ResNet18-based classification of Breast HER2 megapatches
: ventures into ViT with a small dataset model

12.18.2024

Jongwon Lee

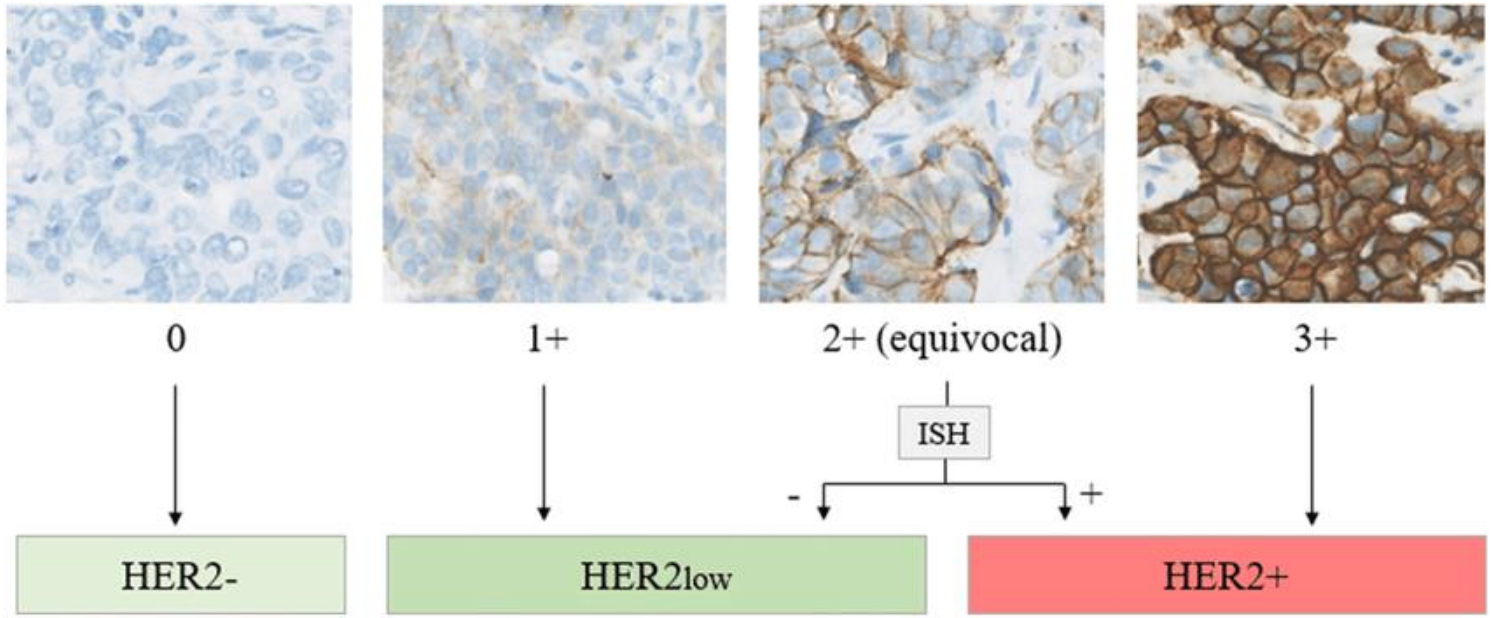
Department of Brain and Cognitive Engineering

Contents

- Introduction:
  - Pathologic image diagnosis of HER2 IHC patches
- Methods: A relatively small dataset (3500 patches) classification
  - ResNet18
  - Add-on: Comparison with ViT, Hybrid ViT
- Results: Confusion matrix, GradCAM
- Discussion and conclusion: The difference between CNN and ViT in scarce data settings.

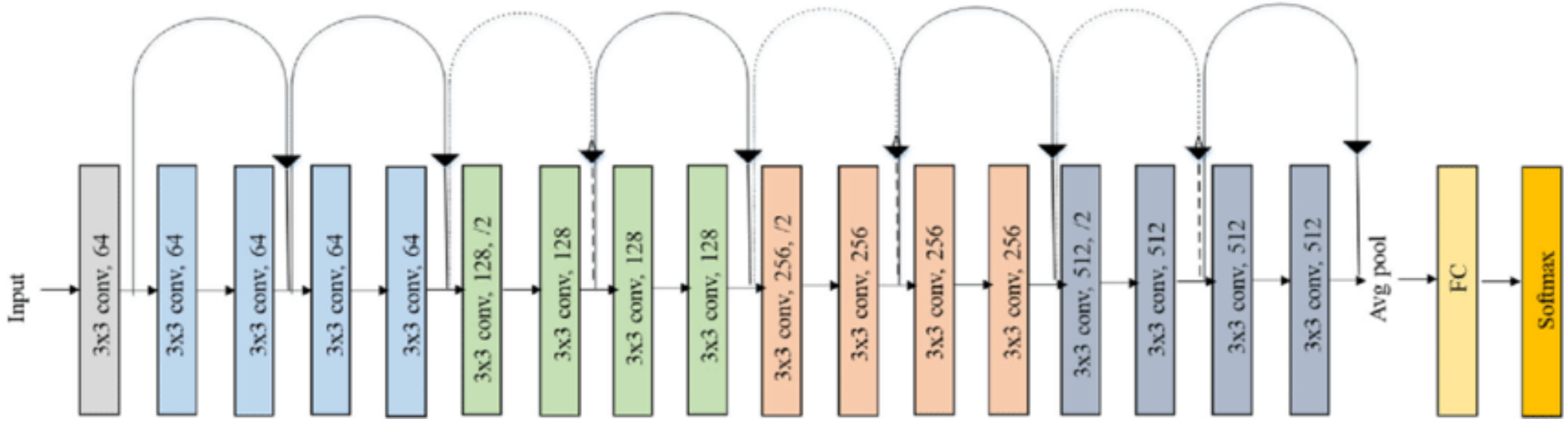
Breast cancer histology image: HER2 immunohistochemistry

- HER2 immunohistochemistry (IHC) scoring is a critical component in breast cancer



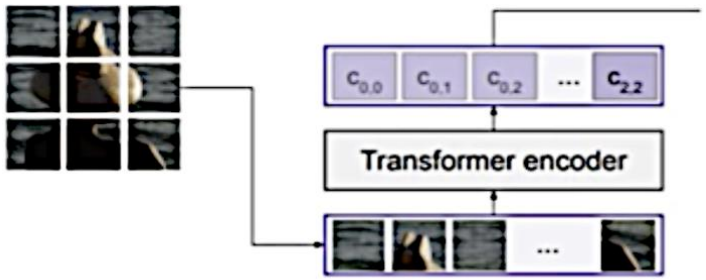
Challenges
1. It has to be carcinoma cells.
2. Membranous staining

https://www.researchgate.net/figure/Original-ResNet-18-Architecture_fig1_336642248

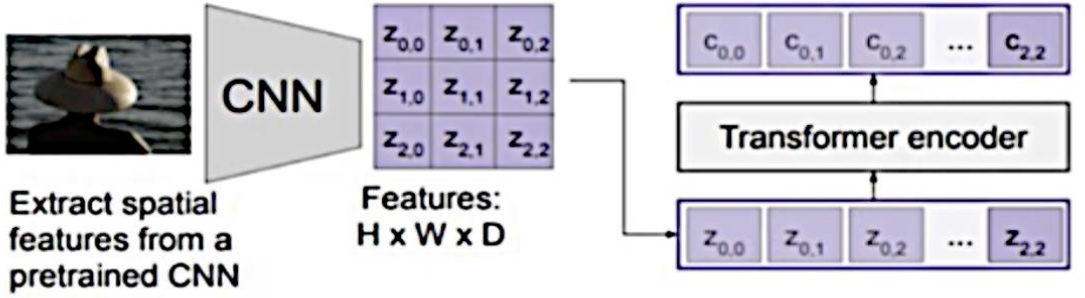


ResNet: skip connection- learning of the residuals: improvement of function.

Vision Transformer



Hybrid vision transformer



Self-attention: patches/pixels far away is still important

Methods

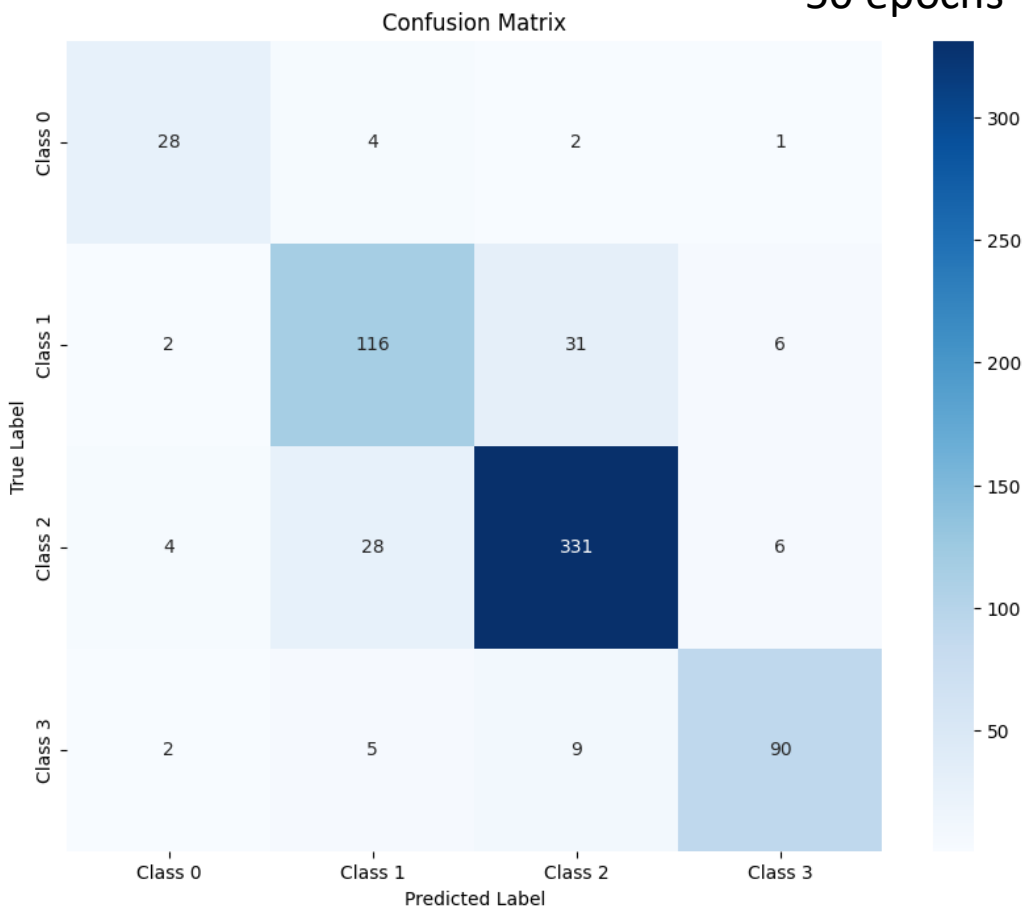
- BCI dataset, provided by Liu et al. (https://github.com/bupt-ai-cz/BCI)
  - 4,630 image patches (1024x1024 pixels) labeled 0, 1+, 2+, and 3+ (1+, 1,153 patches; **2+, 2,142 patches**; 3+, 1,335 patches).
- ResNet18, ViT, Hybrid ViT models were **pretrained with ImageNet,** trained with the data using PyTorch and GPU acceleration.
- Images were **resized to 224x224 pixels** and normalized. Training utilized an 80/20 data split(train-val-test), a batch size of 32, and Cross-Entropy Loss. Adam optimizer was used for ViT and AdamW for ResNet18, both with a learning rate of 0.001 for ResNet18, and scheduler for ViT.
- Conducted all training procedures with PyTorch frameworks on GPU hardware to accelerate computations.

