The Journal of Pathology: Clinical Research
J Pathol Clin Res March 2025; **11**: e70024
Published online in Wiley Online Library
(wileyonlinelibrary.com). DOI: 10.1002/2056-4538.70024

ORIGINAL ARTICLE

Enhancing HER2 testing in breast cancer: predicting fluorescence *in situ* hybridization (FISH) scores from immunohistochemistry images via deep learning

Daniel O Macaulay*, Wenchao Han, Mark D Zarella, Chris A Garcia and Thomas E Tavolara

Division of Computational Pathology and AI, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA

**Correspondence to: Daniel O Macaulay, Division of Computational Pathology and AI, Department of Laboratory Medicine and Pathology, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905, USA. E-mail: macaulay.daniel@mayo.edu*

Abstract

Breast cancer affects millions globally, necessitating precise biomarker testing for effective treatment. HER2 testing is crucial for guiding therapy, particularly with novel antibody–drug conjugates (ADCs) like trastuzumab deruxtecan, which shows promise for breast cancers with low HER2 expression. Current HER2 testing methods, including immunohistochemistry (IHC) and *in situ* hybridization (ISH), have limitations. IHC, a semi-quantitative assay, is prone to interobserver variability. While ISH provides higher precision than IHC, it remains more resource-intensive in terms of cost and workflow. However, turnaround time is typically faster than that of other advanced molecular methods such as next-generation sequencing. We adapted the clustering-constrained-attention multiple-instance deep learning model to improve IHC testing and reduce dependence on reflex fluorescence ISH (FISH) tests. Using 5,731 HER2 IHC images, including 592 cases with FISH testing, we trained two models: one for predicting HER2 scores from IHC images and another for predicting FISH scores from equivocal cases. The HER2 IHC score prediction model achieved 91% ± 0.01 overall accuracy and a receiver operating characteristic (ROC) area under the curve (AUC) of 0.98 ± 0.01. The FISH score prediction model had an ROC AUC of 0.84 ± 0.07, with sensitivity at 0.37 ± 0.13 and specificity at 0.96 ± 0.03. External validation on cases from 203 institutions showed similar performance. The HER2 IHC model maintained a 91% ± 0.01 accuracy and an ROC AUC of 0.98 ± 0.01, while the FISH model had an ROC AUC of 0.75 ± 0.03, with sensitivity at 0.28 ± 0.04 and specificity at 0.93 ± 0.01. Our model advances HER2 scoring by reducing subjectivity and variability in current scoring methods. Despite lower accuracy and sensitivity in the FISH prediction model, it may be a beneficial option for settings where reflex FISH testing is unavailable or prohibitive. With high specificity, our model can serve as an effective screening tool, enhancing breast cancer diagnosis and treatment selection.

Keywords: HER2; deep learning; breast cancer; immunohistochemistry; biomarkers; *in situ* hybridization; fluorescence

Received 23 July 2024; Revised 8 February 2025; Accepted 18 February 2025

No conflicts of interest were declared.

Introduction

Human epidermal growth factor receptor 2 (HER2) is a tyrosine-kinase transmembrane receptor that has been identified as a role player in the aggressive growth of breast cancer cells [1]. About 20% of breast cancer cases exhibit HER2 positivity, and testing is recommended for all primary invasive breast cancers, as HER2 status is critical for guiding targeted therapies [1,2]. Thanks to the discovery of new anti-HER2 therapies, the prognosis for patients with HER2-positive breast cancer has significantly improved [3]. Antibody-drug conjugates (ADCs) like trastuzumab deruxtecan have changed the landscape significantly [2,4]. A subclass of HER2-negative tumors, known as HER2-low (characterized by low HER2 expression that does not meet the criteria for HER2 positivity), has been shown to benefit from targeted ADC therapies [5]. Thus, the emergence of expanded treatment options further demonstrates the importance of precise biomarker measurement for effective therapy selection in breast cancer.

Immunohistochemistry (IHC) and *in situ* hybridization (ISH) assays are mainstays for the quantification of

HER2 expression, each with unique protocols and interpretation challenges. IHC scores range from 0 to 3+, with 0 and 1+ scores denoting HER2-negative cases, and a 3+ score being positive. According to the American Society of Clinical Oncology-College of American Pathologists (ASCO-CAP) guidelines, a score of 2+ is equivocal and would require reflex fluorescence ISH (FISH) testing [1]. IHC readout offers a semi-quantitative assessment of HER2 protein overexpression in cancer cells but is prone to variability in interpretation due to the qualitative nature of the scoring system [6,7]. Such variability can ultimately affect treatment decisions and could potentially lead to poor patient outcomes [7]. Fluorescence ISH, which detects *HER2* gene amplification, provides a more quantitative measure. However, it is significantly more costly and time-intensive to perform. About 25% of breast cancer tumors express borderline/equivocal levels of the HER2 protein and thus require reflex FISH testing [8]. The need for skilled personnel, specialized equipment, and probes all add to operational costs. Furthermore, the turnaround time for FISH results can delay treatment decisions, thus negatively affecting patient care. These limitations present a need for more efficient and cost-effective HER2 testing strategies. An ideal testing approach would be one that addresses these limitations without compromising accuracy.

Artificial intelligence (AI) applications in healthcare have made great strides in recent years, specifically in the diagnostic processes of diseases [9]. Deep learning (DL) techniques have been applied in classification tasks for a substantial number of cancer types, including breast cancer [10,11]. In HER2 biomarker testing, there exists the opportunity to overcome the limitations of traditional diagnostic methods with DL models, with some published reports showing moderate success in their use [12]. Bannier *et al* trained a feature extractor learning model using 675 HER2 IHC tiles and showed improved performance when compared to an extractor trained on H&E tiles [13]. Rasmussen *et al* trained a CNN with EfficientNet B0 architecture on 115 IHC whole slide images (WSIs) and reported an overall accuracy of 79% [14].

Here we propose an integrated DL model that can predict the *HER2* amplification of a given test case by HER2 IHC whole slide imaging alone. The first model predicts the IHC scores, that is, 0, 1+, 2+, or 3+. If a case is determined by the model to have a 2+ score, it is passed to the second model that has been pretrained on HER2 cases that have undergone FISH scoring. This model predicts the final *HER2* amplification score, a binary output, that is, positive or negative. In the context of clinical implementation, our model could be integrated into the diagnostic workflow at the point where HER2 IHC scoring is performed, serving as an intermediary screening tool to prioritize or even replace reflex FISH testing in certain scenarios. We adapted the widely utilized clustering-constrained-attention multiple-instance *learning* (CLAM) model, a weakly supervised multiple-instance learning model [15]. To achieve all this, we leveraged our extensive dataset of HER2 IHC WSIs spanning nearly a decade's worth of histopathological data collected and scanned at the Mayo Clinic. To our knowledge, this dataset is the largest that has ever been used to train a HER2 breast cancer scoring algorithm.

