Original Research Article

Hybrid Supervised-Unsupervised CycleGAN for Virtual HER2 Immunohistochemistry from Hematoxylin and Eosin Stains

Aneesh Dantuluri

Branham High School, 1570 Branham Ln, San Jose, CA 95118, United States

ABSTRACT

HER2-status is a vital biomarker for breast cancer diagnosis and treatment, typically assessed using immunohistochemistry (IHC), a technique that is expensive and demanding of laboratory experience. Hematoxylin and eosin (H&E) staining, in contrast, is widely available and inexpensive, motivating approaches that can computationally translate H&E images into IHC. While previous work has explored translating H&E stains of breast tissue into IHC using purely unsupervised methods, this study introduces a hybrid CycleGAN framework that combines unsupervised cycle-consistency with supervised paired reconstruction objectives. By leveraging the paired structure of the BCI dataset, this approach significantly improves quantitative metrics (PSNR: 16.203 → 17.807 (Adam); SSIM: 0.373 → 0.4061 (AdamW) and visual fidelity compared to unsupervised-only baselines, narrowing the performance gap with supervised-only architectures while maintaining CycleGAN's flexibility. These findings show that incorporating limited supervision into cycle-consistent adversarial training enhances H&E-to-IHC translation quality, offering a more affordable and accessible pathway to HER2 screening.

Keywords: CycleGAN; histopathology; HER2; Generative-AI; image-to-image translation

INTRODUCTION

Breast cancer is a leading cause of mortality among women worldwide, impacting over 2.1 million women annually (1). One critical factor in treatment planning is assessing HER2 protein overexpression, which occurs in roughly 20% of breast cancers (2). HER2-positive tumors can be effectively treated with Herceptin (Trastuzumab), a monoclonal antibody that targets the HER2 protein (3, 4). This makes accurate HER2 status determination essential for guiding therapeutic decisions.

Immunohistochemistry (IHC) is one of the primary methods for HER2 assessment (5, 6). It uses specific antibodies to visualize and quantify HER2 expression in tumor tissue samples (7). But IHC is expensive, technically demanding, and requires specialized laboratory equipment and expertise (8). IHC scoring can also be subjective, leading to inter-laboratory variability (9). Given these limitations and recent advances in computer vision, researchers have investigated whether inexpensive H&E staining—a routine histopathological technique—could be computationally translated into IHC images using machine learning methods.

Liu *et al*. (10) collected a dataset of paired (H&E, IHC) images from adjacent breast tumor tissue sections and evaluated various image-to-image translation models for H&E-to-IHC conversion. They assessed CycleGAN as an unsupervised baseline (disregarding

Corresponding author: Aneesh Dantuluri, E-mail: aneesh72583@gmail.com.

Accepted October 23, 2025
https://doi.org/10.70251/HYJR2348.365360

pairing information) and proposed a supervised pyramid pix2pix architecture that outperformed CycleGAN. However, they didn't evaluate whether CycleGAN could benefit from incorporating both its original unsupervised cycle-consistency objective and a supervised pairing loss when paired data are available.

This study addresses that gap by training CycleGAN with a hybrid objective that combines unsupervised cycle-consistency with supervised paired reconstruction. The results show this hybrid approach improves both quantitative metrics and visual quality over the unsupervised-only baseline, striking a better balance between the flexibility of unpaired methods and the accuracy gains of supervision.

METHODS AND MATERIALS

BCI Dataset

The BCI dataset (10) consists of 4870 paired (H&E, IHC) image patches obtained by differential staining of adjacent breast tumor sections. While these sections are anatomically similar, the tissue sectioning process introduces small spatial translations and rotations, the images aren't perfectly aligned. This partial misalignment is actually what motivates using cycle-consistency objectives alongside supervised losses.

Before model training, I performed a preprocessing step to identify and remove corrupted or low-quality images. Images with pixel-wise variance below predefined thresholds (variance < 29 for H&E images, variance < 10 for IHC images) were flagged as corrupted and removed from both domains to maintain pairing integrity. This quality control step removed 185 corrupt image pairs from the training set.