Results

ResNet18

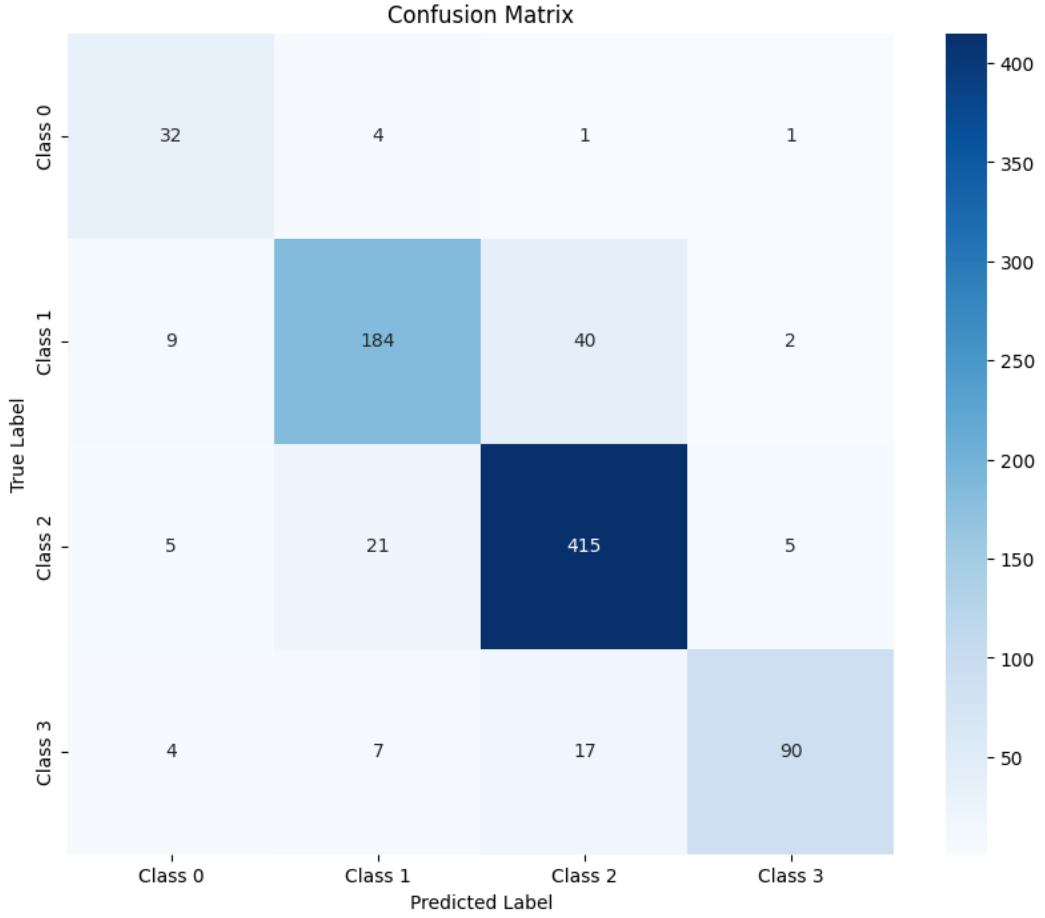
Validation set

50 epochs



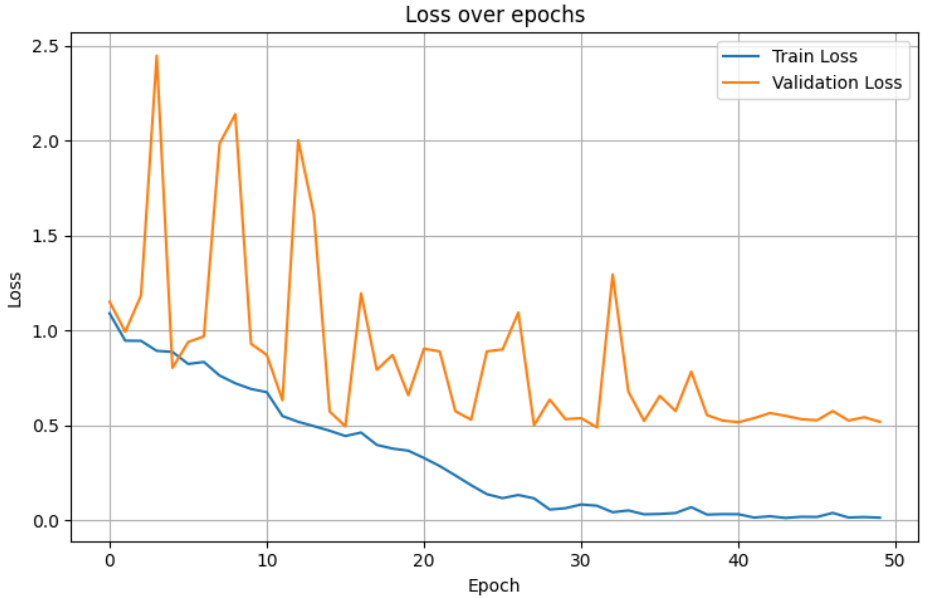
The validation set achieved an overall accuracy of about 85%.

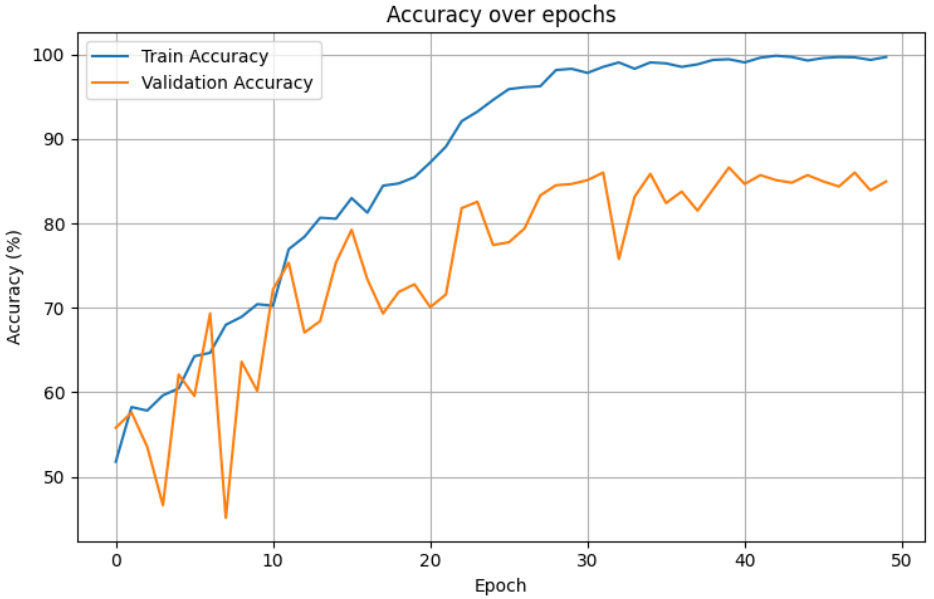
Test set



Overall test accuracy reached approximately 86.1%.

Significant fluctuations in the validation loss and accuracy





- **Insufficient Data or Class Imbalance**
- **Learning Rate Too High**:
  - Even with LR=0.001, the model may oscillate.- small dataset
  - Solutions: Reduce LR, apply LR scheduling, or switch optimizers.
- **Lack of Regularization and Augmentation**:
  - Inadequate dropout/weight decay → overfitting.

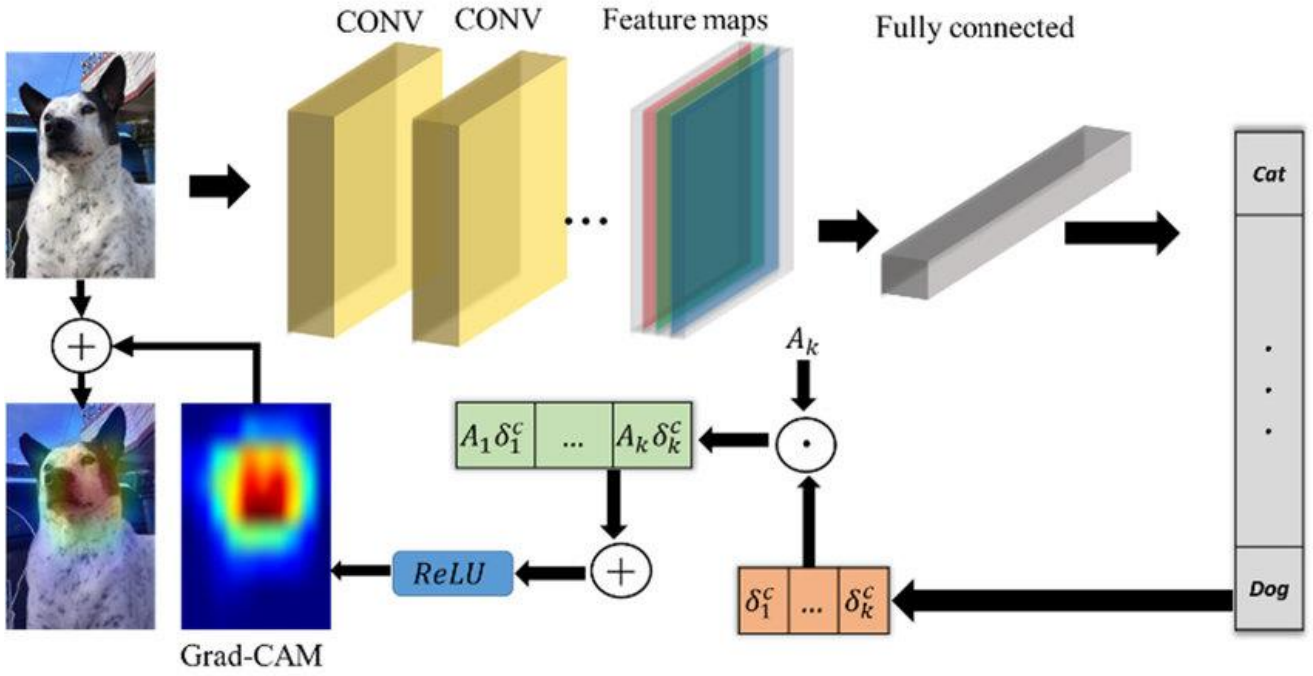
Grad-CAM (Gradient-weighted Class Activation Mapping)