By utilizing IHC images as a predictive tool for FISH results, we also hope to help address the economic challenges of current HER2 testing protocols by offering a viable option for limited-resource settings [16].

Materials and methods

Dataset acquisition and preparation

A comprehensive dataset comprising 5,731 HER2 IHC unique cases was sourced from the Mayo Clinic Department of Laboratory Medicine and Pathology's Aperio eSlide manager database (Leica Biosystems, Nussloch, Germany), spanning from June 2014 to June 2023. This included 592 cases that had undergone reflex FISH assays. This dataset predominantly included invasive breast carcinoma specimens, covering a range of HER2 scores from 0 to 3+. Collection and analysis of all tissue image data was conducted with approval from the institutional research ethics board. Patient clinical characteristics data is shown in Table 1.

Slides were scanned with the Aperio ScanScope AT2 at ×20 magnification with a quality factor of 70. WSIs underwent manual annotation by skilled technicians under a breast pathologist's supervision, with invasive carcinoma regions delineated using Aperio ImageScope software. Annotations were saved in XML format, facilitating the extraction of regions of interest (ROIs) for the model training process. Mayo Clinic sees a high volume of consult cases, where tissue slides are sourced from 203 secondary and tertiary institutions. Consequently, the diversity represented in the slide and staining attributes, as well as the patient demographics, provides a unique resource for AI/machine learning (ML) development not previously available at this scale. From these cases specifically,

Table 1. Patient clinicopathologic characteristics data

Characteristic	Internal cases ($n = 5,796$) Count	%	External cases ($n = 1,000$) Count	%
Age				
20–39	315	5	43	4
40–49	785	14	117	12
50–59	1,252	22	200	20
60–69	1,606	28	282	28
70–79	1,235	21	240	24
80–89	499	9	100	10
90–99	68	1	19	2
Sex				
Female	5,719	99	990	99
Male	77	1	10	1
Tumor histotype				
Ductal carcinoma *in situ*	35	1	1	<1
Invasive carcinoma of no special type	4,921	85	980	98
Invasive carcinoma with other features	170	3	2	<1
Invasive lobular carcinoma	543	9	4	<1
Other	128	2	12	1
Specimen type				
Biopsy	4,500	78	964	96
Excision	1,087	19	34	3
Fine-needle aspiration	47	1	0	0.0
Mastectomy	162	3	2	<1
HER2 status by IHC				
0	2,175	37	320	32
+1	1732	30	449	45
+2	1,321	23	127	13
+3	568	10	104	10
FISH status				
Negative	506	8.8	1712	87
Positive	86	1.5	255	13

external validation was done using 1,000 invasive breast carcinoma cases for the HER2 score predictor model and 1967 cases with reflex FISH scores for validation of the FISH scoring model. Consistency in scoring criteria was ensured in reviews conducted by Mayo Clinic pathologists. The methodology for model assessment paralleled that of the training set, with a focus on annotated ROIs and their classification via the trained model. The stepwise process for case selection and screening is depicted in Figure 1A.

HER2 IHC and FISH protocols

HER2 testing protocols were applied to 5-μm thick formalin-fixed, paraffin-embedded (FFPE) tissue sections. Likewise, fine-needle aspirates were processed from 10% neutral-buffered formalin specimens. The IHC staining process used the PathWay HER2 4B5 rabbit monoclonal primary antibody (Ventana Medical Systems, Oro Valley, AZ, USA).

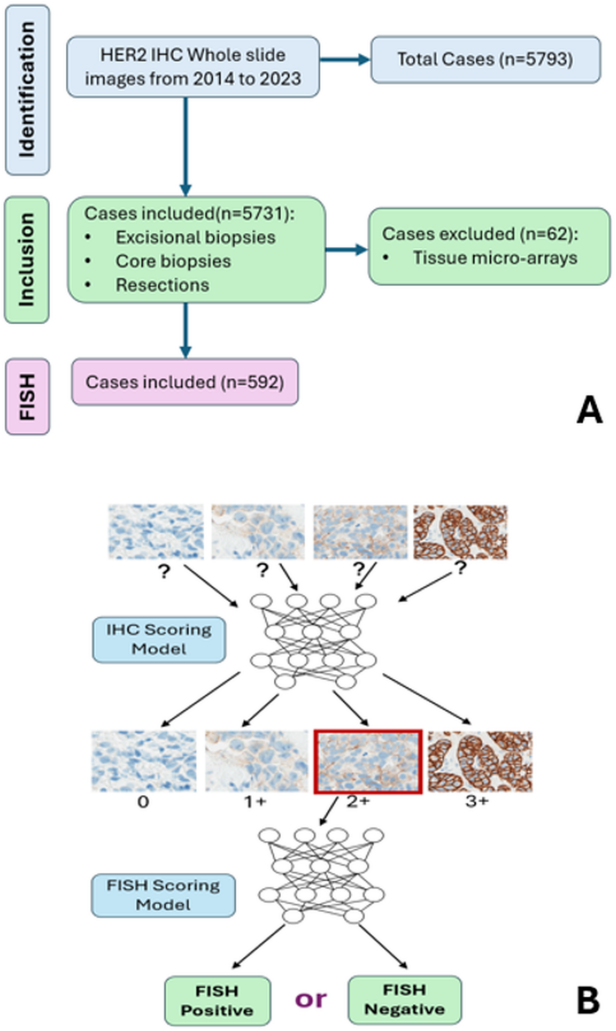

Figure 1. (A) Flowchart showing case selection process. (B) Framework of IHC and FISH score predictor models.

This comes in a ready-to-use predilute concentration as supplied by the manufacturer. Antibody detection was done using the UltraView detection kit (Ventana) with heat-induced epitope retrieval (CC1 36 min). Positive and negative controls were affixed to each tissue slide that underwent IHC staining, with interpretations aligning with the 2018 ASCO-CAP guidelines. FISH testing employed the PathVysion *HER2* DNA probe kit (Abott Molecular, Des Plaines, IL, USA), featuring both a *HER2*-specific probe and a chromosome 17 centromere probe (D17Z1). Tissue target areas were selected by a pathologist and then etched with a diamond-tipped engraving tool. The probe was then hybridized to the target areas and subsequently analyzed by two laboratory technologists.

Model architecture and training

The CLAM model is a versatile DL framework adaptable to various WSI classification tasks. It integrates a pretrained ImageNet base and leverages an attention-based multiple-instance learning strategy, identifying diagnostically significant regions through an attention mechanism. Additionally, it incorporates an instance-level loss by pseudo-labeling highly attended and unattended regions identified by the attention mechanism. One of its defining features is the lack of need for tile-level labels, requiring only slide-level labels to predict HER2 scores. The dataset was randomized and divided into training (80%), validation (10%), and test (10%) sets, with a 10-fold Monte Carlo cross-validation. Model implementation was done using the PyTorch framework, supported by a dedicated 6-GPU cluster with V100 GPUs. Figure 1B demonstrates the framework of approach, consisting of two basic stages: IHC score prediction and FISH score (*HER2* amplification) prediction.