Following preprocessing, the filtered dataset was organized into domain-specific directories. H&E images were designated as domain A (trainA, testA) and IHC images as domain B (trainB, testB) following standard CycleGAN conventions. The dataset was split into training (80%, n=3896 after filtering) and test (20%, n=974 after filtering) sets. A separate validation split wasn't explicitly created; instead, model checkpoints were saved at regular intervals and the checkpoint with the best visual quality on held-out test samples was selected for final evaluation.

CycleGAN Architecture

CycleGAN (11) is a method for learning image-to-image translation between two domains, A and B, without needing perfectly aligned paired data. The architecture uses two generators G_AB: A → B and G_BA: B → A, and two discriminators D_A: A → [0,1] and D_B: B → [0,1].

For this study, the generator architecture used ResNet-based networks with 9 residual blocks (resnet_9blocks), configured with 64 base filters (ngf=64). The discriminators used the basic PatchGAN architecture with 3 convolutional layers (n_layers_D=3) and 64 base filters (ndf=64). Instance normalization was applied throughout the networks, and dropout was disabled (no_dropout=True) as is standard for CycleGAN training. The original CycleGAN objective has two main components. The cycle-consistency loss makes sure that translating from domain A to B and back to A reconstructs the original image:

$$L_{cyc}(G_{AB}, G_{BA}, a, b) = \|G_{BA}(G_{AB}(a)) - a\|_1 + \|G_{AB}(G_{BA}(b)) - b\|_1$$

where $a \sim A$ and $b \sim B$. The cycle-consistency loss was weighted with $\lambda_A = \lambda_B = 10.0$ to balance reconstruction fidelity with adversarial objectives. Adversarial discriminator losses push generated images to resemble real samples from the target domain. The study used least-squares GAN (LSGAN) objectives rather than vanilla GAN losses for better training stability:

$$L_D(G_{AB}, G_{BA}, a, b) = L_{D_A} + L_{D_B}$$
$$= -\log D_A(1 - G_{BA}(b)) - \log D_B(1 - G_{AB}(a))$$

Additionally, an identity mapping loss with weight $\lambda_{identity} = 0.5$ was included to encourage color consistency, a standard CycleGAN practice for tasks involving appearance transfer rather than geometric transformation. The full unsupervised generator loss is:

$$L_G = L_{cyc} + L_D + L_{identity}$$

Proposed Hybrid Supervised-Unsupervised Objective

Since the BCI dataset provides paired (H&E, IHC) samples, this study introduces a supervised reconstruction loss that leverages the pairing information. Let a denote an H&E image and b denote its corresponding IHC image from the paired distribution. The supervised loss is:

$$L_{sup}(G_{AB}, G_{BA}, a, b) = \|G_{BA}(b) - a\|_2^2 + \|G_{AB}(a) - b\|_2^2$$

This loss directly penalizes deviations between generated and ground-truth paired images. The choice of L2 (mean squared error) over L1 for the supervised loss comes from wanting to penalize large errors more heavily, which improved visual fidelity in preliminary experiments. L1 is kept in the cycle-consistency loss to maintain robustness to the misalignment noise that's inherent in the dataset. The full hybrid training objective becomes:

$$L'_G = L_G + \beta L_{sup}$$

where β is a scalar hyperparameter that balances the contribution of supervised and unsupervised objectives. The supervised loss was implemented by adding a command-line flag (supervised) to the training script, which makes switching between unsupervised and hybrid training modes seamless (Figure 1).

Training Configuration

All the models were trained using the pytorch-CycleGAN implementation using custom modifications which support supervised loss integration. Model training was done on Google Colab with GPU acceleration (specific GPU allocation varied but typically consisted of Tesla T4 or P100 instances with 16GB VRAM).

Typical training time was around 6-8 hours per model for 100 epochs on Colab's allocated GPUs, though the exact timing varied based on GPU availability. The model checkpoints were periodically backed up to a Google Drive to prevent data loss during long training runs.