1.Input & Output

- •Starts with a dog image (input)
- •Model predicts classes (e.g., "Cat", "Dog") at the end

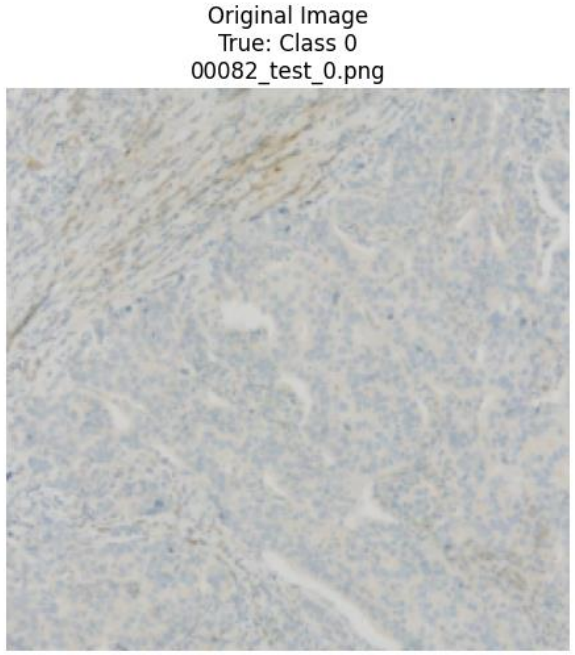
2. Grad-CAM Process

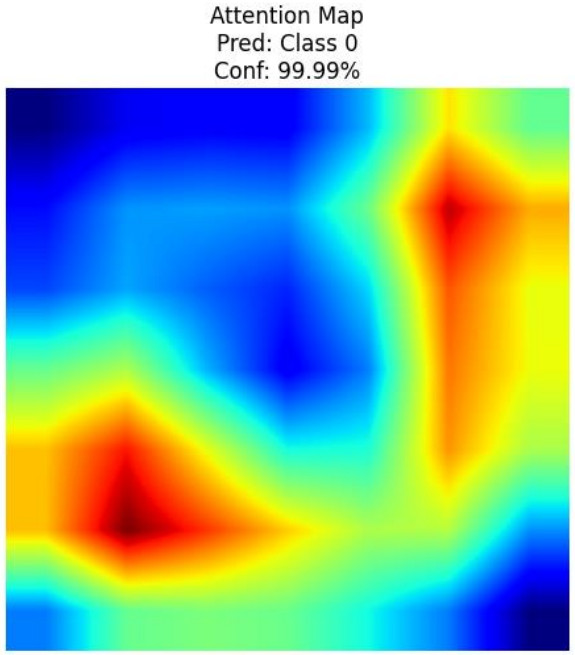
- •Gets feature maps from last convolutional layer
- •Computes gradients with respect to "Dog" class (δ^k)
- •Multiplies feature maps by importance weights (Ak) (parameter)
- •Applies ReLU to highlight positive contributions
- •Overlay for the most important areas of feature selection

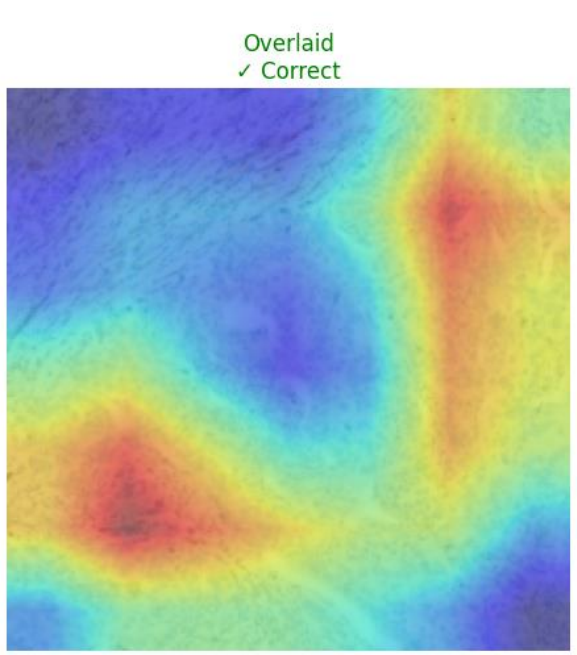


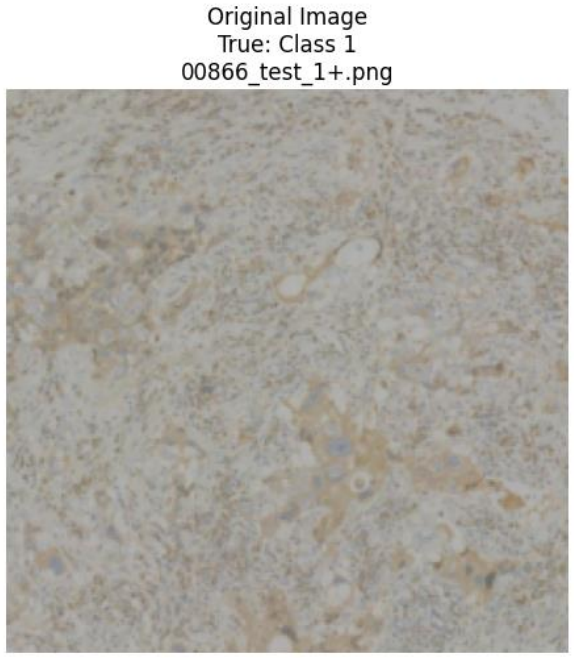
ResNet18

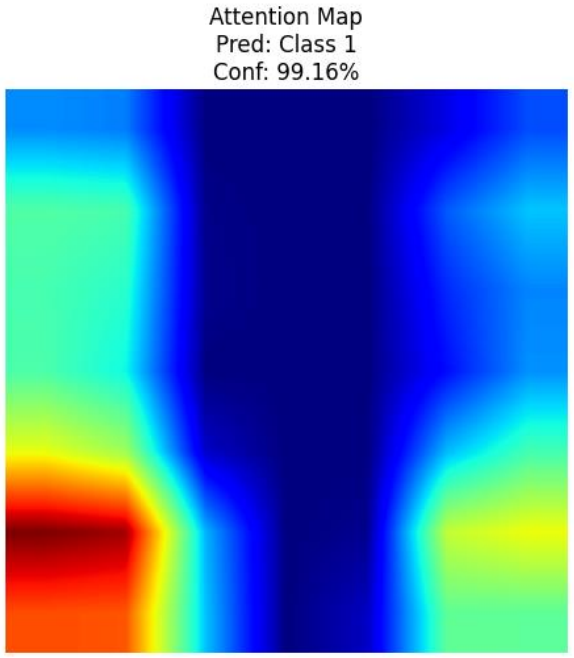
- **Classification Success**
- Focused on the edges of the tumor.
- Ignores the center..