As a baseline comparison, we replicated the method from Rasmussen *et al* [14] and trained it using our institutional data. EfficientNet B0 is a widely recognized convolutional neural network architecture that has been extensively used for image classification tasks in various domains. It uses weights from pretrained ImageNet data.

Additionally, we implemented both pixel-based and cell-based IHC quantification algorithms commonly used in commercial and open-source software – HALO, Visiopharm, and QuPath. The hyperparameters for these methods were independently optimized on a random selection of 100 HER2 slides (with equal 0, 1+, 2+, and 3+ cases) using simulated annealing to maximize overall classification accuracy. These optimized algorithms were then applied to the whole dataset.

Image processing and annotation

Tumor ROIs were extracted from previous annotations done by skilled laboratory personnel under the supervision of a breast pathologist. ROIs from each WSI were then extracted into 256 × 256 pixel patches at 0.4958 μm/pixel magnification without overlap. All images contained in the dataset had been annotated prior to splitting into training, validation, and test sets. Features were extracted using an ImageNet-pretrained Resnet50 encoder, as in the typical CLAM framework, generating 1,024-dimensional features for each patch.

Evaluation

Model performance was evaluated at the WSI individual case level, assessing the ability to predict HER2 scores and FISH amplification scores. Performance metrics, including the receiver operating characteristic area under the curve (ROC AUC), sensitivity, and specificity, were calculated for each cross-validation fold. Mean and standard deviation (SD) for each metric were calculated across folds. We validated the HER2 score predicted by our model against the HER2 score from the diagnostic report by our pathologist. For external validation cases, the models were likewise assessed at the slide level.

Results

Model performance metrics

The evaluation of the models' proficiency in predicting HER2 score with slide-level labels yielded promising results. The base HER2 scoring model demonstrated a weighted mean ROC AUC of 0.98 (SD ± 0.002), with an overall accuracy of 0.91 (SD ± 0.014). Figure 2 depicts the confusion matrix and ROC curves for the IHC scoring model.

Three experimental designs were set up and run for *HER2* FISH score prediction. The best performing model demonstrated a mean ROC AUC of 0.84 (SD ± 0.07) with a mean sensitivity of 0.37 (±0.13) and a mean specificity of 0.96 (±0.03). This was achieved by oversampling the positive class of the training and validation sets of the FISH prediction model. For the second experiment, we replicated other published methods for FISH prediction from HER2 IHC, using our own IHC imaging data. This yielded a mean ROC AUC of 0.65 (±0.05), with sensitivity of 0.21 (±0.18) and specificity of 0.90 (±0.06). Finally, we experimented with using incremental amounts of our training data to see if there was a minimum threshold of training data that would yield a peak overall prediction accuracy. Our findings indicate that model accuracy improves proportionally with an increase in the amount of training data, suggesting that access to more data may boost model performance (supplementary material, Table S1). Figure 3 depicts the confusion matrices and ROC AUC curves for all the experiments. Table 2 compares the evaluation metrics for all experiments. To further evaluate the performance of the FISH scoring model, we analyzed predictions for a subset of nonequivocal cases, especially HER2 IHC scores of 1+ and 3+. Understanding the model's behavior across these categories is essential to assess its generalizability beyond equivocal (2+) cases. Confusion matrices for these subsets have been included in the supplementary section (supplementary

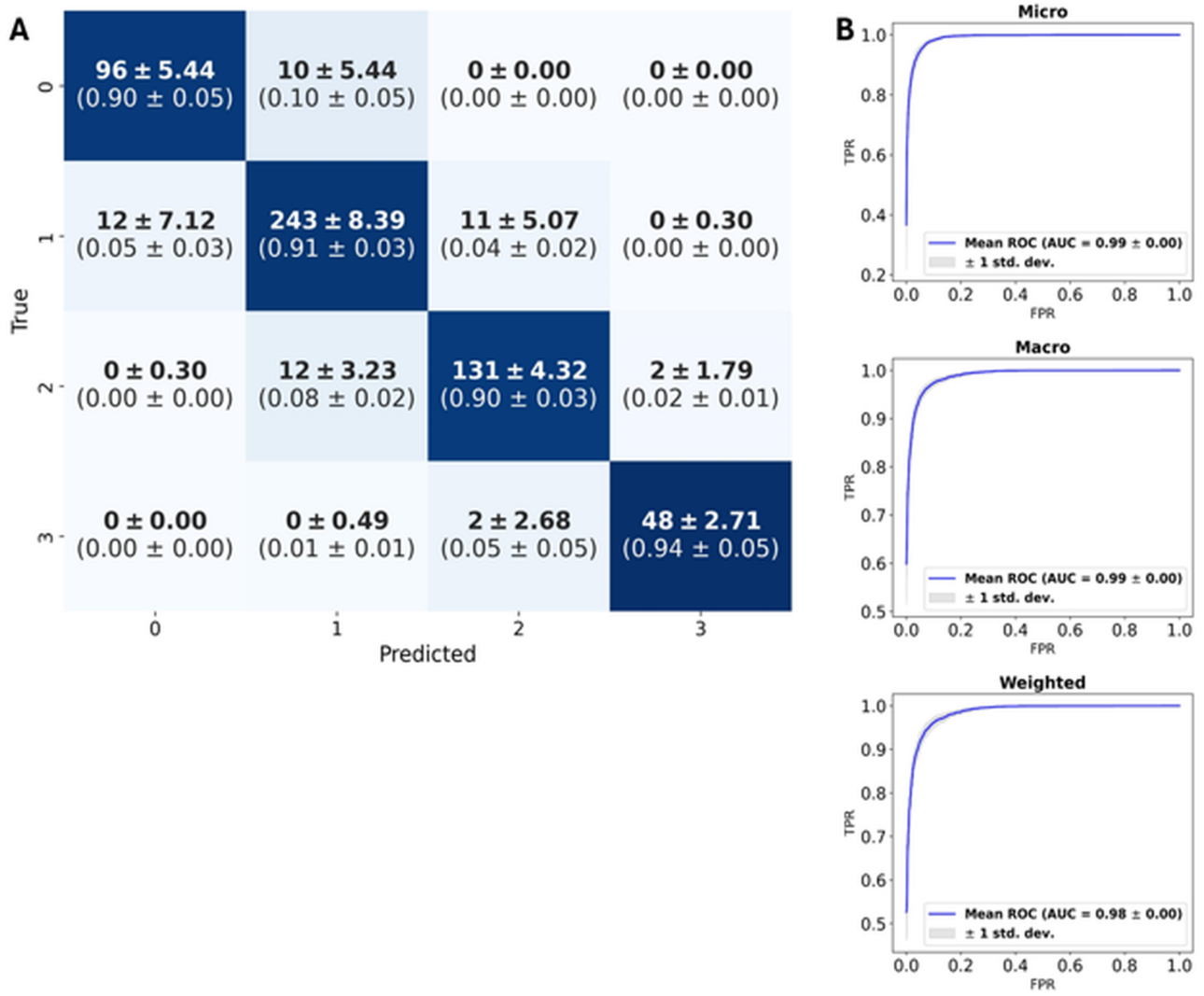

Figure 2. (A) Row-normalized confusion matrix showing the performance of the IHC scoring model based on 10-fold cross-validation data. (B) Multi-class ROC AUC curves showing model performance across 10-folds of test data.

material, Figure S1 for institutional data and Figure S2 for external data).