The supervised loss weight was set to β=2 based on initial testing (a grid search) with values of 0.5, 1, 2, and 5. While β=2 resulted in the best qualitative results on validation samples, a more detailed study will have to wait for future work. Detailed training parameters and configurations are documented in Table 1.

Evaluation Metrics

Model performance was measured using two standard image similarity metrics computed on the held-out test set:

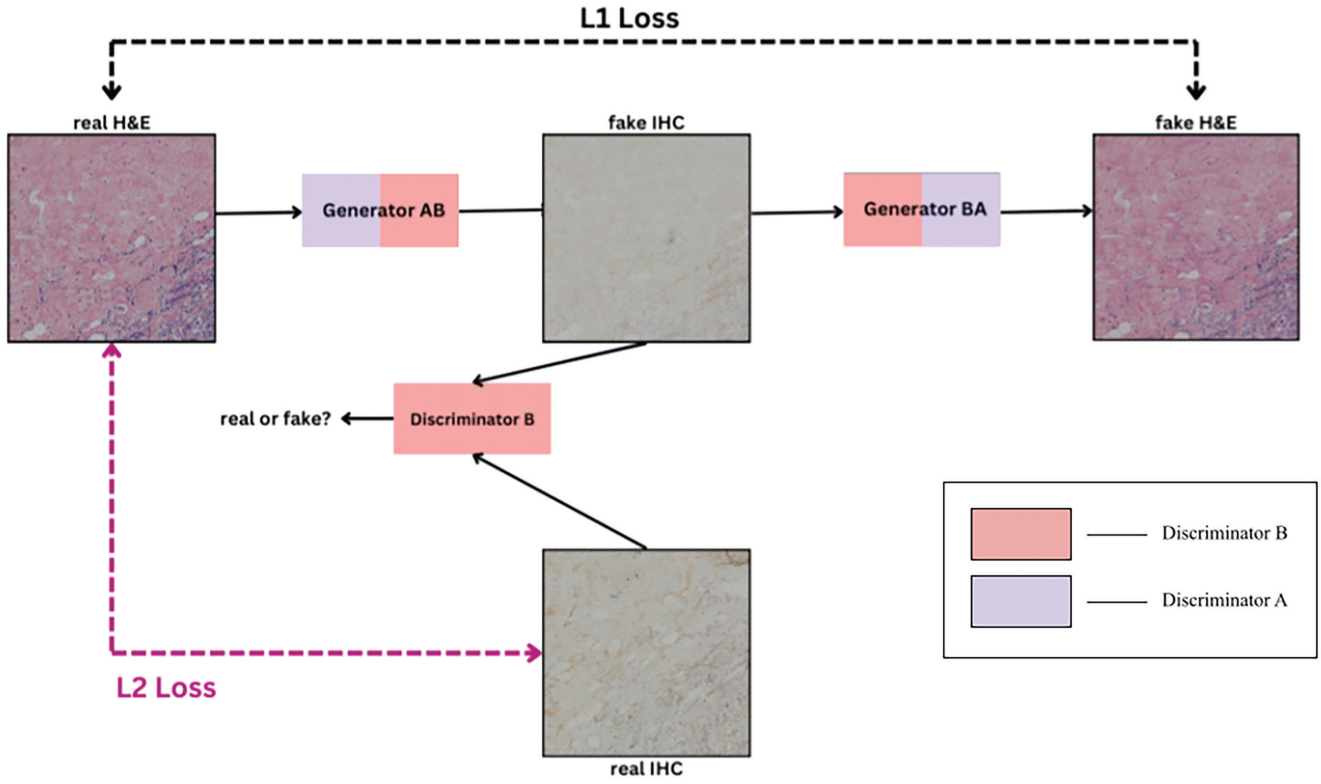

Figure 1. A schematic of the CycleGAN supervised training objective. Only one of the terms of L_{cyc} is shown. In addition, since we have paired data. we include a supervised L2 loss between real H&E and IHC pair.