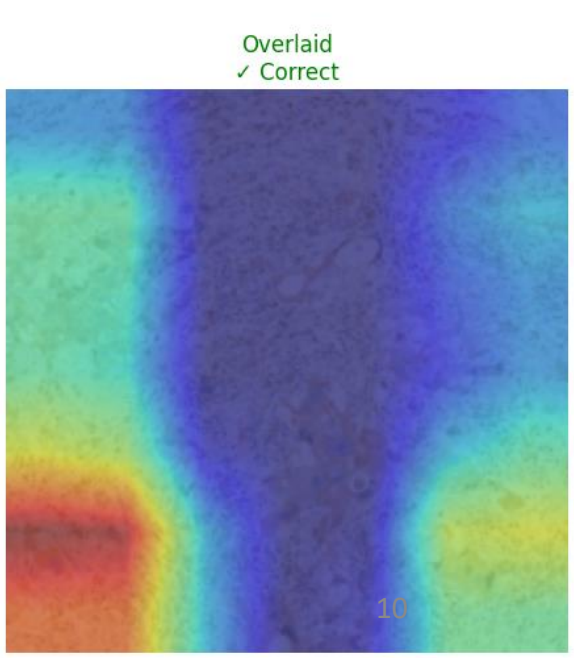






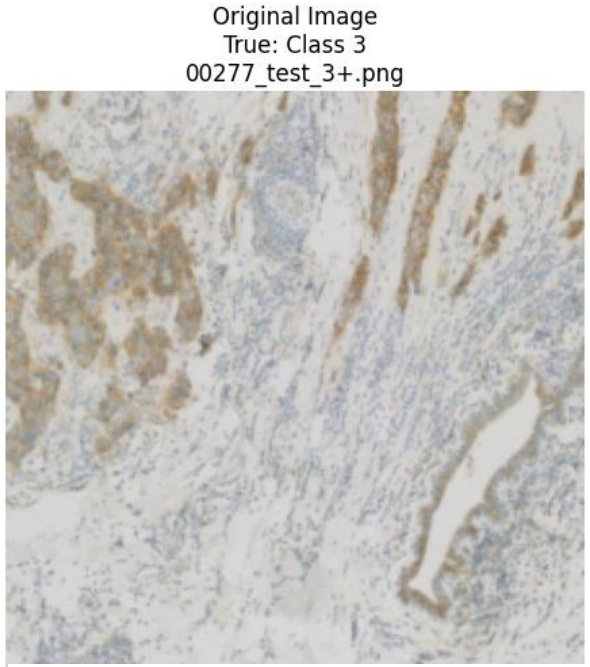


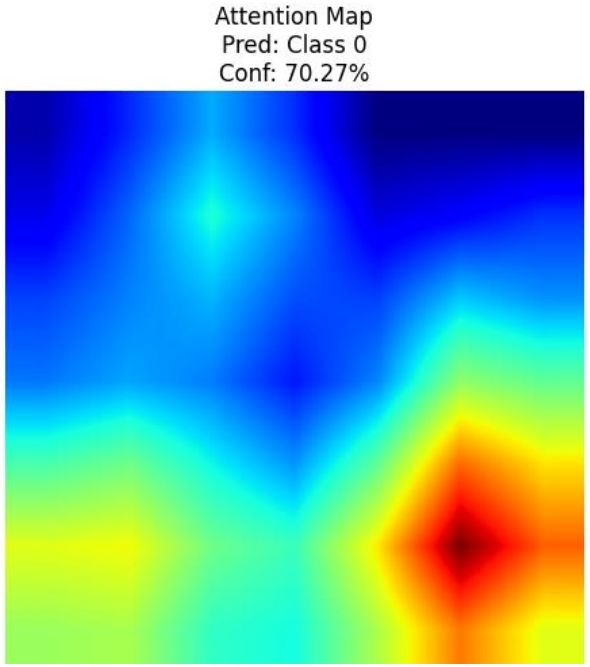


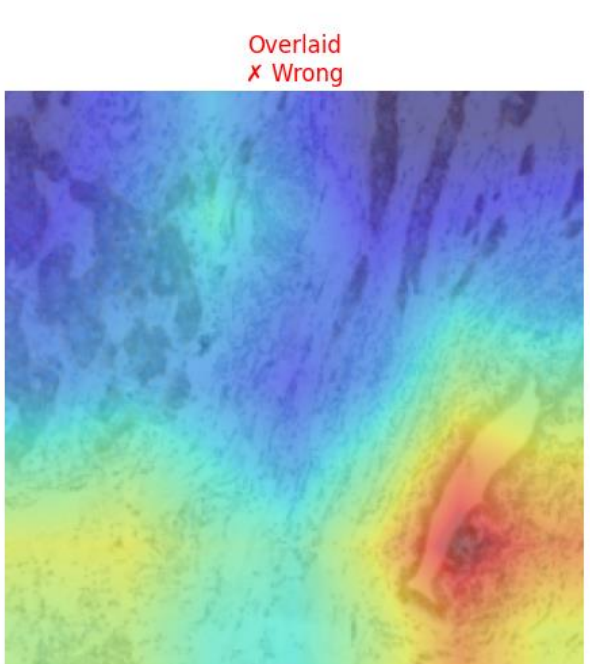


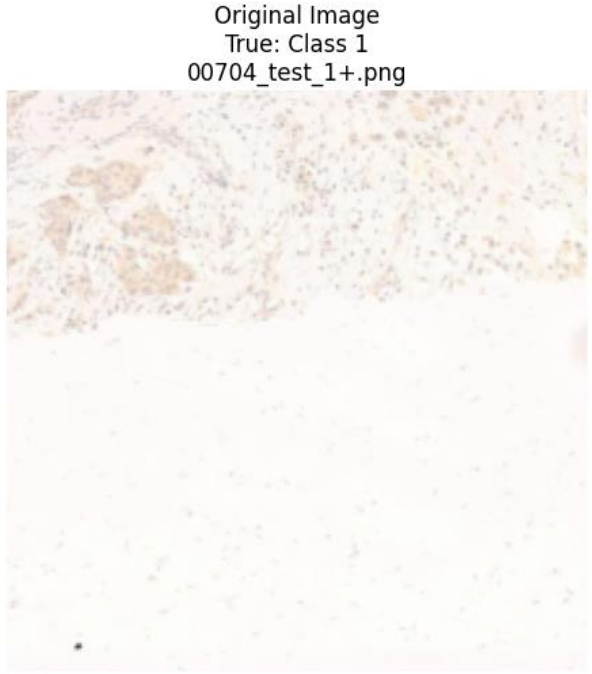
ResNet18

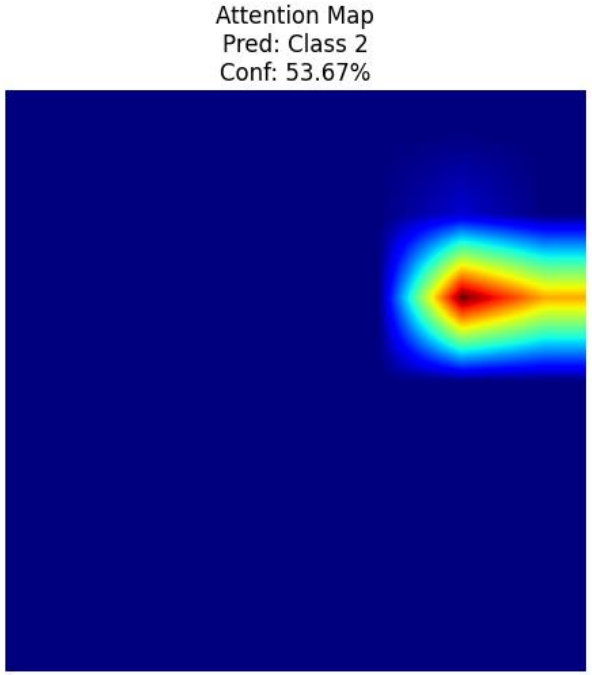
- **Failed classification:**
- Ignored tumor and focused on edge

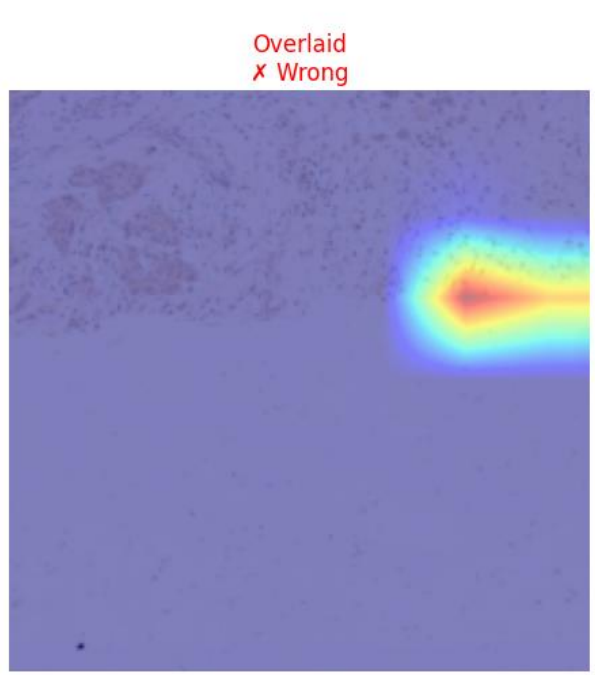




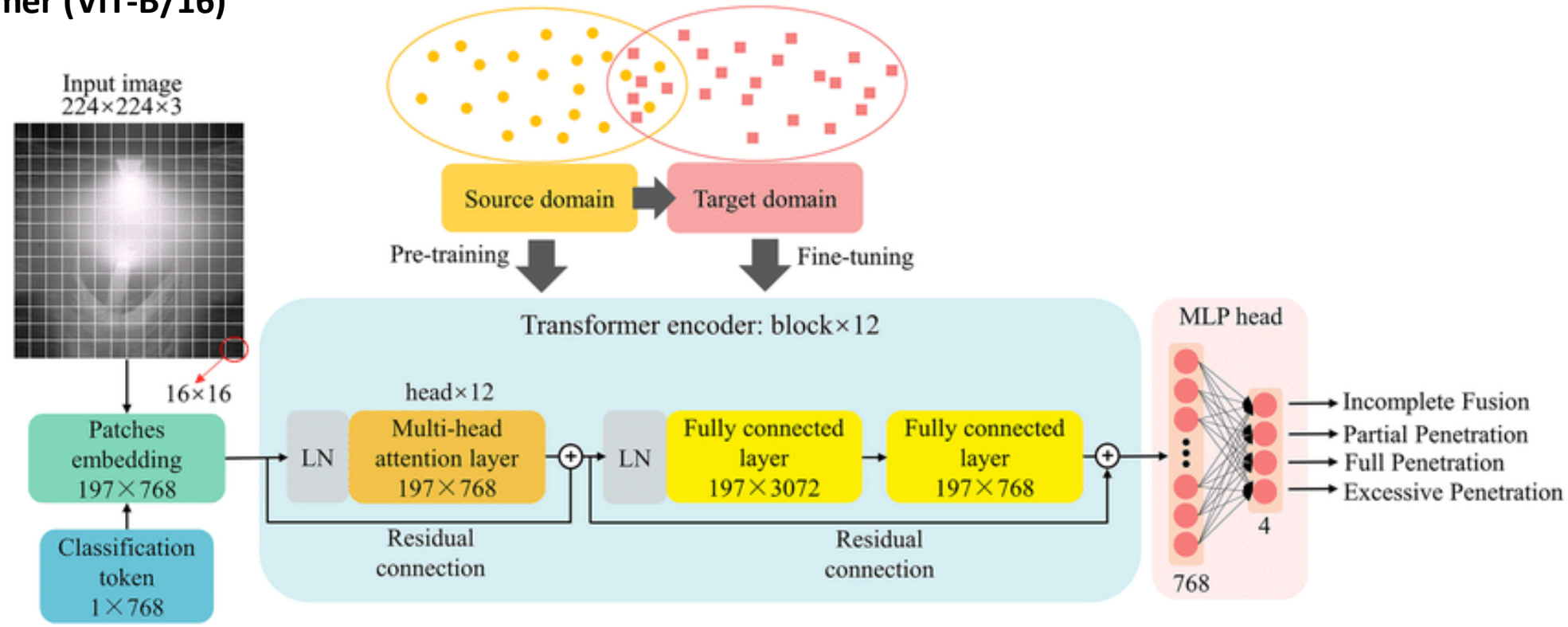








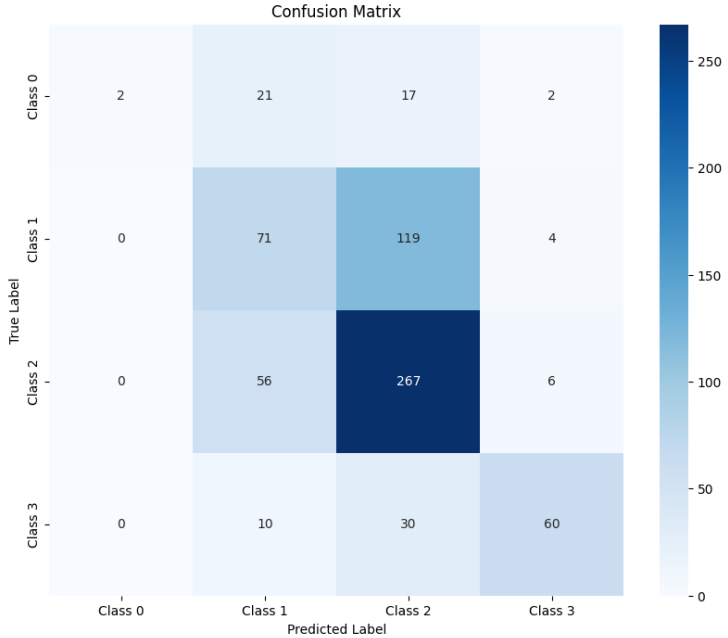
Vision Transformer (ViT-B/16)



- **Model**: Patch Size 16x16 (not 14x14!!)
- **Input**: Image divided into 14x14 patches (for 224x224 resolution)
- **Architecture**:
  - **Layers**: 12 Transformer Encoder Blocks
  - **Attention Heads**: 12
  - **MLP Size**: 3072
- **Parameters: 86M**
- **Pretrained**: ImageNet

