MF162), partly supported by the National Natural Science Foundation of China (No. 62171007, 61906058, 61901018, and 61771031), partly supported by Beijing Natural Science Foundation (Grant No. 7242270), partly supported by the Fundamental Research Funds for the Central Universities of China (No. YWF-23-Q-1075), partly supported by Scientific Research Project of Anhui Provincial Education Department (No. 2023AH040404), partly supported by undergraduate college student innovation and entrepreneurship training program (No. S202210358114).

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

This study was approved by Medical Research Ethics Committee of the First Affiliated Hospital of the University of Science and Technology of China (No. 2023KY-378). Patient consent was not required because all samples were archival.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Pathology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230036, Anhui, China. [2]Intelligent Pathology Institute, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230036, Anhui, China. [3]School of Software, Hefei University of Technology, Hefei 230601, China. [4]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, Anhui, China. [5]School of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230027, Anhui, China. [6]School of Engineering Medicine, Beijing Advanced Innovation Center on Biomedical Engineering, Beihang University, Beijing 100191, China. [7]Department of Pathology, Centre for Leading Medicine and Advanced Technologies of IHM, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230001, Anhui, China.

Received: 19 December 2024 Accepted: 10 March 2025
Published online: 18 April 2025

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 Countries. CA Cancer J Clin. 2024. https://doi.org/10.3322/caac.21834.
2. Cao W, Chen HD, Yu YW, Li N, Chen WQ. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. Chin Med J (Engl). 2021;134:783–91.
3. Xu Y, Gong M, Wang Y, Yang Y, Liu S, Zeng Q. Global trends and forecasts of breast cancer incidence and deaths. Sci Data. 2023;10:334.
4. Tai W, Mahato R, Cheng K. The role of HER2 in cancer therapy and targeted drug delivery. J Control Release. 2010;146:264–75.
5. Marchiò C, Annaratone L, Marques A, Casorzo L, Berrino E, Sapino A. Evolving concepts in HER2 evaluation in breast cancer: heterogeneity, HER2-low carcinomas and beyond. Semin Cancer Biol. 2021;72:123–35.
6. Wolff AC, Hammond MEH, Allison KH, Harvey BE, Mangu PB, Bartlett JMS, et al. Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. J Clin Oncol. 2018;36:2105–22.
7. Swain SM, Shastry M, Hamilton E. Targeting HER2-positive breast cancer: advances and future directions. Nat Rev Drug Discov. 2023;22:101–26.
8. Tarantino P, Hamilton E, Tolaney SM, Cortes J, Morganti S, Ferraro E, et al. HER2-low breast cancer: pathological and clinical landscape. J Clin Oncol. 2020;38:1951–62.
9. Ahh S, Woo JW, Lee K, Park SY. HER2 status in breast cancer: changes in guidelines and complicating factors for interpretation. J Pathol Transl Med. 2020;54:34–44.
10. Modi S, Jacot W, Yamashita T, Sohn J, Vidal M, Tokunaga E, et al. Trastuzumab deruxtecan in previously treated HER2-low advanced breast cancer. N Engl J Med. 2022;387:9–20.
11. Gavrielides MA, Gallas BD, Lenz P, Badano A, Hewitt SM. Observer variability in the interpretation of HER2/neu immunohistochemical expression with unaided and computer-aided digital microscopy. Arch Pathol Lab Med. 2011;135:233–42.
12. Hsu CY, Ho DMT, Yang CF, Lai CR, Yu IT, Chiang H. Interobserver reproducibility of Her-2/neu protein overexpression in invasive breast carcinoma using the DAKO HercepTest. Am J Clin Pathol. 2002;118:693–8.
13. Bussolati G, Annaratone L, Maletta F. The pre-analytical phase in surgical pathology. Recent Results Cancer Res Fortschritte Krebsforsch Progres Dans Rech Sur Cancer. 2015;199:1–13.
14. Farahmand S, Fernandez AI, Ahmed FS, Rimm DL, Chuang JH, Reisenbichler E, 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;35:44–51.
15. Yousif M, Huang Y, Sciallis A, Kleer CG, Pang J, Smola B, et al. Quantitative image analysis as an adjunct to manual scoring of ER, PgR, and HER2 in invasive breast carcinoma. Am J Clin Pathol. 2022;157:899–907.
16. Nguyen H, Kieu LM, Wen T, Cai C. Deep learning methods in transportation domain: a review. IET Intell Transp Syst. 2018;12:998–1004.
17. Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K. Speech recognition using deep neural networks: a systematic review. IEEE Access. 2019;7:19143–65.
18. Mukundan R. Analysis of image feature characteristica for automated scoring of HER2 in histology slides. Journal of Imaging. 2019;5:35–46.
19. Saha M, Chakraborty C. Her2net: a deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. IEEE Trans Image Process. 2018;27:2189–200.
20. Qaiser T, Rajpoot NM. Learning where to see: a novel attention model for automated immunohistochemical scoring. IEEE Trans Med Imaging. 2019;38:2620–31.
21. Lu W, Toss M, Dawood M, Rakha E, Rajpoot N, Minhas F. Slidegraph+: whole slide image level graphs to predict HER2 status in breast cancer. Med Image Anal. 2022;80:102486–98.
22. Wang J, Zhu X, Chen K, Hao L, Liu Y. Hahnet: A convolutional neural network for HER2 status classification of breast cancer. BMC Bioinform. 2023;24:353–68.
23. Shovon MSH, Islam MJ, Nabil MNAK, Molla MM, Jony AI, Mridha M. Strategies for enhancing the multi-stage classification performances of HER2 breast cancer from hematoxylin and eosin images. Diagnostics. 2022;12:2825–45.
24. Si Wu, Yue M, Zhang J, Li X, Li Z, Zhang H, et al. The role of artificial intelligence in accurate interpretation of HER2 immunohistochemical scores 0 and 1+ in breast cancer. Mod Pathol. 2023;36(3):100054–63.
25. Zheng Y, Li J, Shi J, Xie F, Huai J, Cao M, et al. Kernel attention transformer for histopathology whole slide image analysis and assistant cancer diagnosis. IEEE Trans Med Imaging. 2023;42:2726–39.
26. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual-language foundation model for pathology image analysis using medical Twitter. Nat Med. 2023;29:2307–16.
27. Fernandez AI, Liu M, Bellizzi A, Brock J, Fadare O, Hanley K, et al. Examination of low ERBB2 protein expression in breast cancer tissue. JAMA Oncol. 2022;8:1–4.
28. Moelans C, de Weger R, Van der Wall E, van Diest P. Current technologies for HER2 testing in breast cancer. Crit Rev Oncol. 2011;80:380–92.
29. Moutafi M, Robbins CJ, Yaghoobi V, Fernandez AI, Martinez-Morilla S, Xirou V, et al. Quantitative measurement of HER2 expression to subclassify *ERBB2* unamplified breast cancer. Lab Invest. 2022;102:1101–8.
30. Anand D, Kurian NC, Dhage S, Kumar N, Rane S, Gann PH, et al. Deep learning to estimate human epidermal growth factor receptor 2 status from hematoxylin and eosin-stained breast tissue images. J Pathol Inform. 2020;11:19–26.
31. Yao Q, Hou W, Wu K, Bai Y, Long M, Diao X, et al. Using whole slide gray value map to predict HER2 expression and FISH status in breast cancer. Cancers. 2022;14:6233–46.

Publisher's Note