Table 1. Training hyperparameters and configuration settings for supervised-unsupervised CycleGAN model, including optimizer specifications, learning rate schedules, loss weights, and infrastructure

Parameter	Setting/Description
Framework	PyTorch CycleGAN implementation with custom modifications for supervised loss integration
Hardware	Google Colab (Pro) GPU acceleration (Tesla T4 or P100, 16 GB VRAM)
Batch Size	1 (standard for CycleGAN to maximize image resolution within memory limits)
Input/Output Channels	3 (RGB images)
Image Preprocessing	Resize to 286 × 286 (load_size = 286), random crop to 256 × 256 (crop_size = 256), random horizontal flip for augmentation
Optimizer	Adam and AdamW ($\beta_1 = 0.5$, $\beta_2 = 0.999$)
Initial Learning Rate	0.0002 (for both generators and discriminators)
Learning Rate Schedule	Linear decay (lr_policy = 'linear') starting after 50 epochs (lr_decay_iters = 50)
Training Duration	Unsupervised model: 200 epochs (100 + 100 decay); Hybrid models: Adam stopped at epoch 89, AdamW at epoch 62 (based on validation monitoring)
GAN Objective	Least-Squares GAN (gan_mode = 'lsgan')
Cycle-Consistency Weights	$\lambda A = \lambda B = 10.0$
Identity Loss Weight	$\lambda identity = 0.5$
Supervised Loss Weight (β)	2.0 (enabled only for hybrid models via --supervised flag)
Image Pool Size	50 (pool_size = 50)
Checkpoint Frequency	Save every epoch (save_epoch_freq = 1); latest checkpoints every 5000 iterations (save_latest_freq = 5000)
Training Time	≈ 6–8 hours per model for 100 epochs (depending on GPU availability)
Backup Procedure	Model checkpoints backed up periodically to Google Drive to prevent data loss
Supervised Loss Tuning	$\beta \in \{0.5, 1, 2, 5\}$ tested; $\beta = 2$ yielded best qualitative results on validation set

Peak Signal-to-Noise Ratio (PSNR): Measures pixel-level reconstruction accuracy, directly related to mean squared error. Higher values mean closer match to ground truth. PSNR is calculated using the formula: $PSNR = 10 \cdot \log_{10}(MAX^2/MSE)$, where MAX is the maximum possible pixel value (255 for 8-bit images) and MSE is the mean squared error between predicted and ground-truth images.

Structural Similarity Index (SSIM): Evaluates perceptual similarity by comparing luminance, contrast, and structure. SSIM values range from 0 to 1, with higher values meaning greater visual fidelity. SSIM is sensitive to structural distortions and is more similar to human perception than PSNR.

Both metrics were computed on the test set using paired ground-truth IHC images as reference. Training PSNR was also monitored on a subset of training samples during model development to check for overfitting and convergence behavior, though formal training metrics aren't reported in the final results table.

RESULTS AND DISCUSSION

Quantitative Performance

Table 2 summarizes the full quantitative results. For both the AdamW and Adam optimizer, the supervised loss substantially increased performance compared to the unsupervised-only baseline. The unsupervised CycleGAN reached PSNR = 16.203 and SSIM = 0.373. With the hybrid objective, the Adam optimizer reached PSNR = 17.807 and SSIM = 0.3906, while AdamW pushed to PSNR = 17.527 and SSIM = 0.4061.

The difference, though it seems minimal at first, actually represents notable improvements in structural fidelity and color accuracy, as you can see from visual inspection (Figure 2). The SSIM jump from 0.373 to

Table 2. Performance metrics comparison of unsupervised baseline CycleGAN and hybrid supervised-unsupervised CycleGAN models trained with Adam and AdamW optimizers. Higher values indicate better performance for all metrics

Loss	Optimizer	β	Batch Size	Epoch	Training PSNR	PSNR	SSIM
Unsupervised (baseline)	Adam	0	2	100	-	16.203	0.373
Unsupervised + Supervised	Adam	2	2	89	21.414	17.807	0.3906
Unsupervised + Supervised	AdamW	2	2	62	21.457	17.527	0.4061