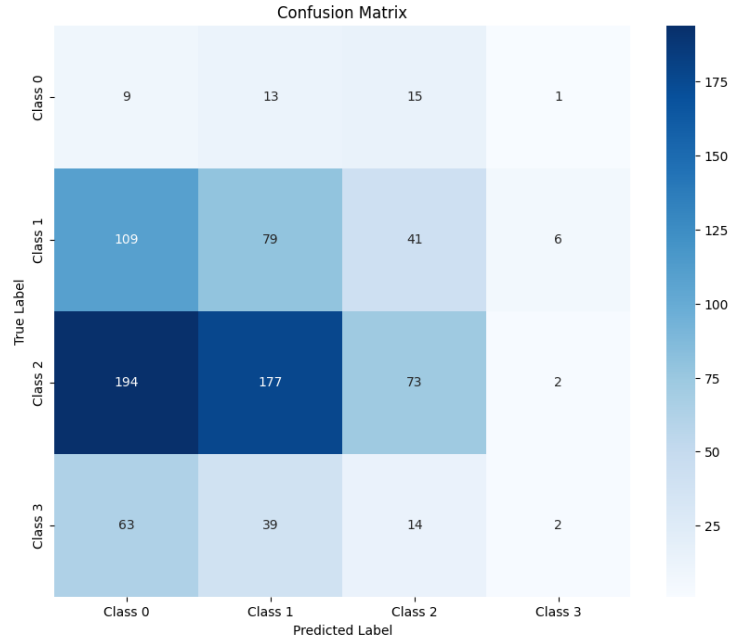
ViT– severe underfitting

Validation set



20/50 epochs

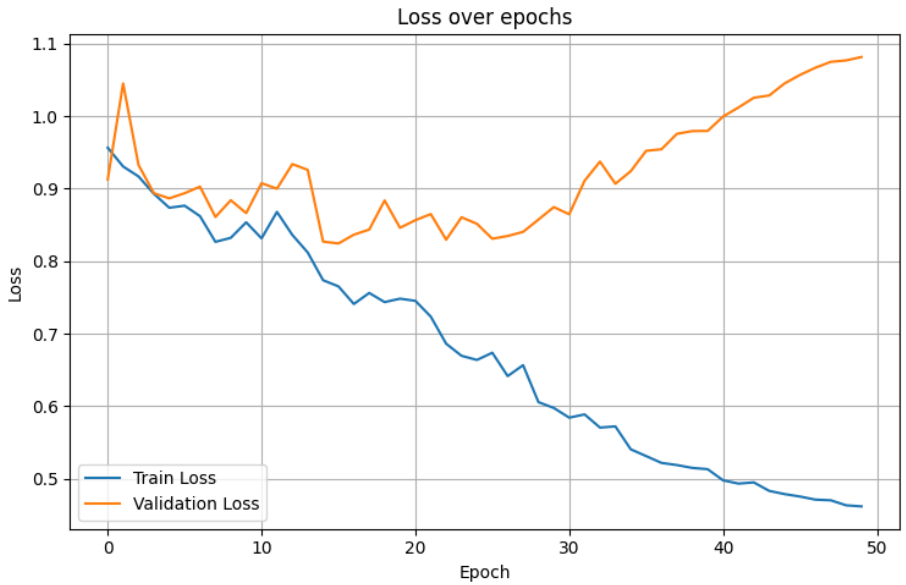
Test set

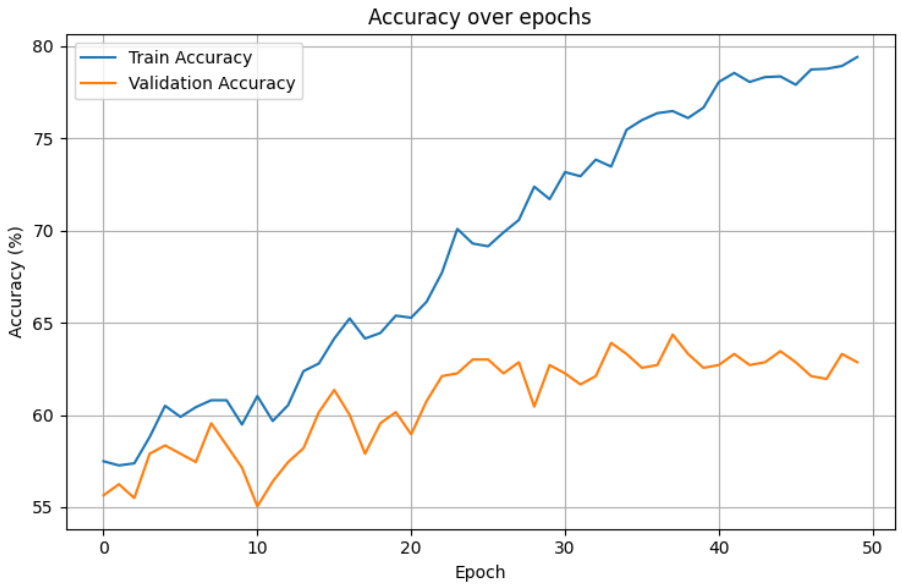


Validation set results: Accuracy of 0.63

Test set Accuracy: 0.19

Loss divergence





ViT GradCAM

- **Loss of focus:**
- Focuses on background, Scatters all over.

Image: 00366_test_3+.png

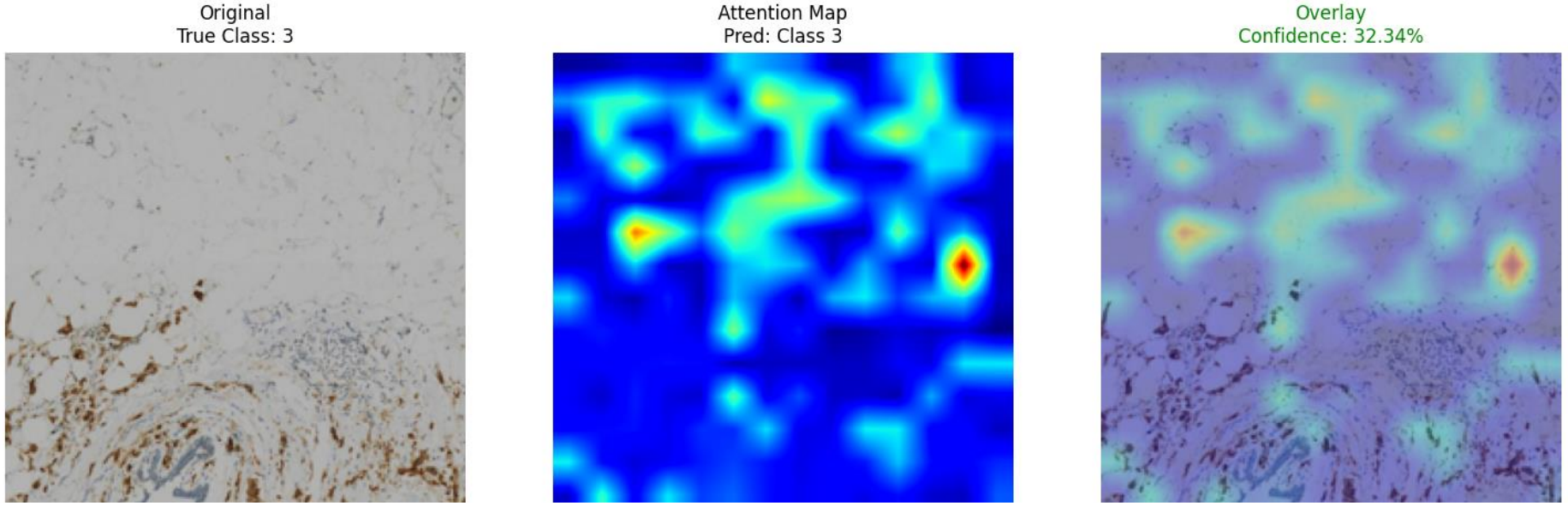


Image: 00318_test_3+.png

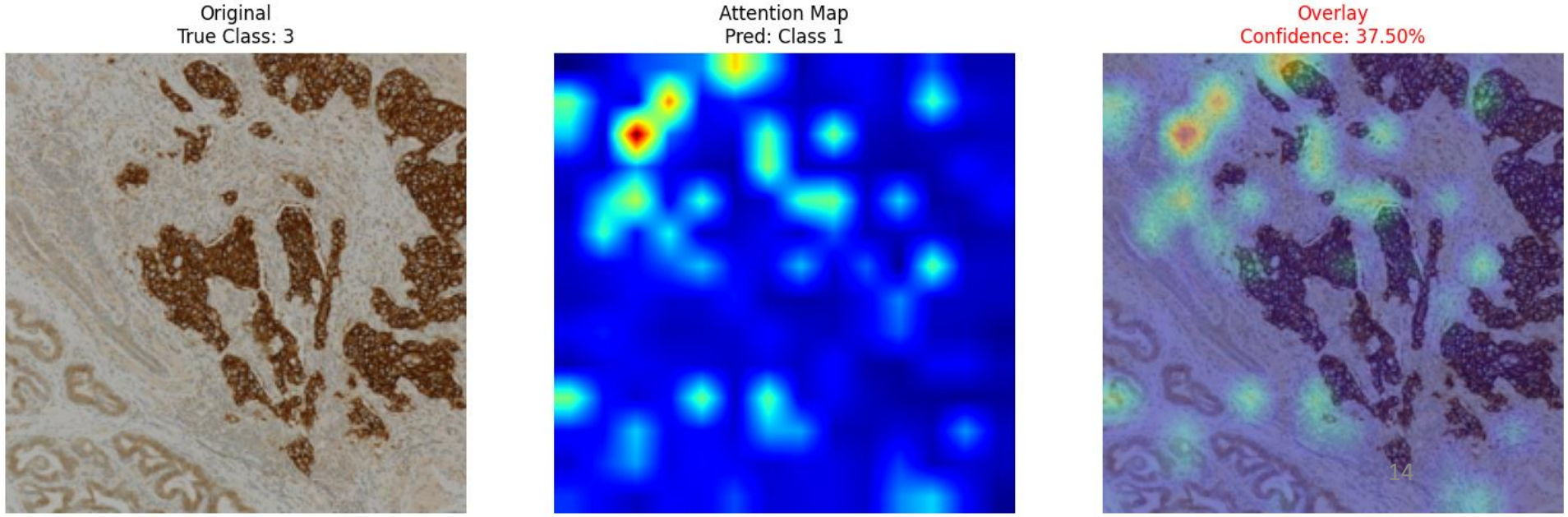
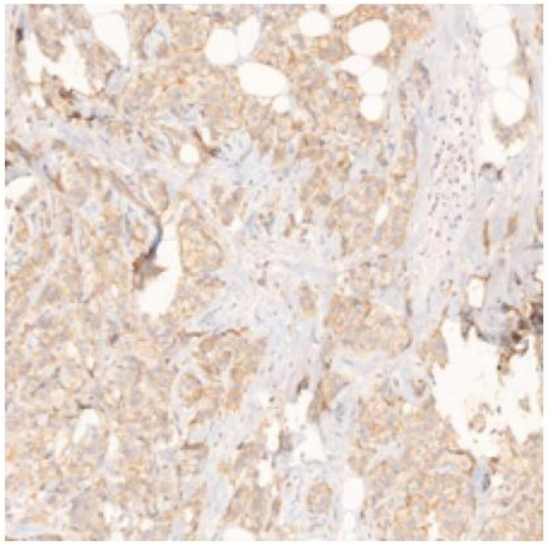
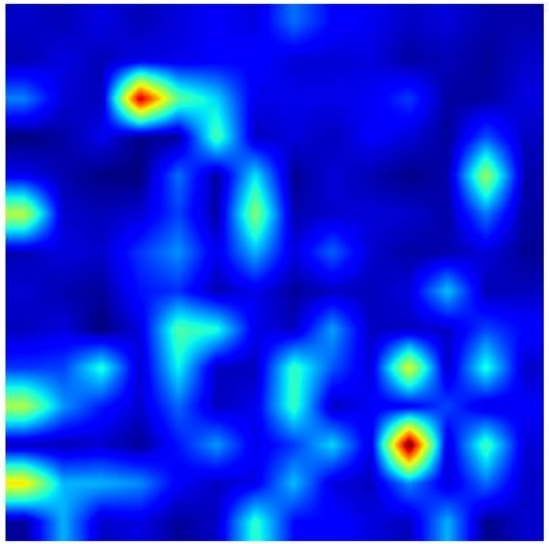