Error analysis

Of the 387 misclassified WSIs/cases in our base model, 98% (383) were adjacent to their ground truth scores, that is, true score ± 1. A total of 45% (178) cases were either 0+ misclassified as 1+ or 1+ misclassified as 0+. Another 82 WSIs/cases were 1+ misclassified as 2+, and 81 images were 2+ misclassified as 1+. We did not observe a pattern of misclassification due to tumor size, tumor infiltration pattern, or tissue size. Some images that were misclassified with a higher score had the presence of hemosiderin granules, which may have influenced prediction. Slide overviews with the representative regions highlighted along with the magnified ROIs are demonstrated in Figure 4.

Previous studies on interobserver concordance have shown a high variance in agreement between pathologists, depending on the agreement criteria used [17]. Rates of discordance are typically higher when IHC scores range between 0 and 1+, with these cases accounting for nearly 50% of the discordant cases [6,18]. While our error rates mirror the literature in terms of percentage cases that contribute to discordant rates, the overall agreement (about 91%) between our base model and the ground truth far supersedes the interobserver agreement that has been observed in

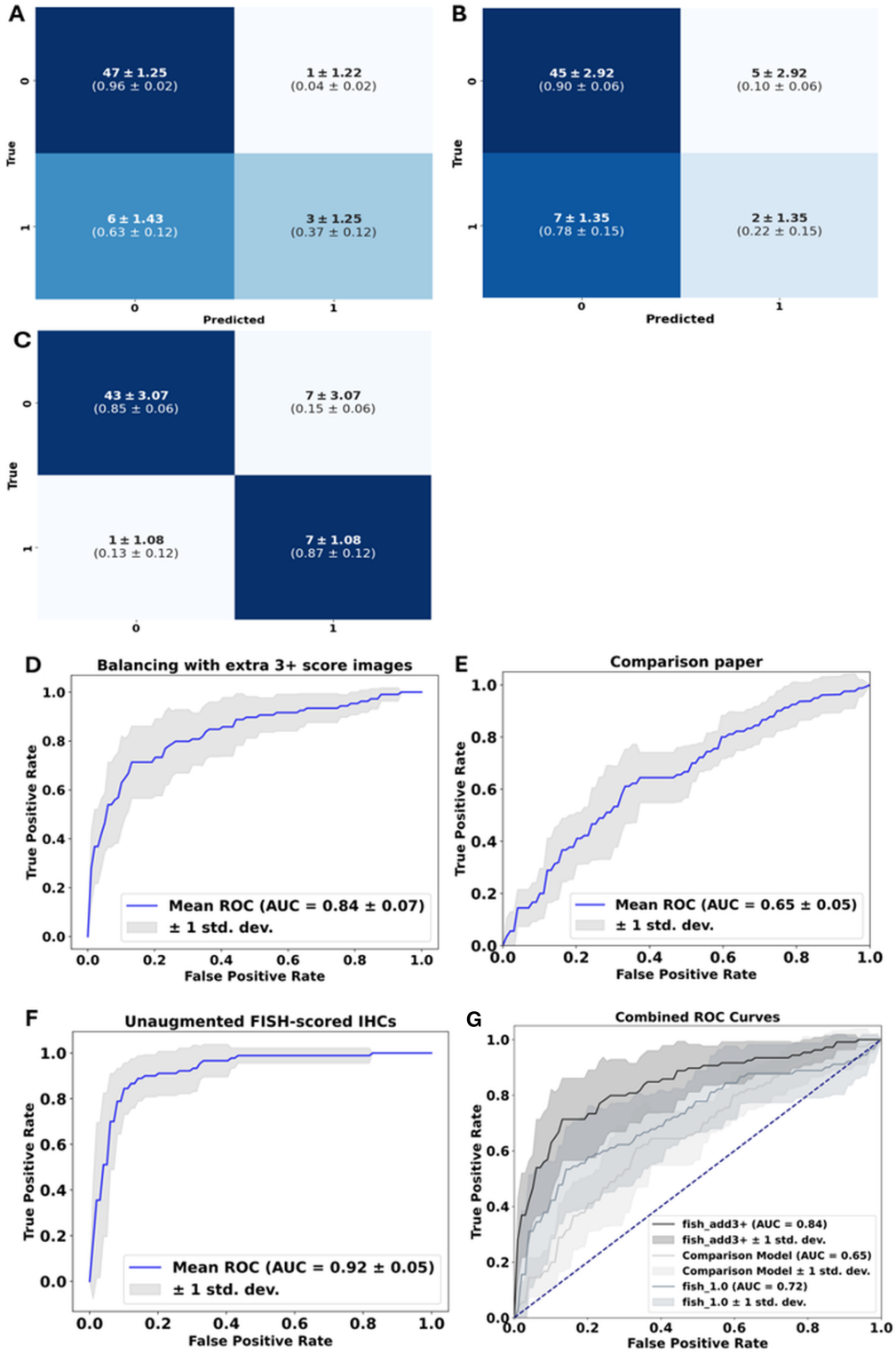

Figure 3. (A–C) Row-normalized confusion matrices showing FISH score prediction performances for (A) the model with added 3+ images, (B) the comparison paper method, and (C) the baseline model. (D–F) ROC curves for FISH score performances for the three experiments shown in A, B, and C, respectively. (G) A combined ROC curve for all experiments.

Table 2. Comparing ROC AUC plots for all FISH scoring models

	ROC AUC (SD)	Sensitivity (SD)	Specificity (SD)
Comparison paper	0.65 (±0.05)	0.21 (±0.18)	0.90 (±0.06)
Baseline FISH-scored IHC images	0.92 (±0.05)	0.32 (±0.23)	0.93 (±0.06)
Adding extra 3+ images	0.84 (±0.07)	0.37 (±0.13)	0.96 (±0.03)

previous studies. Figure 5 shows a single randomly selected WSI with model attention maps. The regions with the highest attention exhibit a greater concentration of membrane staining compared to the least attended regions.

