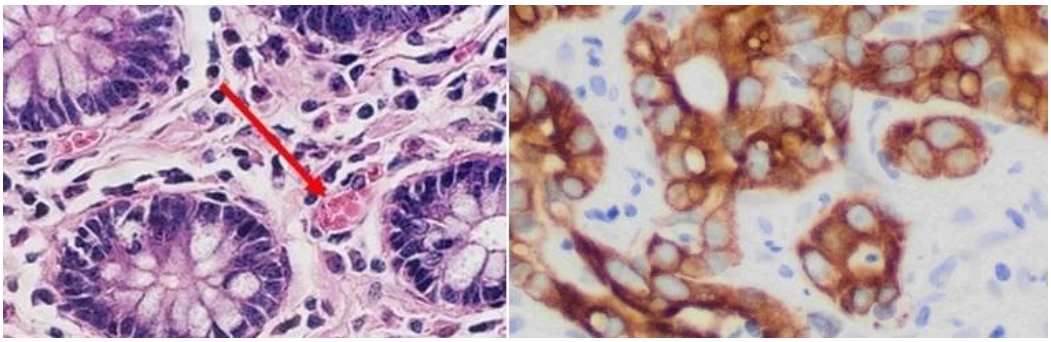
Deep Learning Models for Tumor Ratio Estimation in Histological Images

Project Overview: Hematoxylin and eosin (H&E) staining is the most widely used histological technique for visualizing tissue morphology in both clinical diagnostics and cancer research. While H&E slides provide rich structural information, they lack molecular specificity and do not directly label tumor cells. In contrast, immunohistochemistry (IHC) staining—such as Ki-67—offers explicit tumor cell labeling but is more expensive, less scalable, and often unavailable for large datasets. As a result, accurately estimating tumor burden directly from H&E images remains a central challenge in computational pathology.



Left: H&E result, red arrow indicates tumor cells Right: IHC result, tumor cells are brown, normal cells are blue.

Most existing deep learning approaches applied to H&E whole-slide images (WSIs) focus on binary classification (tumor-present vs tumor-absent) or coarse stage-level categorization, which are suitable for clinical workflows but insufficient for cancer biology research. In experimental metastasis models, researchers must quantitatively estimate tumor burden, often expressed as a tumor-to-tissue ratio or number of metastatic clusters, to evaluate treatment efficacy. This task requires higher precision than standard diagnostic models and must operate under conditions where pixel-perfect ground truth annotations are unavailable.

This project addresses the problem of tumor ratio estimation from H&E images under imperfect supervision, using Ki-67 IHC slides as an approximate ground truth despite spatial misalignment between consecutive tissue sections.

Key Challenge: Imperfect Ground Truth Alignment

A fundamental difficulty in using IHC as ground truth lies in the fact that H&E and IHC images are typically obtained from adjacent but non-identical tissue sections. While global tissue structure is preserved, cell-level correspondence is not guaranteed. Pixel-wise supervision is therefore unreliable, making direct segmentation-based learning infeasible.