Image: 00303_test_3+.png

Original
True Class: 3



Attention Map
Pred: Class 0



Overlay
Confidence: 40.23%

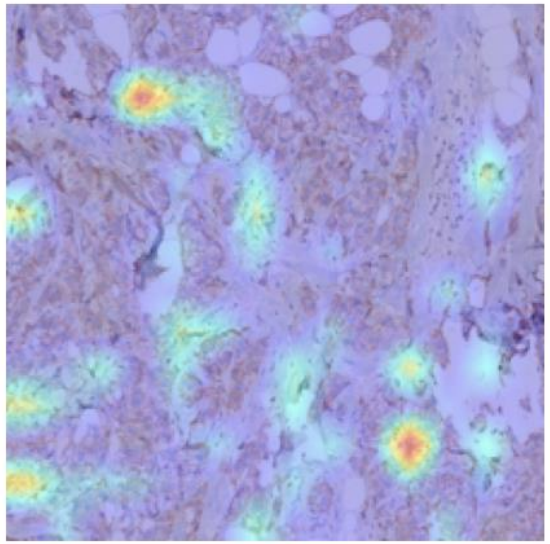
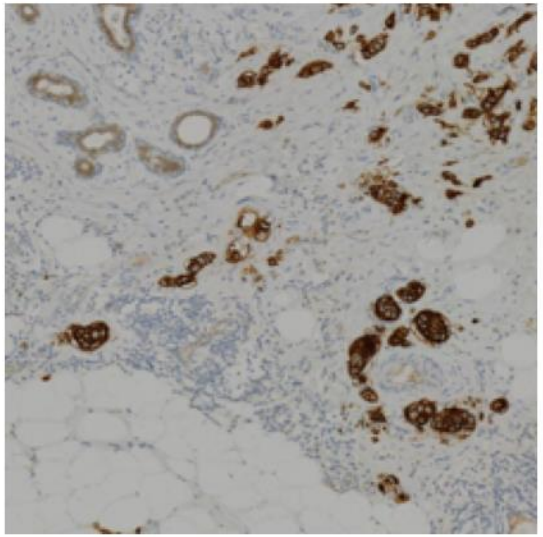
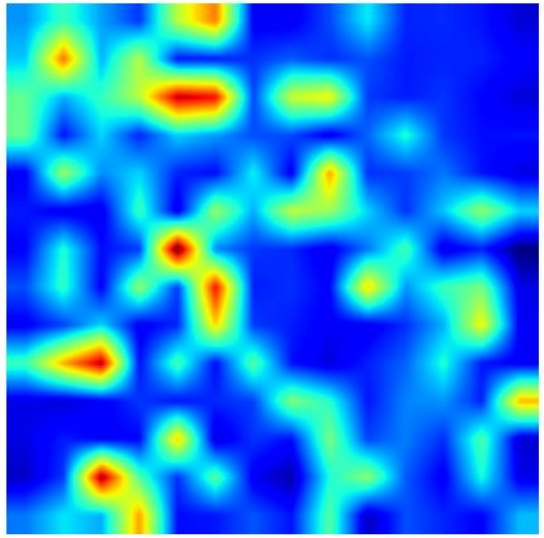


Image: 00280_test_3+.png

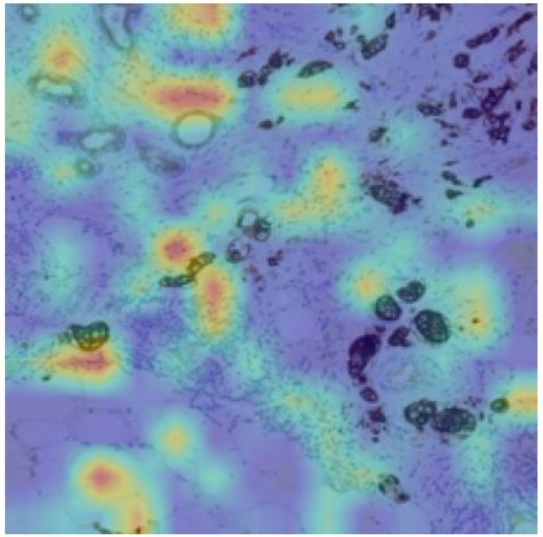
Original
True Class: 3

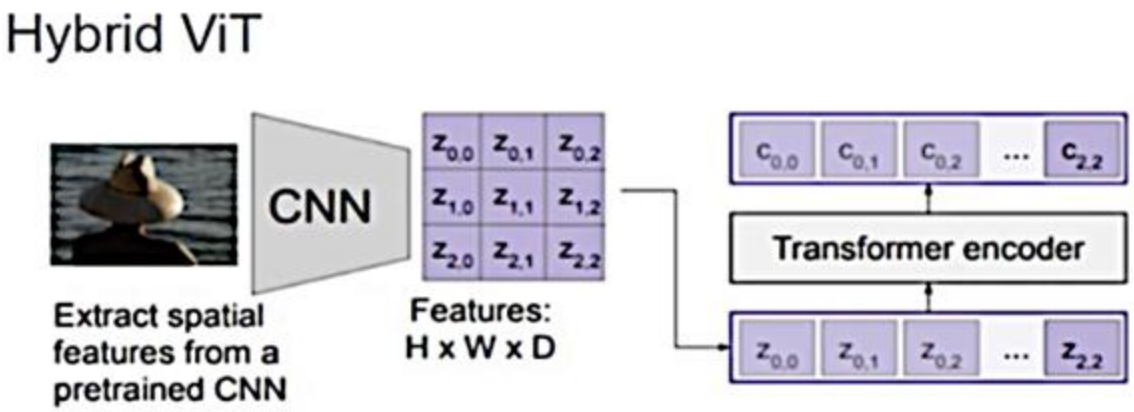


Attention Map
Pred: Class 1



Overlay
Confidence: 43.08%





Combines ResNet18 CNN with Vision Transformer

- CNN backbone: ResNet18 CNN extracts **spatial features** of the input image.
- Output feature map has dimensions
  - **H x W x D**:**H**: Height **W**: Width. **D**: Depth (number of channels).

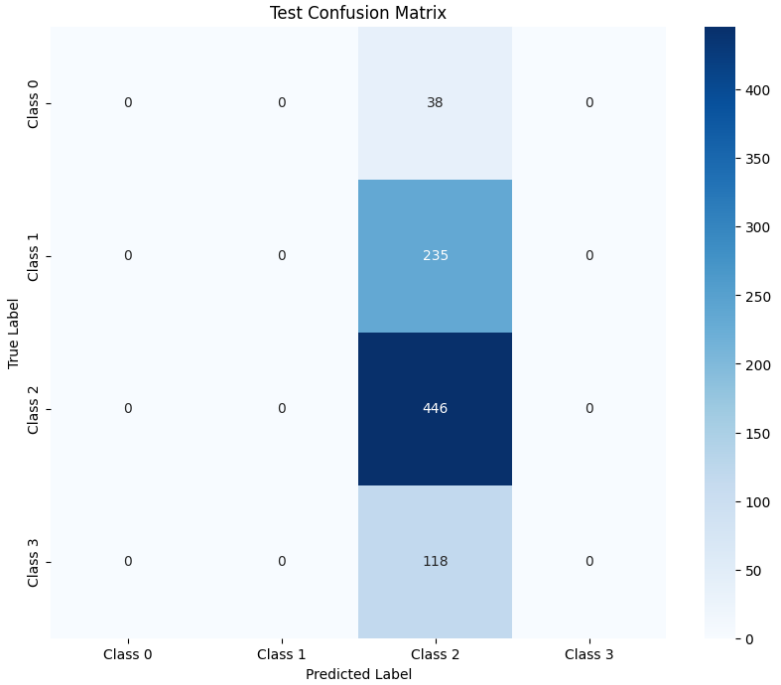
* Transformer: 4 layers, 8 attention heads

* Embedding dimension: **512**

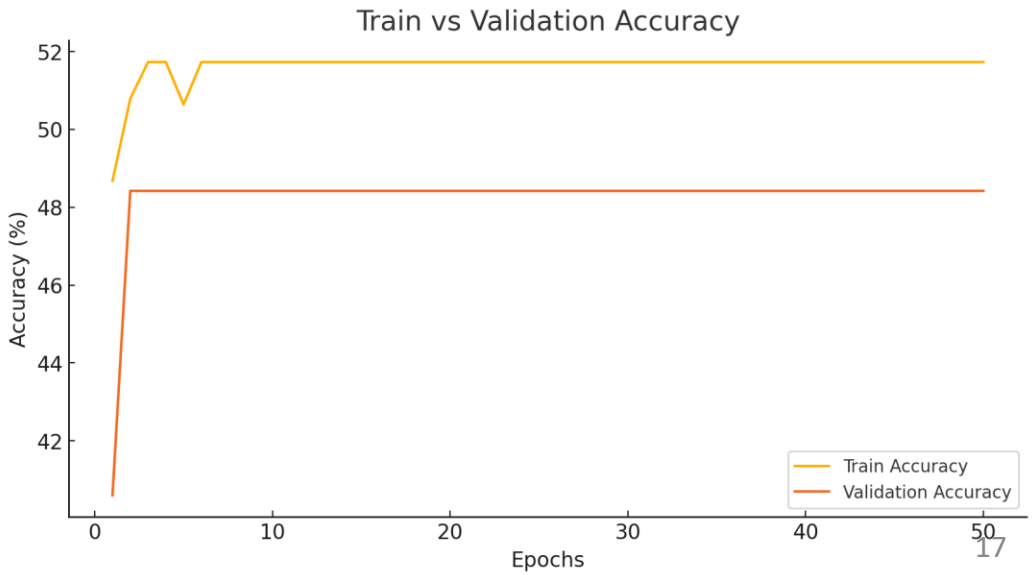
* Final layer: Linear classifier with 4 outputs

