A Comparative Study on HER2 Immunohistochemical scoring performance using Vision Transformer and DenseNet-201

Jongwon Lee[1], Ji Won Hwang[2], GunHee Lee[1], and Hee Jin Lee [1]

[1]Department of Pathology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea;

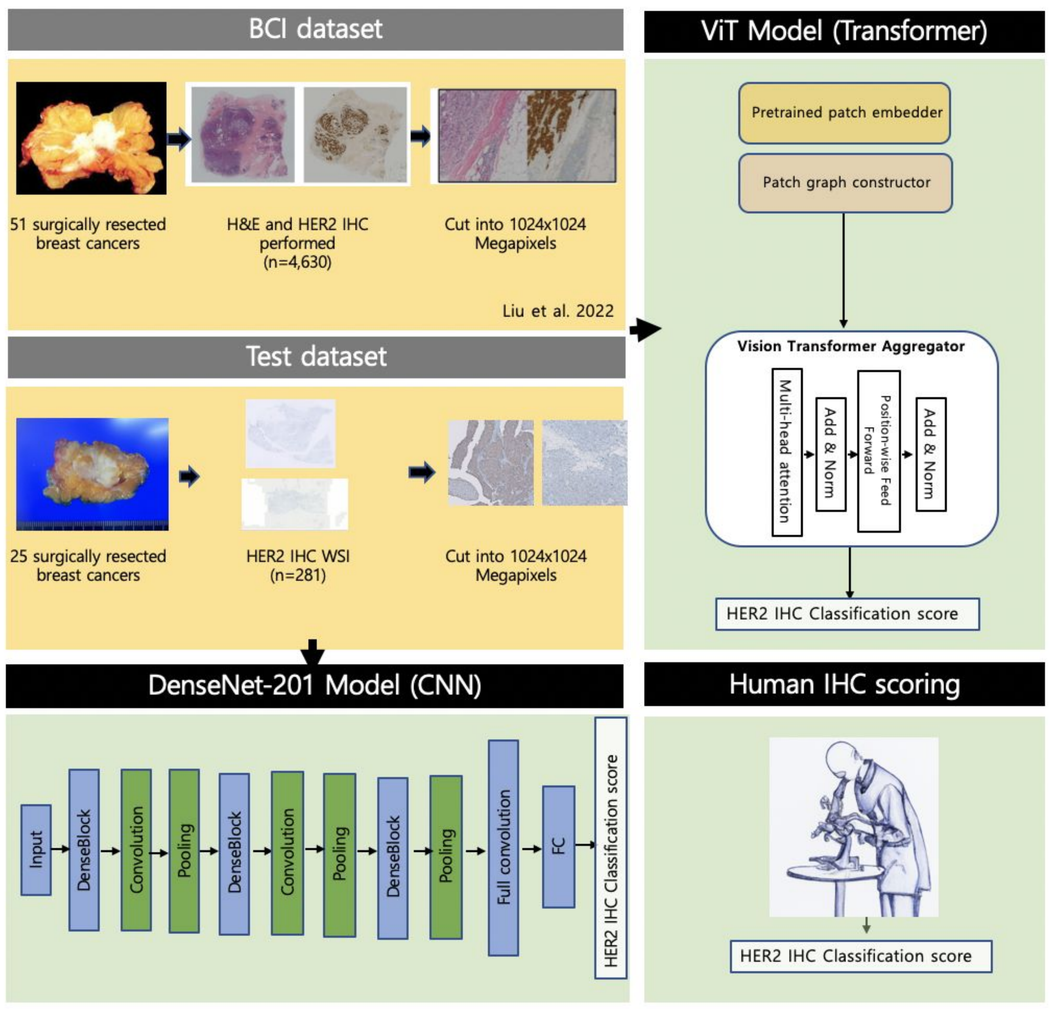
[2]University of Ulsan College of Medicine, Seoul, Republic of Korea

INTRODUCTION

- HER2 immunohistochemistry (IHC) scoring is a critical component in breast cancer. With the advent of machine learning models, particularly convolutional neural networks and transformer architectures there are new opportunities for automated image classification.
- However, the extent to which these models can accurately interpret specific images so as to score them, and their alignment with pathologist interpretations, remains to be fully comprehended.

METHODS

- BCI dataset, provided by Liu et al. (https://github.com/bupt-ai-cz/BCI) was used. The data consisted of 4,630 image patches (1024x1024 pixels) labeled by two pathologists according to ASCO-CAP guidelines, as 1+, 2+, and 3+ (1+, 1,153 patches; 2+, 2,142 patches; 3+, 1,335 patches).
- Vision Transformer (ViT) and DenseNet-201 models were trained with the data using PyTorch and GPU acceleration. Images were resized to 224x224 pixels and normalized. For both models, training utilized an 80/20 data split, a batch size of 32, and Cross-Entropy Loss. Adam optimizer was used for ViT and AdamW for DenseNet-201, both with a learning rate of 0.001. Post-training validation measured classification accuracy on the validation set.
- For the test dataset, HER2 IHC slides obtained from surgically resected 19 breast cancer patients were obtained from Asan Medical Center and broken into 281 patches (1024x1024 pixels). Two pathologists and one student independently evaluated the slides and constructed a consensus dataset following a joint meeting (1+, 134 patches; 2+, 38 patches; 3+, 109 patches) (**Figure 1**).