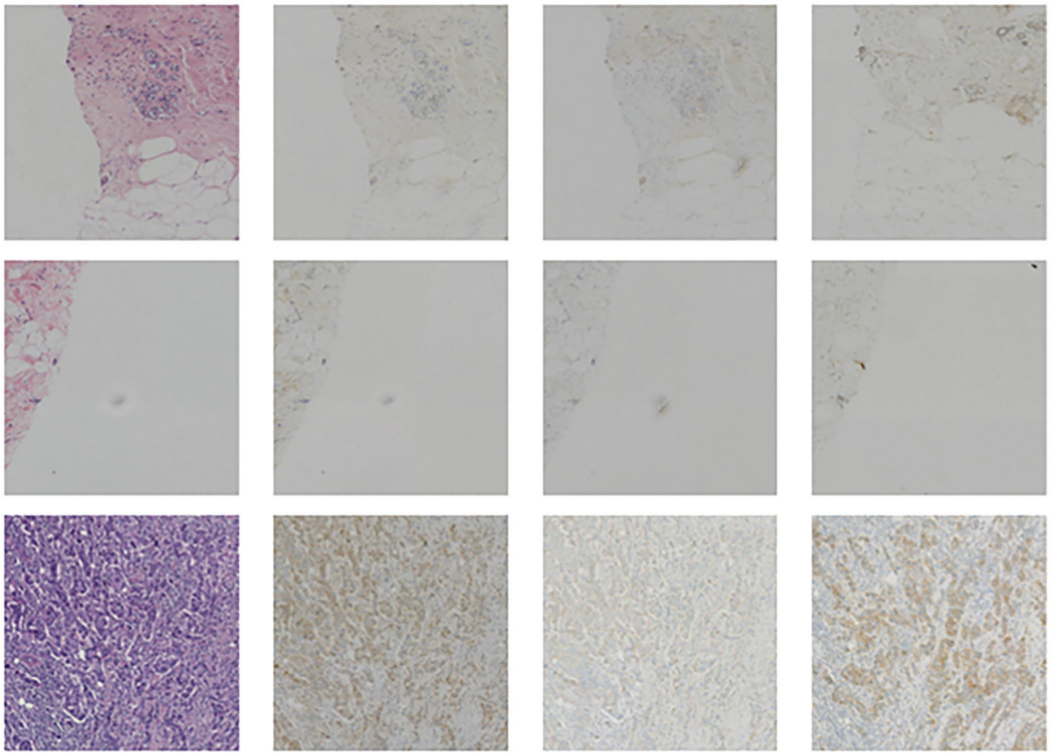

Figure 2. Visual comparison of H&E to IHC translation results showing real H&E input images (leftmost column), IHC images generated by hybrid CycleGAN with Adam optimizer (second column), IHC images generated by hybrid CycleGAN with AdamW optimizer (third column), and ground-truth IHC images (rightmost column). The hybrid AdamW model demonstrates improved color accuracy compared to the hybrid Adam model, particularly in regions with strong HER2 staining (brown coloration).

0.4061 (+8.8%) could translate to better interpretability for diagnostic tasks, however, professional pathologist review is needed to confirm true clinical usability.

The training PSNR values (21.414 for Adam, 21.457 for AdamW) are quite a bit higher than the test PSNR, which signals some degree of overfitting. This gap likely comes from the partial misalignment between paired sections, which causes noise that supervised losses try to fit during training but can't fully generalize in testing. The unsupervised baseline overcomes this issue by ignoring pairing altogether, which results in blurrier but more generalizable outputs.

The superior SSIM performance of the AdamW optimizer (0.4061 vs 0.3906 for Adam), despite slightly lower test PSNR (17.527 vs 17.807), conveys that AdamW's weight decay regularization helps preserve structure better than Adam, even if pixel-wise accuracy takes a minimal hit. This is understandable given that AdamW is known to be effective at preventing overfitting and keeping generalization strong.

Qualitative Assessment

Figure 2 shows example translations comparing the unsupervised baseline (Adam), hybrid model (AdamW), and ground-truth IHC. The hybrid AdamW model captures the overall coloration and tissue structures far more accurately than the unsupervised baseline, especially in regions with strong HER2 staining (the dark brown regions in IHC). As expected, the model can't perfectly reproduce fine spatial details because of misalignment between adjacent sections during data collection. This misalignment explains why the unsupervised CycleGAN generates blurrier outputs—it learns to average over spatial uncertainty rather than committing to precise but potentially misaligned reconstructions.