We also randomly selected 100 samples from the cases misclassified by the HER2 IHC model. We removed the ground-truth scores and model-derived scores, and we asked an expert pathologist to re-score them. The goal was to see what the rate of agreement was between our model and a second pathologist when considering only wrongly classified cases. The results showed that the pathologist agreement with the ground-truth score was much higher (81%) than the pathologist agreement with the model (17%), strongly indicating that misclassifications were far more likely to be due to AI error rather than original pathologist error. See supplementary material, Figure S3 for more information.

Impact of data augmentation on model accuracy

A few data augmentation techniques were tried to bolster the model's accuracy. These included bootstrapping techniques involving randomly combined patch-level instances from positive and negative samples to create novel training samples. The iBOT method, which uses a masked image modeling framework for self-supervised learning, is thought to outperform pretrained ImageNet models and other pretrained contrastive learning models in learning histological patch-level image representations [19]. Nonetheless, these methods did not yield encouraging results (supplementary material, Table S2). We report the only successful attempt at augmentation, which was the oversampling of the positive class in the training data, with HER2 3+ cases. While these additional image cases were never tested for *HER2* amplification, we anticipated that the model might benefit from learning more feature representations from this subset of images.

External validation results

When applied to an external dataset, the IHC scoring model achieved an ROC AUC of 0.98 (SD ± 0.01) and an overall accuracy of 0.91 (SD ± 0.01). The external dataset yielded a sensitivity of 0.50 and a specificity of 0.82. The FISH score prediction model achieved an overall accuracy of 0.85 (SD ± 0.01) and an ROC AUC of 0.75 (SD ± 0.03). This overall accuracy represents the best performance of all variations of the models that were trained. Figure 6 shows the confusion matrices and ROC AUC values for the external validation dataset. Table 3 contains the accuracy metrics for the two main model variants that the external validation set was tested on.

To benchmark our model's performance against widely used digital software, we compared its HER2 IHC scoring capabilities to QuPath, Visiopharm, Halo Pixel, Visiopharm Pixel, and Average threshold. While these tools offer cell-based quantification, they are limited by their reliance on predefined segmentation rules and thresholding algorithms. Our model can leverage global slide-level features without requiring manual region-of-interest selection. In our experiments, the DL achieved a higher overall accuracy (91%) than the traditional digital pathology methods tested (highest overall accuracy of traditional models being 52% for Halo). See supplementary material, Figures S4–S8 for visualization. See supplementary material, Table S3 for overall accuracies only.

Discussion

The findings from this study highlight the efficacy of the CLAM model in accurately predicting HER2 scores from IHC images. Achieving a ROC AUC of 0.98 in HER2 score prediction and an overall accuracy of 95% shows the potential for DL to enhance diagnostic accuracy in breast cancer treatment. Our FISH score prediction model achieved an ROC AUC of 0.84 and overall accuracy of 86%. While this model is considerably less accurate than the base HER2 score prediction model, when deployed in tandem, we believe our approach offers a way to significantly streamline the HER2 testing process by reducing reliance on expensive and time-consuming FISH tests. This may be particularly beneficial in resource-limited settings where access to FISH testing may not be readily available.

Earlier efforts in this domain have been mostly focused on providing HER2 score predictions from IHC, H&E, and FISH images, with varying amounts of success [20,21]. These have typically relied on tissue-level annotations for a supervised DL model approach [14]. Here, we have attempted to create a two-in-one tool that can be used both as a base test

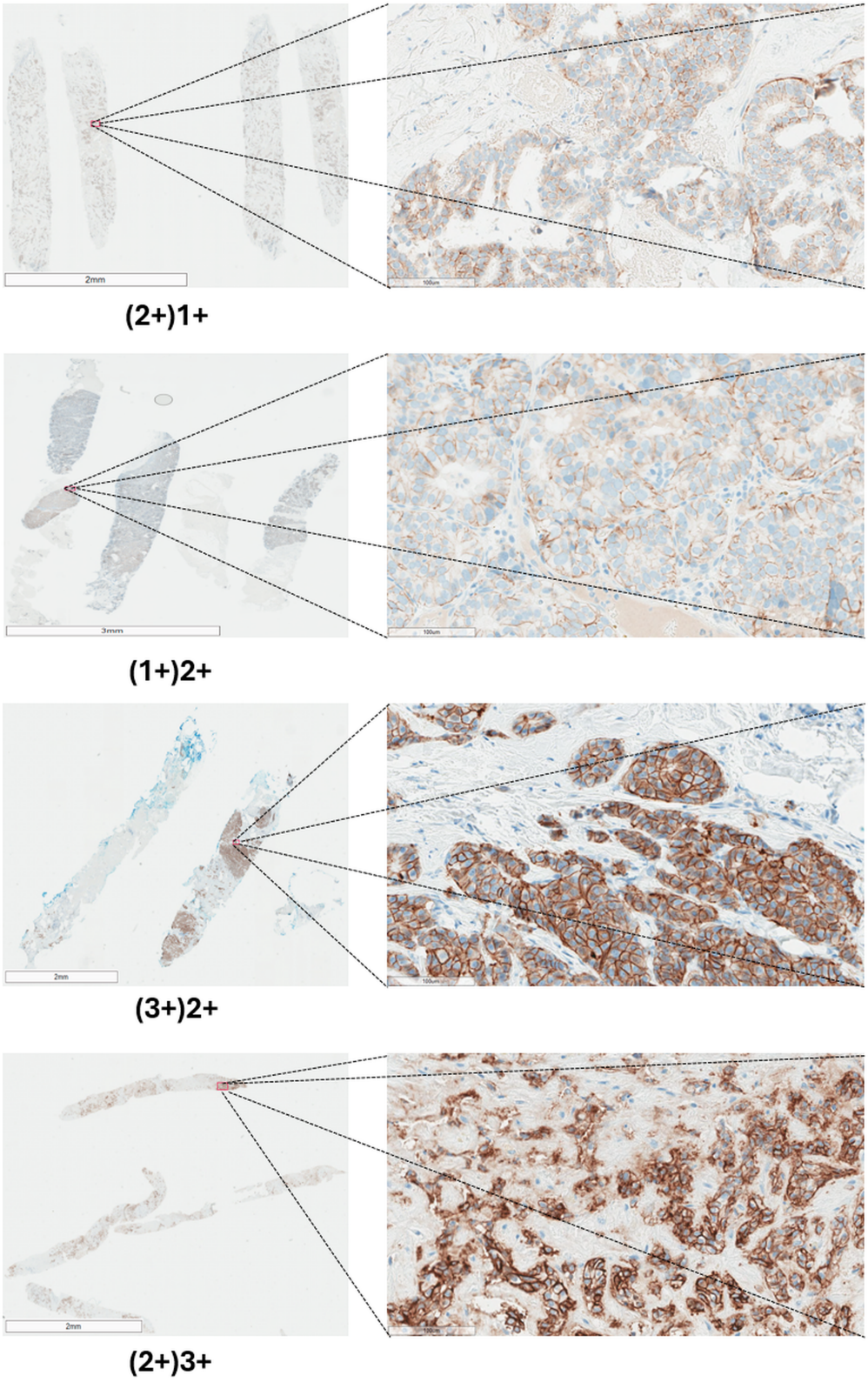

Figure 4. The left column shows slide overviews of incorrectly classified WSIs using the IHC scoring model. The right column shows the regions of interest identified by the model for incorrectly classified WSIs. The correct classes are given in parentheses.