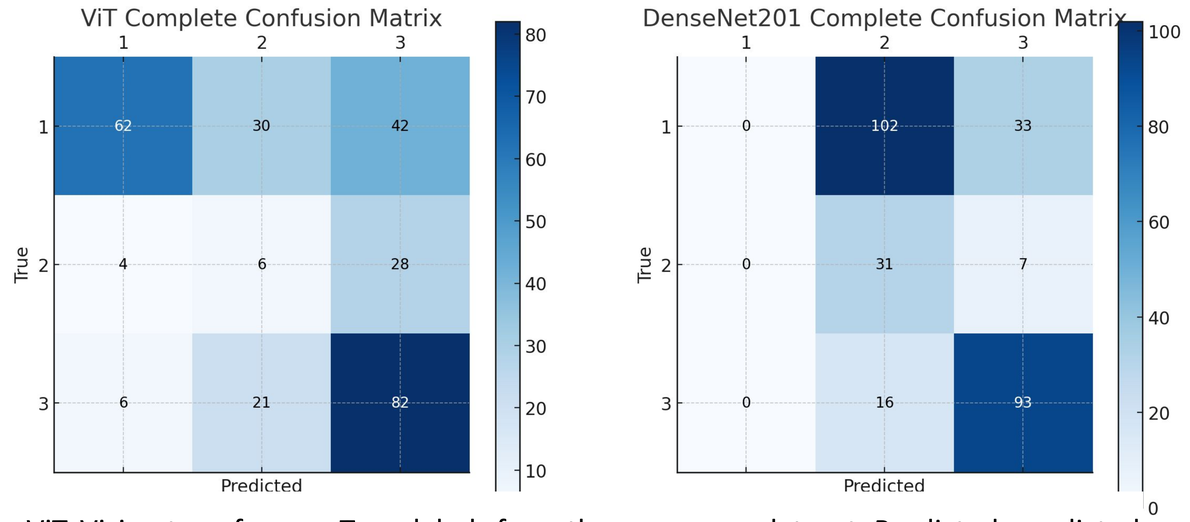


H&E, Hematoxylin and eosin; IHC, immunohistochemistry; WSI, Whole slide image; ViT, Vision transformer; DenseNet, DenseNet-201; FC, Full convolution

Figure 1. **The AI models trained with BCI dataset and tested with a separate test set. ViT:** The ViT is trained using the BCI dataset. Multi-head attention facilitates the parallel processing of diverse features. Proportionwise feed-forward is employed for sequence-to-sequence transformations, while Add&Norm normalizes following the addition of the input and its residual.; **DenseNet-201:** DenseNet-201 initiates feature extraction within DenseBlocks, succeeded by convolution layers for mapping. Subsequently, pooling reduces the spatial dimensions. Full convolution is applied to preserve spatial feature transformations.

RESULTS

- For validation accuracy, ViT and DenseNet-201 each achieved 91.0% and 65.6% on the BCI dataset. When tested on the consensus dataset, ViT outperformed DenseNet-201 with an overall accuracy of 58.7% compared to 40.1%. Both models showed varied performance across different HER2 IHC levels (**Figure 2**).



ViT, Vision transformer; True, labels from the consensus dataset; Predicted, predicted scores by AI models

Figure 2. **Comparative confusion matrices for IHC classification against the consensus dataset.** The ViT model achieves an overall concordance rate of 58.7% with the consensus, registering 45.9% (62 out of 135) for 1+, 15.8% (6 out of 38) for 2+, and 67.0% (73 out of 109) for 3+. DenseNet-201 records an overall match rate of 40.1% with the consensus, showing 0% (0 out of 135) for 1+, 81.6% (31 out of 38) for 2+, and 85.3% (93 out of 109) for 3+.

Table 1. Inter-rater reliability: Cohen's kappa scores among pathologists and AI models

	Consensus	P1	P2	ViT	DenseNet
Consensus	1.000	0.816	0.787	0.360	0.240
P1	0.816	1.000	0.623	0.335	0.259
P2	0.787	0.623	1.000	0.289	0.130
ViT	0.360	0.335	0.289	1.000	0.133
DenseNet	0.240	0.259	0.130	0.133	1.000

P1, Pathologist 1, P2, Pathologist 2, ViT, Vision transformer, DenseNet, DenseNet-201

- High concordance was observed among human raters with kappa scores 0.623-0.816. However, compared to AI models, agreement levels were significantly lower with kappa scores from 0.130 to 0.360 (**Table 1**).

DISCUSSION

- DenseNet-201 demonstrated impressive predictive capabilities for HER2 IHC scores of 2+ and 3+, but it failed entirely in accurately predicting 1+. In subsequent research, reducing layers within DenseNet-201 could enhance computational efficiency and potentially improve its predictive accuracy.
- Given ViT's nature as a transformer model, a larger dataset could be particularly beneficial. For future studies, augmenting the training set could bolster the performance of both ViT and DenseNet-201.
- For future studies, we recommend testing the model on an external cohort to validate its efficacy across diverse datasets

CONCLUSION

- The effectiveness of automated HER2 IHC classification relies heavily on the specific characteristics of the AI model and human supervision.
- Discrepancies in kappa scores between the models and pathologists highlight the need for further model optimization.