The hybrid model shows noticeably improved ability to capture IHC-specific colorimetric properties, including the characteristic brown DAB (3,3'-diaminobenzidine) staining that indicates HER2 protein presence. The model also does a better job preserving tissue architecture, keeping cellular boundaries and glandular structures intact, features that are absolutely critical for pathological assessment. In regions with weak or heterogeneous HER2 expression, though, the model sometimes produces artifactual staining patterns that aren't in the ground truth, suggesting that further tweaking of the supervised loss weighting or training duration could boost fidelity in these trickier cases.

These observations point to something important: while the hybrid approach improves image fidelity, clinical validation with professional pathologists is needed to figure out whether the improvements actually enhance diagnostic accuracy for HER2 scoring. Metrics like PSNR and SSIM measure image similarity but don't tell us directly about clinical interpretability, reliability, or usability.

Limitations and Future Directions

There are several limitations of this study worth noting. The choice of $\beta=2$ (the supervised loss weight) was made based on a limited grid search across four values (0.5, 1, 2, and 5). A more indepth ablation study that varies β alongside learning rate schedules, cycle-consistency weights, and training duration would provide more insight into CycleGAN's behavior and ideal settings. The different number of training epochs used for the Adam (89 epochs) and AdamW (62 epochs) models suggests that optimizer choice interacts with supervised loss in ways which should be investigated further.

Next, due to the model training being limited to a single dataset of breast tumor sections from a single staining protocol, the model's performance on different tissues, staining protocols, or clinical settings remains unknown. Additionally, the data preprocessing step that removed 185 corrupted images might have introduced selection bias, though this is only about 3.8% of the original dataset and likely had an insignificant impact on overall results.

Another major limitation is the lack of clinical validation. Image similarity metrics like PSNR and SSIM do not equate to diagnostic usefulness, which is what is ultimately valuable for clinical application. Future work needs to examine blinded pathologist readings that compare HER2 scoring agreement (0, 1+, 2+, 3+ scores) between real IHC, generated IHC, and H&E-only interpretation. Interrater reliability metrics like Cohen's kappa or intraclass correlation would help quantify whether generated IHC actually contributes clinically useful insight not available from H&E alone.

While this work is an improvement compared to an unsupervised baseline CycleGAN, comparison directly to the pyramid pix2pix architecture from Liu *et al.* would serve to clarify the position hybrid approaches relative to purely unsupervised and purely supervised methods. From the reported PSNR and SSIM values, the hybrid CycleGAN likely still remains behind pyramid pix2pix performance but perhaps not by as wide a margin as previously described.

The variable training times (62-100 epochs) and reliance on manual checkpoint selection based on visual quality introduces some subjectivity in the analysis. Future work should employ automated early stopping based on quantitative validation metrics and report confidence intervals over multiple independent training runs to establish just how robust these results really are. This would also help with reproducibility concerns and provide a better sense of model variance over different random initializations.

Finally, the hybrid approach in this study can be extended to other immunohistochemical stains like ER (estrogen receptor), PR (progesterone receptor), and Ki-67 to evaluate generalizability beyond HER2 (12). Multitarget translation, where several IHC outputs for different biomarkers are generated from one H&E input, can also enhance clinical utility by reducing the need for multiple separate staining routines. Recent advancements in transformer-based architectures and diffusion models have shown superior performance

in image-to-image translation tasks. Incorporating these modern techniques while maintaining the hybrid supervised-unsupervised architecture may lead to additional performance gains.

Despite these limitations, the results of this study show that incorporating supervised losses in the training objective of CycleGAN notably improves translation quality. This hybrid approach highlights a valuable middle ground between unsupervised flexibility and supervised accuracy and is a promising direction for applications in computational pathology where paired data can be acquired but might be imperfectly aligned.