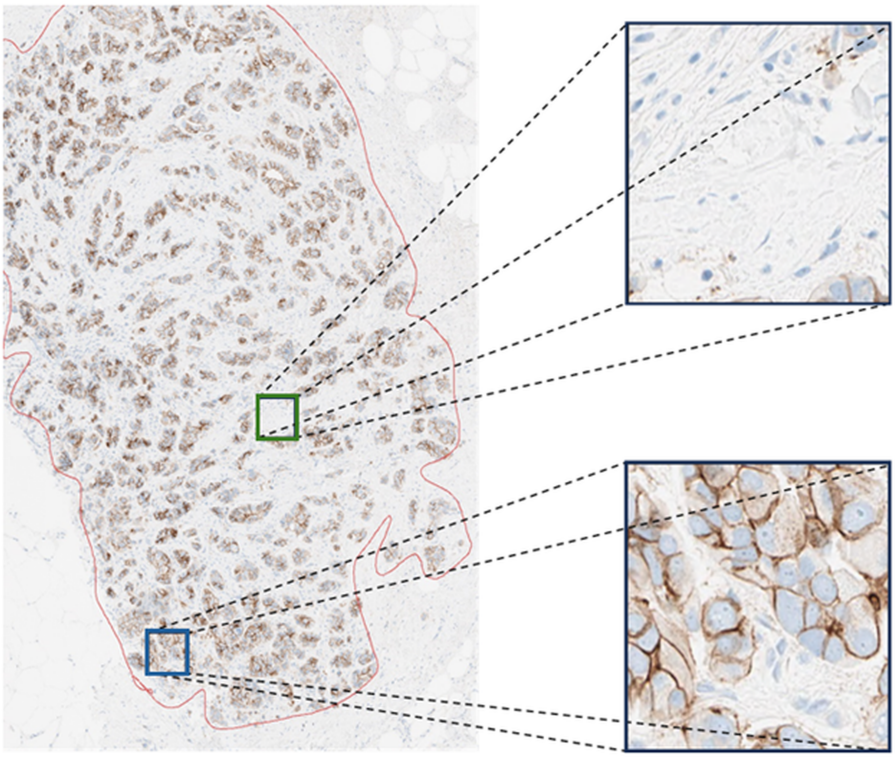

Figure 5. Attention maps for the CLAM model. The least attended patch is bounded by the green box, and the most attended area is bounded by the blue box. The annotated region is in red. Regions with the highest attention exhibit a greater concentration of membrane staining compared to the least attended regions. Though the attention weights correlate to tile-level HER2 positivity, they do so implicitly as a result of training on slide-level HER2 labels.

predictor and as a reflex test predictor. We have also employed a weakly supervised DL approach that can make predictions using only slide-level labels. This method should, in theory, bypass the need for a pathologist to re-annotate a ROI on a slide that has been ruled as equivocal (2+) in base HER2 testing [22]. If the model is sufficiently accurate in its prediction, this would reduce the need for FISH testing. A reliable prediction would also address the issue of intra- and interpathologist variability in the interpretation of IHC images by offering a more standardized and less subjective alternative [17]. One ideal integration approach for this model would be as a decision-support tool for pathologists. Rather than replacing pathologist-driven IHC interpretation, the model could assist in standardizing equivocal (2+) cases and identifying those that are most likely to require FISH confirmation.

FISH testing is known for its high per-test cost due to expensive reagents, consumables, and specialized labor, with turnaround times ranging from several days to over a week. In contrast, AI model implementation costs depend largely on an institution's existing infrastructure. Many large hospitals already possess slide-scanning capabilities, making AI adoption relatively inexpensive, with operational costs as low as cents to a few US dollars per slide, primarily for computing resources and software maintenance. However, institutions without digital pathology infrastructure face significant upfront costs, typically ranging from $50,000 to $300,000. Once implemented, AI models offer a cost-effective alternative, reducing reliance on consumables and specialized personnel while providing rapid predictions within minutes of slide digitization. While not intended to replace FISH testing entirely, the model could prioritize reflex FISH use in equivocal cases, reducing overall test volume, costs, and turnaround times at a systemic level.

Unlike many ML studies in digital pathology that are constrained by the small size of the training data, our study benefits from a large dataset that boosts the effectiveness of our model. The benefits of a large training sample size in the computational pathology domain are well studied and documented [22–24]. Our