Hybrid ViT

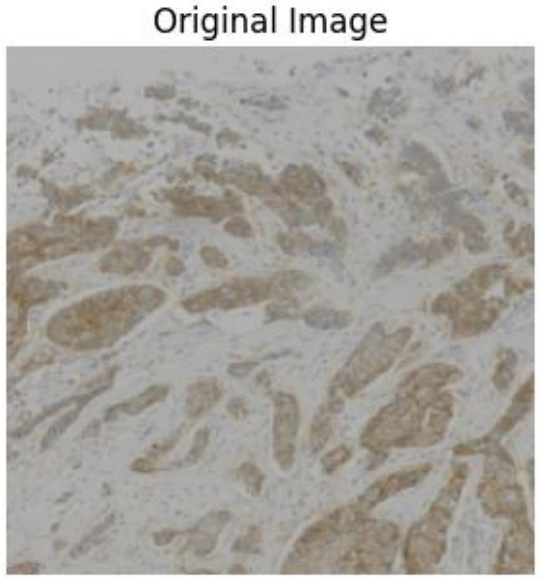
Test accuracy: 53%

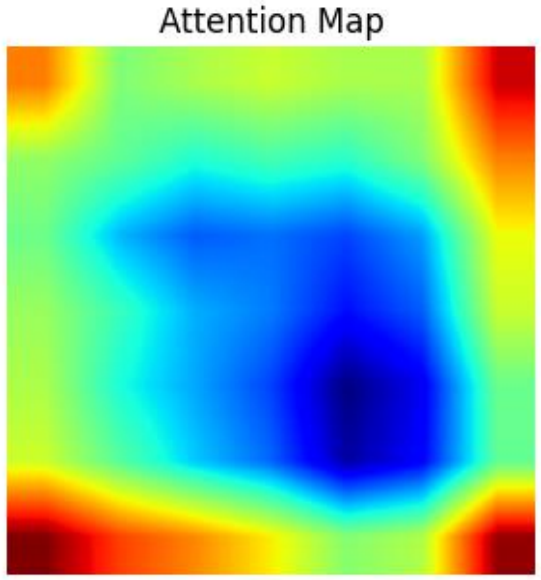


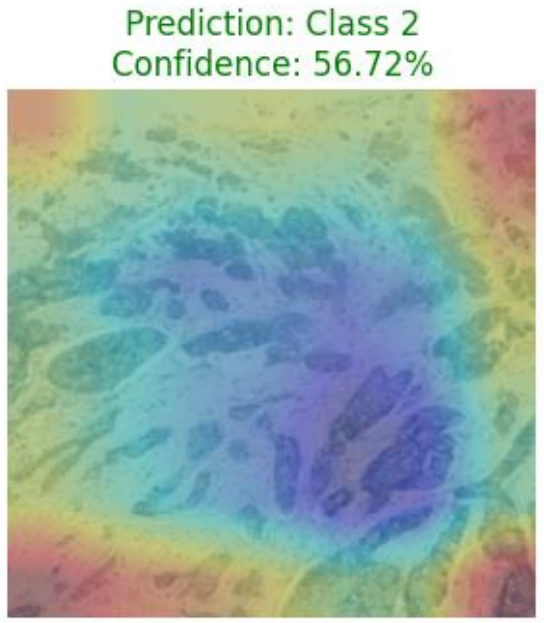
- **Validation set confusion matrix:**
  - **Serious underfitting**
- Training accuracy: 51.73%
- Validation accuracy: 48.42%
- **Test accuracy: 53%**



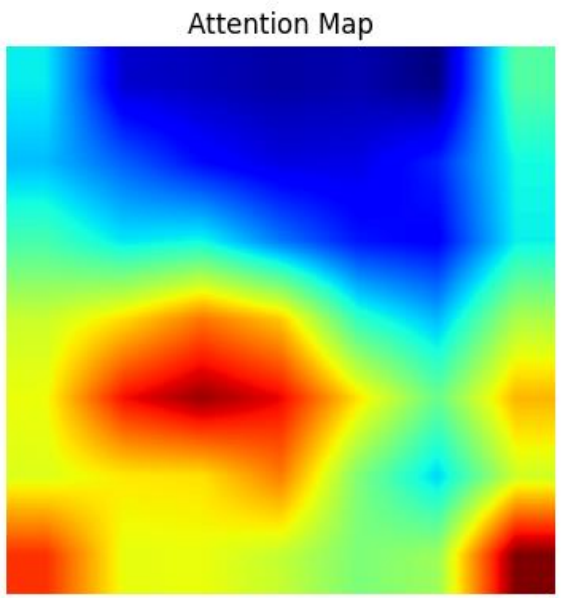
Hybrid ViT

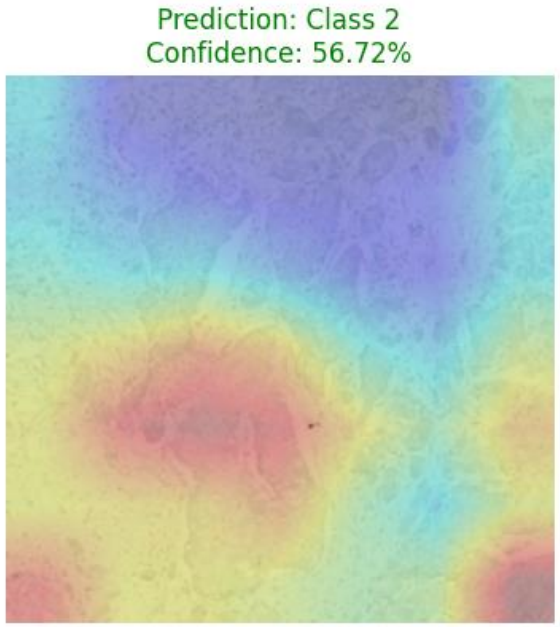




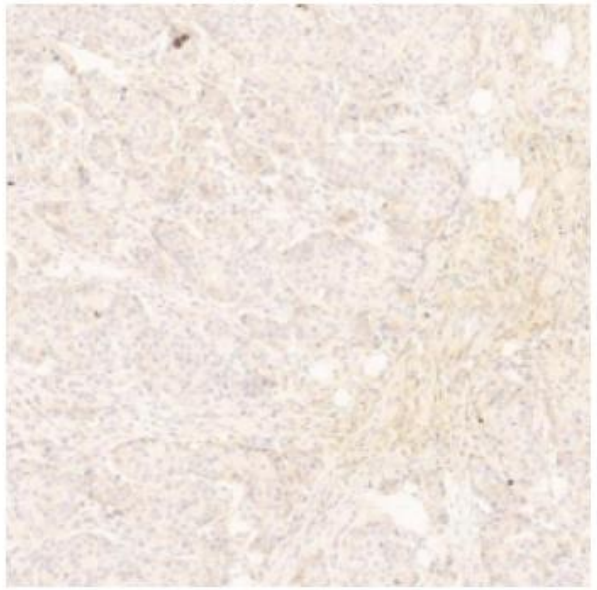


- Better than ViT: not scattered focus
- Edge-focusing is back.