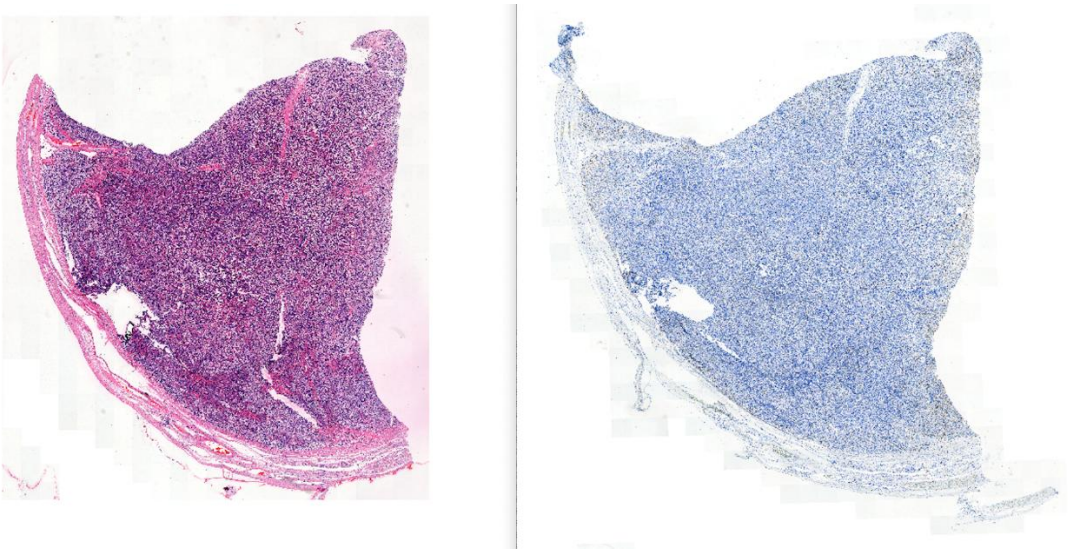
To address this limitation, we adopt the assumption—supported by prior histopathology studies—that tumor-to-tissue ratios remain locally consistent across consecutive sections, even when precise spatial alignment is lost. Based on this premise, the project reframes tumor estimation as a ratio prediction problem rather than a pixel-level segmentation task.

Pipeline and Baseline:

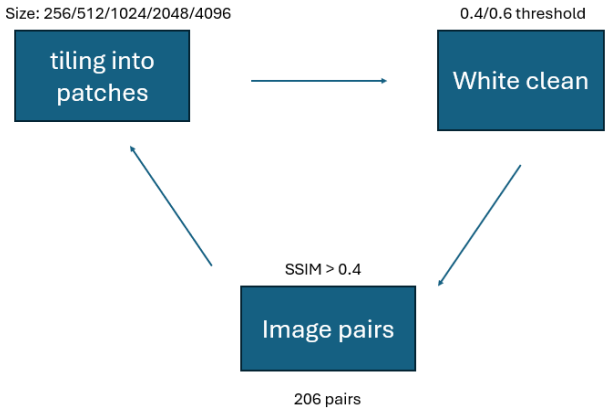
1. Data Preprocessing:

1.1 Download the Datasets

77 pairs of whole slide images of H&E staining and Ki67 stained IHC images are downloaded from https://zenodo.org/records/11218961. These slides are registered and aligned with method described in the paper. All slides are of size 31104 pixel * 31938 pixel



1.2 Data pre-processing



The preprocessing phase consisted of several iterative steps, including tiling, white cleaning, and similarity evaluation between image pairs, aimed at ensuring high similarity between H&E and IHC image pairs. This was a critical requirement for generating IHC images from H&E slides using GAN-based models.

Smaller patch sizes, such as 256-pixel tiles used in previous studies, proved inadequate for preserving discernible similarities between the image pairs. Larger patch sizes were then used to better

capture meaningful features and enhance the quality of the generated IHC images. The white cleaning step was particularly important for removing patches containing mostly or entirely background areas, as these patches caused crashes when processed with StarDist.

The Structural Similarity Index Measure (SSIM) was used to evaluate the similarity between H&E and IHC image pairs. A threshold of 0.4 was applied, and only pairs exceeding this threshold were retained for further analysis. Afterward, unsuitable images were manually removed, resulting in a refined dataset of 206 image pairs for subsequent studies.

2. Developing ground truth

The original study developed ground truth by recoloring IHC slides and calculating the ratio of brown pixels to the total of brown and blue pixels. However, this approach posed two significant challenges:

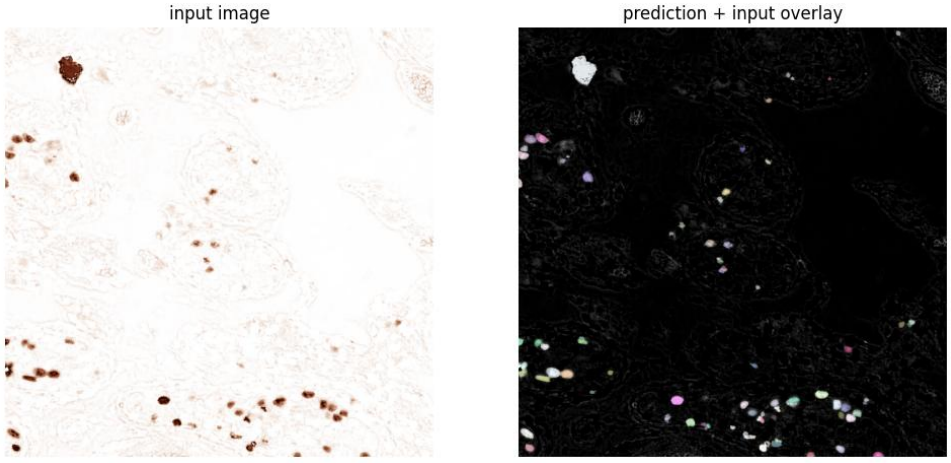
1.The threshold for recoloring is subjective and can vary significantly depending on the implementation.

2. The blue channel, which stains for total cell nuclei, often gives a slight blue tint across the tissue, leading to an overestimation of the nucleus area.

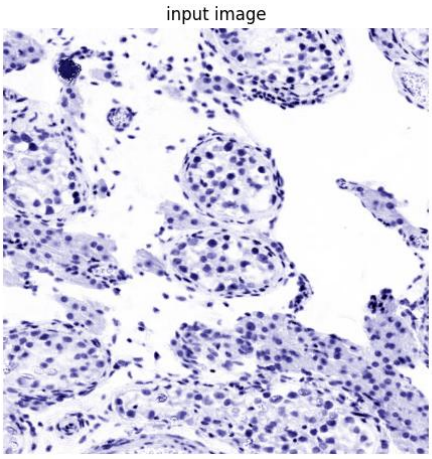
To address these limitations, StarDist was used for segmentation instead. StarDist specifically calculates regions with star-convex shapes, effectively segmenting only on cell nuclei. This approach minimizes the overestimation issue and provides more accurate segmentation compared to recoloring.

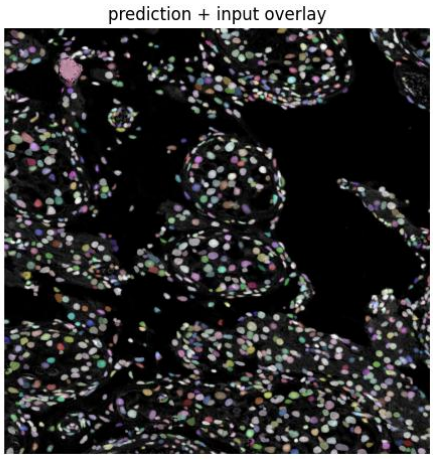
The process involved converting IHC patches from RGB to HED format, separating the image into different channels. Cell nuclei were segmented in both channels, and the ratio of tumor to total tissue was calculated by summing up the segmented regions. While simple thresholding methods did not yield satisfactory results, StarDist provided accurate segmentation that was visually validated. The segmentation code was adapted from a YouTube tutorial https://www.youtube.com/watch?v=L3dZ6fgmlII using a pre-trained StarDist model for efficient and precise cell segmentation.

Ki-67 (tumor) channel segmentation result:

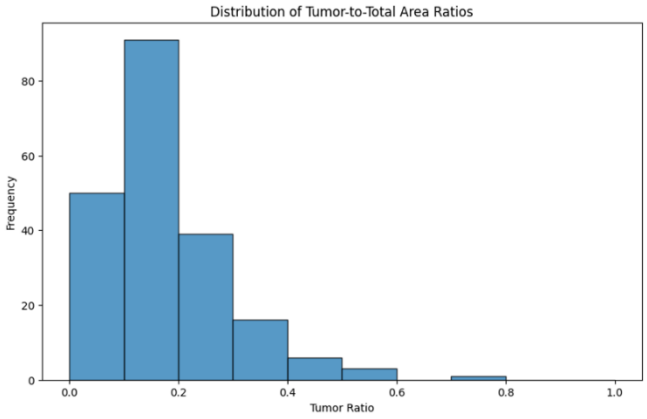


H (total tissue) channel segmentation result:





Tumor ratio with StarDist segmentation:

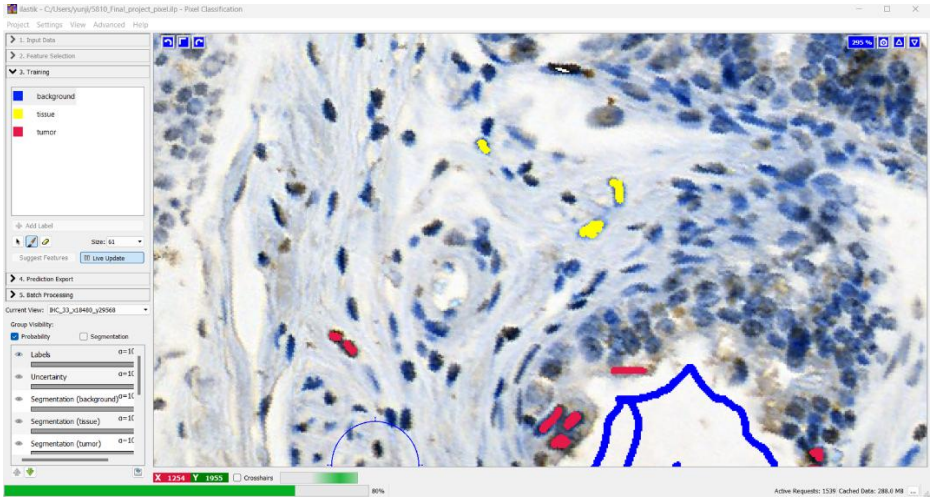


For H&E images, the same approach didn't work. The final decision is influenced by multiple factors beyond just color. Nuclear size, color balance, and nuclear arrangement are key contributors to identifying and classifying cell types and regions. Unlike IHC images, where specific stain colors (e.g., brown and blue) play a central role, H&E relies on a more holistic evaluation of morphological and spatial features.

3. Baseline development with *ilastik*

The original paper divided IHC patches into three groups based on their tumor area ratio: 0-0.2, 0.2-0.5, and 0.5-1. They then trained H&E patches using ResNet with these labels, achieving a maximum accuracy of 0.79 after optimization. However, this method simplifies the problem by categorizing tumor area ratios into just three classes. To achieve more precise outputs from H&E slides for our research, a different approach was necessary.