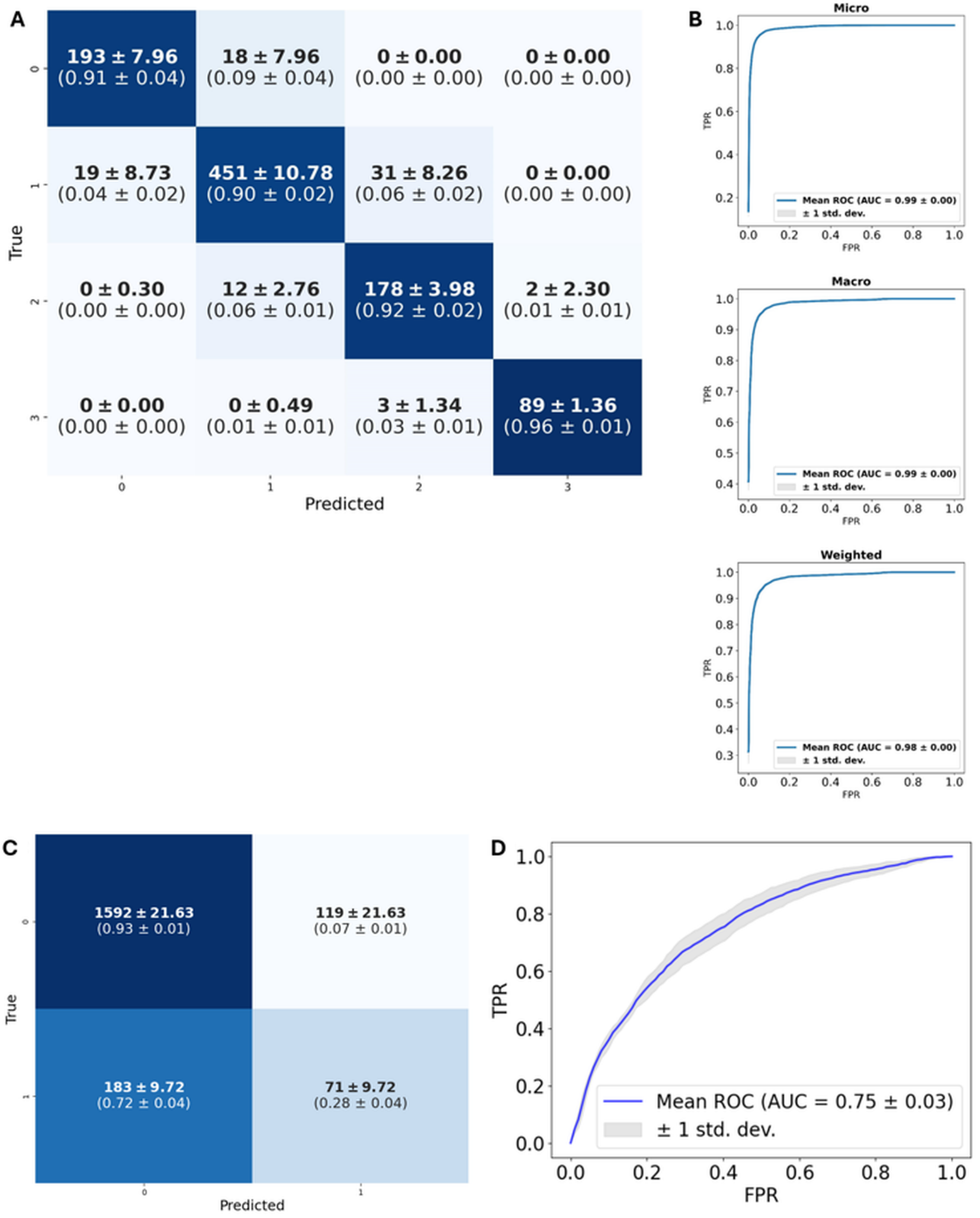

Figure 6. Legend on next page.

Table 3. Comparing the model performance of three experimental variants on external validation data

	Sensitivity (SD)	Specificity (SD)	ROC AUC (SD)
Adding extra 3+ images	0.28 (±0.04)	0.92 (±0.01)	0.75 (±0.03)
Baseline FISH-scored IHC images	0.38 (±0.10)	0.88 (±0.04)	0.73 (±0.02)
Comparison paper	0.32 (±0.11)	0.89 (±0.07)	0.69 (±0.03)

external dataset was obtained from tissue samples derived from a large variety of outside institutions. However, these external samples were mostly processed, tested, and scanned at the Mayo Clinic. One of the key challenges in predicting FISH positivity from IHC images is that the ground truth label/markers are not directly visible in the IHC stain. Traditional DL applications typically rely on training models on readily apparent features in H&E images. To explore whether any discernible features correlate with FISH positivity, a pathologist conducted a review of the most attended regions in cases predicted as FISH-positive or FISH-negative. However, no consistent histological patterns were identified that could reliably distinguish these two categories, suggesting that DL models may be leveraging subvisual cues not readily apparent to human experts. Further research is needed to determine whether certain IHC staining characteristics could inform new HER2 classification criteria or justify additional testing in select cases.

A few notable examples of AI tools that have been adopted by the digital pathology community include QuPath (open source) and Visiopharm (widely used commercial software). These are also known for their relevance in HER2 scoring tasks. We implemented HER2 scoring using both tools to offer a comparison with readily available HER2 scoring tools. The results suggest that our system is at least noninferior to these tools, with superior performance in key classification metrics against every implementation iteration of these AI tools.

Despite its strengths, the study acknowledges certain limitations. The sensitivity of the FISH prediction model was relatively low, indicating difficulties in identifying all positive test cases. This could lead to under-treatment if used in isolation. However, by adjusting the model cutoff for positive prediction from 0.5 to a threshold favoring very high specificity, and thus a high negative predictive value (NPV), we believe the FISH model could be useful in helping to rule out negative amplification status predictions for reflex FISH testing. This justification is supported by the fact that the prevalence of positive *HER2* amplification results upon reflex FISH testing is only about 10%. With such a small percentage of cases likely to test positive and a specificity of 96%, the NPV is 93.2%. In a hypothetical situation where the FISH prediction model is tuned to correctly predict all negative cases in the validation set (i.e., a 100% NPV), when those tuning parameters are then applied to the test set, we attain an NPV of 98% and specificity of 39%. As an example, when 100 FISH tests are to be conducted, only about 10 are expected to be positive based on the prevalence. With a specificity of 39%, we can expect 35 cases to be ruled out with high confidence by our model. Another key limitation is the need for a detailed exploration of uniquely challenging case series, including special histological subtypes, DCIS-rich carcinoma, and metastatic lesions to the breast. Although these cases were included in the overall dataset, they represented less than 2% of the total cases, and none were specifically represented in the test set.

While IHC-based HER2 testing has known limitations, H&E-based DL models remain inadequate for clinical implementation. The HEROHE Challenge demonstrated only moderate success in predicting HER2 scores from H&E –stained images [25]. Tavolara *et al* showed that models trained on HER2 IHC images performed significantly better than models trained on H&E images alone [26]. Larger and more diverse training datasets, along with rigorous validation, are needed to improve the robustness of H&E-based approaches. Despite their current limitations, such models may eventually complement IHC testing, particularly in cases with high interobserver variability. Future studies will focus on improving FISH scoring accuracy, particularly for reliably classifying FISH-positive cases. Furthermore, enriching the dataset with rare histological subtypes and refining image augmentation techniques may enhance performance. Expanding the patient demographic representation could further improve model generalizability.

Figure 6. Results on the external validation set. Error bars represent variation across folds of the original dataset. (A) Normalized confusion matrix showing model performance on external validation samples. (B) Multi-class ROC AUC curves showing model performance on external validation data. (C) Confusion matrix showing model performance for FISH score prediction on external validation samples. Values in parentheses are row-normalized. (D) ROC AUC for the best performing model for FISH score prediction on external validation data.

Our DL model has demonstrated potential in predicting HER2 FISH amplification from IHC images. By integrating AI into the HER2 testing workflow, this approach may help refine the decision-making process, particularly for equivocal (2+) cases where reflex testing is required. While not intended to replace FISH testing, the model could serve as a complementary tool to assist pathologists in identifying cases most likely to benefit from further analysis. In settings where FISH testing is less accessible or associated with long turnaround times, AI-based predictions may provide an additional layer of diagnostic support. The ability to standardize HER2 assessment and improve the consistency of scoring could contribute to more reliable biomarker evaluation. As we continue to refine our model, we are optimistic about its role in shaping future diagnostic processes and making precision medicine more accessible across diverse clinical settings.