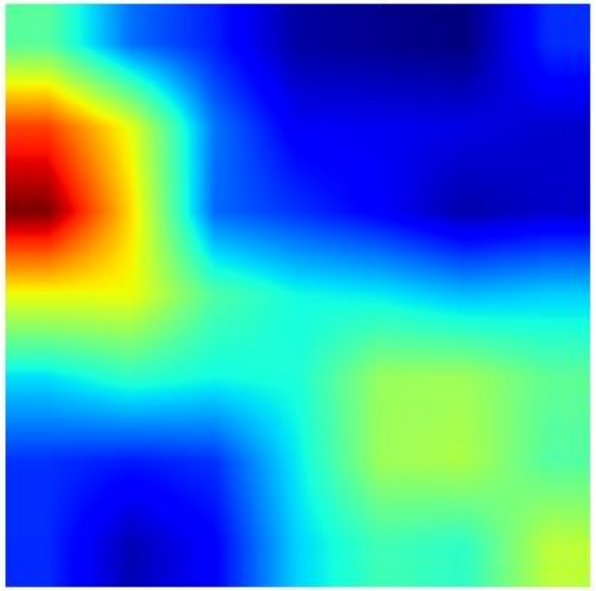




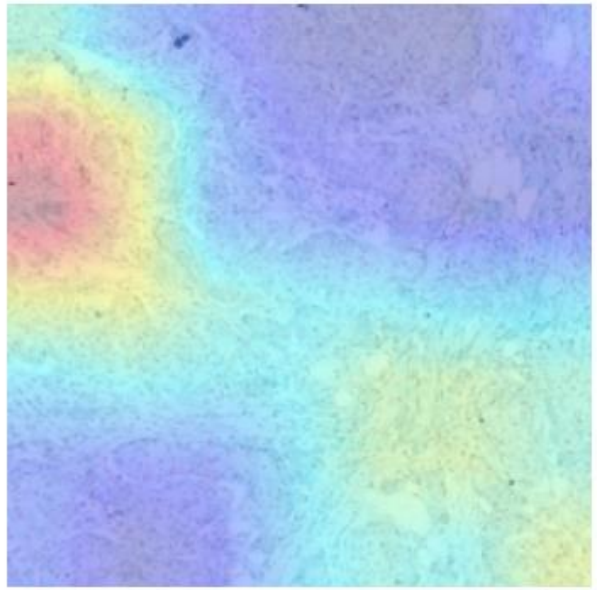
Original Image



Attention Map

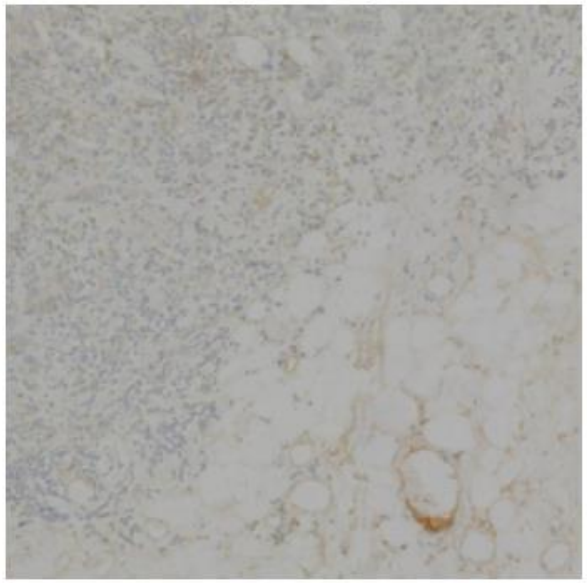


Prediction: Class 2
Confidence: 56.72%

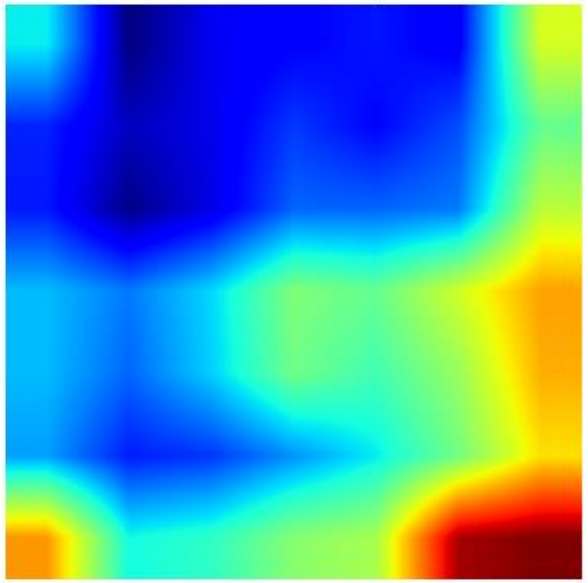


- **Wrongly predicted case:**
- Grade 1 or 2..? Image mislabeling!
- Still focues on the edge.

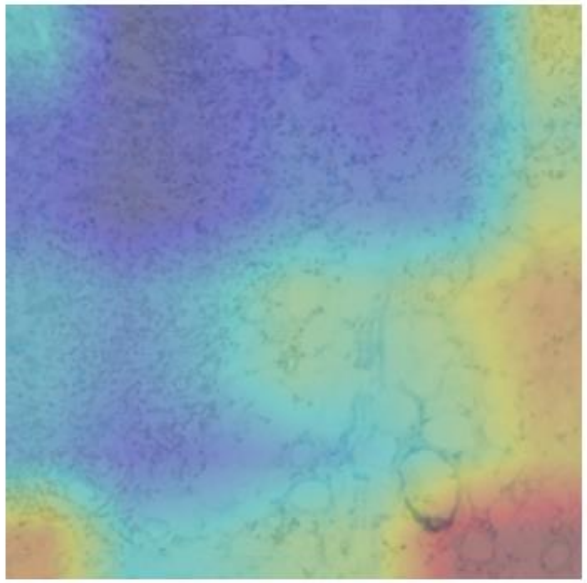
Original Image



Attention Map

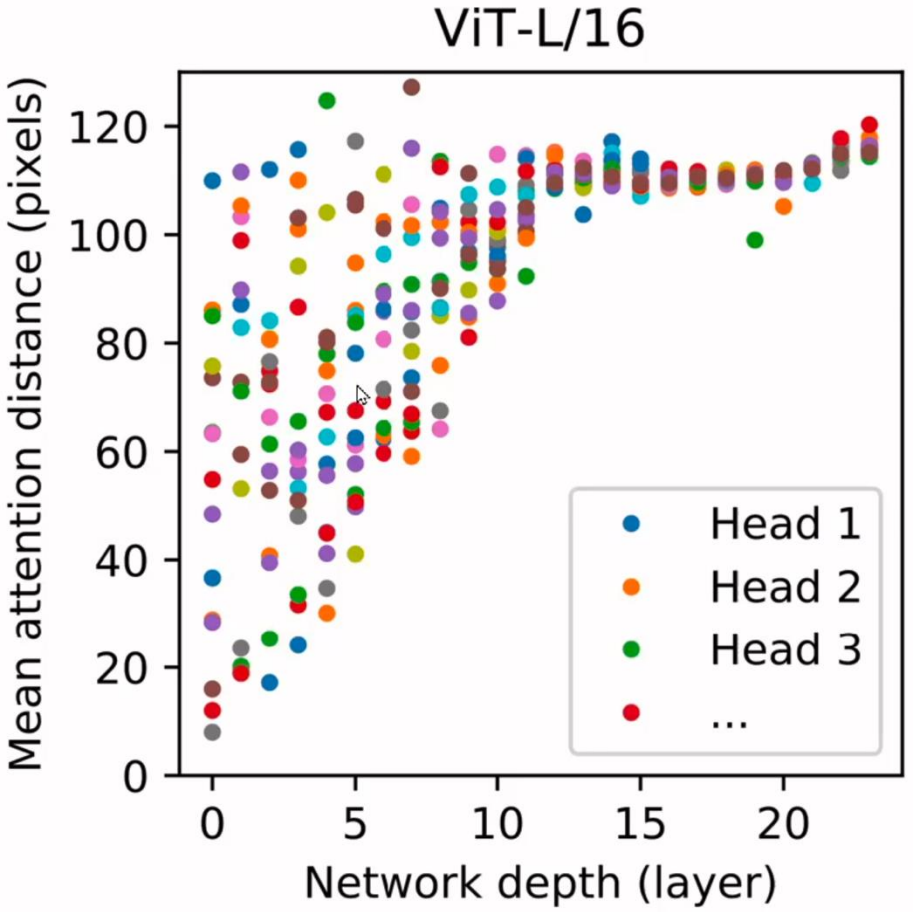


Prediction: Class 2
Confidence: 56.72%



Discussion and Conclusion

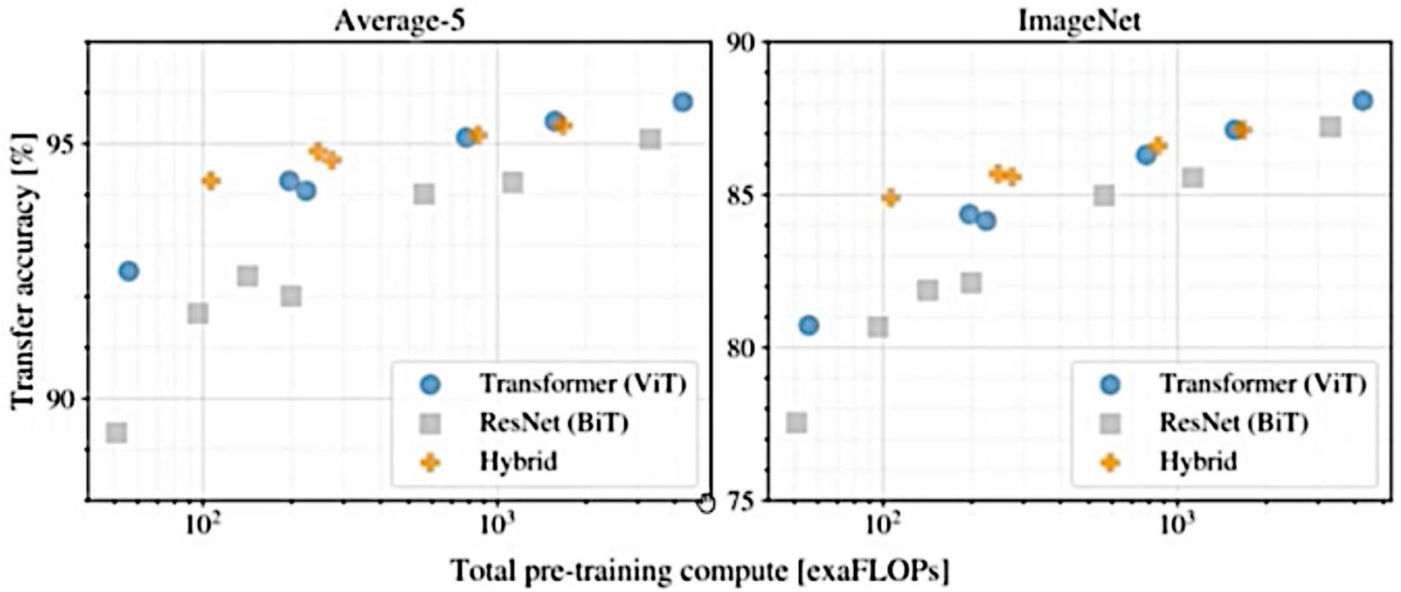
- **ViT comparison with CNNs**:
  - CNNs impose **inductive bias** (e.g., looking at local regions first).
  - ViT lets the model decide **when** and **where** to focus, **learning patch relationships flexibly.**
- **Scalability**:
  - Like LLMs, ViT performance improves with larger models and data.
  - With **small datasets**, **CNN's inductive bias helps.**
  - With **large datasets**, less inductive bias is better, enabling ViT to outperform.



From the eariest layers, ViT looks at faraway heads

Conclusion

- CNN focuses on local features, and position data is much well preserved, even in small datasets
- ViT, and Hybrid ViT does not do well with small dataset, because they have to train tons of parameters due to self attention
- Hybrid ViT, because it uses feature map from CNN, can focus more on specific and local features



Importance of
well-labeled dataset is paramount!

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

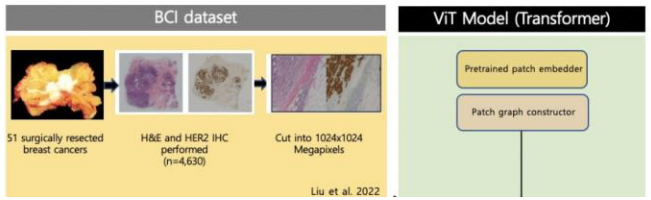
Jongwon Lee[1], Ji Won Hwang[2], GunHee Lee[1], and Hee Jin Lee [1]
[1]Department of Pathology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea;
[2]University of Ulsan College of Medicine, Seoul, Republic of Korea

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.

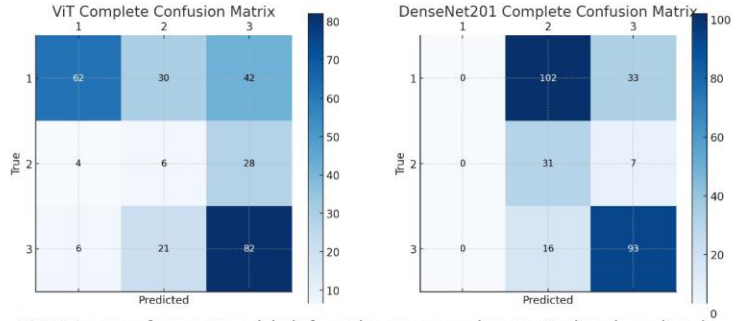
METHODS

- 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**).



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

Thank You