CONCLUSION

This study introduced a hybrid CycleGAN framework that combines an unsupervised cycle-consistency with supervised paired reconstruction for H&E-to-IHC translation in HER2-positive breast cancer assessment. This hybrid approach substantially improved both performance metrics (PSNR, SSIM) and visual fidelity compared to unsupervised-only baselines, showing that leveraging paired data enhances CycleGAN's performance without sacrificing its architectural simplicity.

While the improvements are promising, clinical translation requires further validation through blinded pathologist evaluation to understand diagnostic utility for HER2 scoring. Future research should explore hyperparameter sensitivity, extend evaluation to diverse datasets and staining protocols, compare performance directly with state-of-the-art supervised architectures, and investigate extension to additional immunohistochemical biomarkers other than HER2.

With more research, hybrid generative approaches hold potential to advance cost-effective, accessible diagnostic tools, broadening the reach of precision oncology by reducing reliance on expensive immunohistochemical testing while maintaining the diagnostic accuracy necessary for treatment planning decisions.

ACKNOWLEDGEMENTS

Thank Roger Gin from MIT for mentoring me through this research process.

FUNDING SOURCES

None/NA.

CONFLICT OF INTERESTS

The author declares that there are no conflicts of interest related to this work.

REFERENCES

1. Misganaw M, Zeleke H, Mulugeta H, Assefa B. Mortality rate and predictors among patients with breast cancer at a referral hospital in northwest Ethiopia: A retrospective follow-up study. *PLoS One*. 2023; 18 (1). https://doi.org/10.1371/journal.pone.0279656
2. Slamon DJ, Clark GM, Wong SG, Levin WJ, *et al*. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*. 1987; 235 (4785): 177-182. https://doi.org/10.1126/science.3798106
3. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, *et al*. Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer. *N. Engl. J. Med.* 2005; 353 (16): 1659-1672. https://doi.org/10.1056/NEJMoa052306
4. Romond EH, Perez EA, Bryant J, Suman VJ, *et al*. Trastuzumab plus Adjuvant Chemotherapy for Operable HER2-Positive Breast Cancer. *N. Engl. J. Med.* 2005; 353 (16): 1673-1684. https://doi.org/10.1056/NEJMoa052122
5. Wolff AC, Hammond MEH, Schwartz JN, Hagerty KL, *et al*. American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for HER2 Testing in Breast Cancer. *J. Clin. Oncol.* 2016.
6. Wolff AC, Hammond MEH, Hicks DG, Dowsett M, *et al*. Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: ASCO/CAP Clinical Practice Guideline Update. *J. Clin. Oncol.* 2013.
7. Press MF, Pike MC, Chazin VR, Hung G, *et al*. Her-2/neu expression in node-negative breast cancer: direct tissue quantitation by computerized image analysis and association of overexpression with increased risk of recurrent disease. *Cancer Res.* 1993; 53 (20): 4960-4970.
8. Bartlett JM, Going JJ, Mallon EA, Watters AD, *et al*. Evaluating HER2 amplification and overexpression in breast cancer. *J. Pathol*. 2001; 195 (4): 422-428. https://doi.org/10.1002/path.971
9. Paik S, Bryant J, Tan-Chiu E, Romond E, *et al*. Real-World Performance of HER2 Testing-National Surgical Adjuvant Breast and Bowel Project Experience. *J. Natl. Cancer Inst.* 2002; 94 (11): 852-

854. https://doi.org/10.1093/jnci/94.11.852
10. Liu S, Zhu C, Xu F, Jia X, *et al*. BCI: Breast Cancer Immunohistochemical Image Generation through Pyramid Pix2pix. arXiv, 2022. https://doi.org/10.1109/CVPRW56347.2022.00198
11. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. arXiv, 2017. https://doi.org/10.1109/ICCV.2017.244
12. Hammond MEH, Hayes DF, Dowsett M, Allred DC, *et al*. American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Immunohistochemical Testing of Estrogen and Progesterone Receptors in Breast Cancer. *J. Clin. Oncol*. 2010. https://doi.org/10.1200/JCO.2009.25.6529