Author contributions statement

DOM contributed to data analysis, data curation, methodology, software, validation, writing the original draft and editing the manuscript. WH contributed to formal analysis, and review and editing of the manuscript. MDZ contributed to conceptualization, validation, and review and editing of the manuscript. CAG contributed to validation, methodology, review and editing of the manuscript. TET contributed to data collection, data curation, formal analysis, conceptualization, project administration and review and editing of the manuscript.

Data availability statement

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request. Due to ethical and privacy considerations, certain data elements may be restricted in accordance with institutional and regulatory guidelines. For further inquiries regarding data access and sharing, please contact Daniel Macaulay at macaulay.daniel@mayo.edu.

References

1. Wolff AC, Somerfield MR, Dowsett M, *et al*. Human epidermal growth factor receptor 2 testing in breast cancer. *Arch Pathol Lab Med* 2023; **147**: 993–1000.
2. Tarantino P, Morganti S, Curigliano G. Targeting HER2 in breast cancer: new drugs and paradigms on the horizon. *Explor Target Anti-Tumor Ther* 2021; **2**: 139–155.
3. Modi S, Jacot W, Yamashita T, *et al*. Trastuzumab deruxtecan in previously treated HER2-low advanced breast cancer. *N Engl J Med* 2022; **387**: 9–20.
4. Nicolò E, Boscolo Bielo L, Curigliano G, *et al*. The HER2-low revolution in breast oncology: steps forward and emerging challenges. *Ther Adv Med Oncol* 2023; **15**: 175883592311528.
5. Modi S, Park H, Murthy RK, *et al*. Antitumor activity and safety of trastuzumab deruxtecan in patients with HER2-low–expressing advanced breast cancer: results from a phase Ib study. *J Clin Oncol* 2020; **38**: 1887–1896.
6. Kaufman PA, Bloom KJ, Burris H, *et al*. Assessing the discordance rate between local and central HER2 testing in women with locally determined HER2-negative breast cancer. *Cancer* 2014; **120**: 2657–2664.
7. Lambein K, Van Bockstal M, Vandemaele L, *et al*. Distinguishing score 0 from score 1+ in HER2 immunohistochemistry-negative breast cancer. *Am J Clin Pathol* 2013; **140**: 561–566.
8. Rossing HH, Talman ML, Laenkholm AV, *et al*. Implementation of TMA and digitalization in routine diagnostics of breast pathology. *APMIS* 2012; **120**: 341–347.
9. Verghese G, Lennerz JK, Ruta D, *et al*. Computational pathology in cancer diagnosis, prognosis, and prediction – present day and prospects. *J Pathol* 2023; **260**: 551–563.
10. Echle A, Rindtorff NT, Brinker TJ, *et al*. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer* 2021; **124**: 686–696.
11. Jaber MI, Song B, Taylor C, *et al*. A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. *Breast Cancer Res* 2020; **22**: 12.
12. Brevet M, Li Z, Parwani A. Computational pathology in the identification of HER2-low breast cancer: opportunities and challenges. *J Pathol Inform* 2024; **15**: 100343.
13. Bannier P-A, Herpin L, Dubois R, *et al*. Abstract PO2-07-05: deep learning model for automated quantification of HER2 expression in invasive breast cancers from immunohistochemical whole slide images. *Cancer Res* 2024; **84**: PO2-07-05-PO02-00.
14. Rasmussen SA, Taylor VJ, Surette AP, *et al*. Using deep learning to predict final HER2 status in invasive breast cancers that are equivocal (2+) by immunohistochemistry. *Appl Immunohistochem Mol Morphol* 2022; **30**: 668–673.
15. Lu MY, Williamson DFK, Chen TY, *et al*. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng* 2021; **5**: 555–570.
16. Holten-Rossing H, Møller Talman M-L, Kristensson M, *et al*. Optimizing HER2 assessment in breast cancer: application of automated image analysis. *Breast Cancer Res Treat* 2015; **152**: 367–375.
17. 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**: 100009.
18. Robbins CJ, Fernandez AI, Han G, *et al*. Multi-institutional assessment of pathologist scoring HER2 immunohistochemistry. *Mod Pathol* 2023; **36**: 100032.

19. Filiot A, Ghermi R, Olivier A, *et al*. Scaling self-supervised learning for histopathology with masked image modeling. medRxiv 2023. https://doi.org/10.1101/2023.07.21.23292757.
20. Oliveira SP, Ribeiro Pinto J, Gonçalves T, *et al*. Weakly-supervised classification of HER2 expression in breast cancer haematoxylin and eosin stained slides. *Appl Sci* 2020; **10**: 4728.
21. Xue T, Chang H, Ren M, *et al*. Deep learning to automatically evaluate HER2 gene amplification status from fluorescence in situ hybridization images. *Sci Rep* 2023; **13**: 9746.
22. Campanella G, Hanna MG, Geneslaw L, *et al*. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019; **25**: 1301–1309.
23. Chen RJ, Ding T, Lu MY, *et al*. Towards a general-purpose foundation model for computational pathology. *Nat Med* 2024; **30**: 850–862.
24. Fuchs TJ, Buhmann JM. Computational pathology: challenges and promises for tissue analysis. *Comput Med Imaging Graph* 2011; **35**: 515–530.
25. Conde-Sousa E, Vale J, Feng M, *et al*. HEROHE challenge: predicting HER2 status in breast cancer from hematoxylin-eosin whole-slide imaging. *J Imaging* 2022; **8**: 213.
26. Tavolara TE, Niazi MKK, Tozbikian G, *et al*. Predicting HER2 scores from registered HER2 and H&E images. In *Proceedings of SPIE 12039 Medical Imaging 2022: Digital and Computational Pathology*, 2022; 120390C. https://doi.org/10.1117/12.2612878.

SUPPLEMENTARY MATERIAL ONLINE

Figure S1. FISH Score prediction model in institutional FISH score data

Figure S2. FISH Score prediction model in external FISH score data

Figure S3. Sankey diagram showing analysis of AI-misclassified HER2 IHC scores for diagnostically relevant misinterpretations

Figure S4. QuPath results confusion matrix, showing per class classification accuracy for HER2 score classification

Figure S5. Halo results confusion matrix, showing per class classification accuracy for HER2 score classification

Figure S6. Average threshold method (ATM) results confusion matrix, showing per class classification accuracy for HER2 score classification

Figure S7. Visiopharm pixel method results confusion matrix, showing per class classification accuracy for HER2 score classification

Figure S8. Halo pix results pixel method results confusion matrix, showing per class classification accuracy for HER2 score classification

Table S1. Preliminary experiments examining the effect of varying training set size for FISH prediction from HER2 IHC images

Table S2. Experiments with imaging encoders pretrained using self-supervised learning on HER2 IHC image patches

Table S3. Comparison of HER2 scoring methods