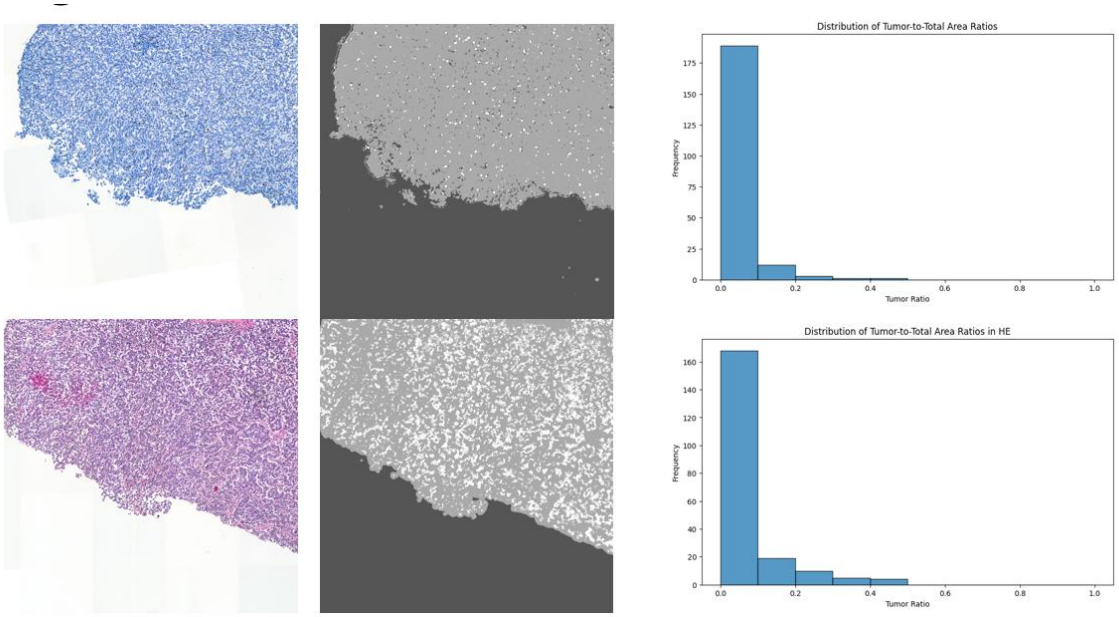
3.1 Using *ilastik* for Segmentation



Ilastik API screen shot

To improve accuracy, we employed *ilastik*, a versatile tool for image analysis and segmentation. Ilastik provides a user-friendly interface that allows manual data labeling for segmentation tasks. By utilizing a random forest algorithm, *ilastik* classifies pixels into user-defined groups with high accuracy.

We selected 10 patches from both groups and manually labeled background, nucleus, and cell regions. This method, in theory, should produce a tumor ratio distribution that more closely aligns with the results presented in the original paper. Our IHC segmentation results confirmed this, demonstrating the effectiveness of *ilastik* in achieving accurate tumor-to-tissue ratio measurements.



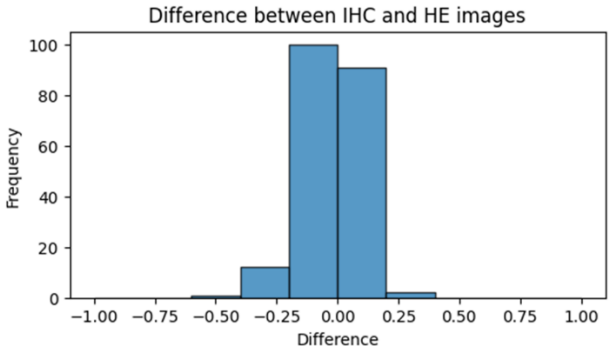
Top row, left to right: IHC patch, segmented result, tumor area ratio from 206 pairs

Bottom row, left to right: H&E patch, segmented result, tumor area ratio from 206 pairs

When labeling H&E-stained slides, we noticed that these slides contain numerous micro-metastases, which are notoriously challenging to detect, the labeling process can be subjective and

somewhat arbitrary. This subjectivity introduces variability, making accurate segmentation and classification more difficult. So, we calculated the difference between H&E and IHC results.

3.2 Accuracy of baseline



If we threshold the **difference at 0.05** between paired images, **accuracy is 0.7524**. If we relax the threshold to **0.1**, accuracy is at **0.8447**, and a further relaxation to **0.2** rendered an accuracy level of **0.9272**. **MSE for the 206 pairs is 0.0084, and MAE is at 0.0455**.

The results demonstrated relatively high accuracy compared to the original report. However, with our approach, the tumor ratio distribution was heavily skewed toward the first bin (0-0.2). As a result, the high accuracy does not necessarily indicate a superior method, as it reflects the imbalanced distribution rather than true robustness. Since our goal is to accurately determine the tumor ratio from H&E-stained slides, we opted to improve both accuracy and robustness using a different approach.

4. Generate IHC images from H&E images by CycleGAN

There have been numerous studies focused on detecting tumor areas in H&E slides. However, most of these approaches either provide binary outputs (tumor positive or tumor negative) or classify results into cancer stages based on the slides. To achieve more precise outputs, we propose generating IHC images from H&E slides for the following reasons:

1. Reduced dependence on expert annotation: Segmentation on H&E slides heavily depend on morphological features like nuclear size and shape, which can vary widely across tissues and cell types, requires experienced pathologists for manual annotation.

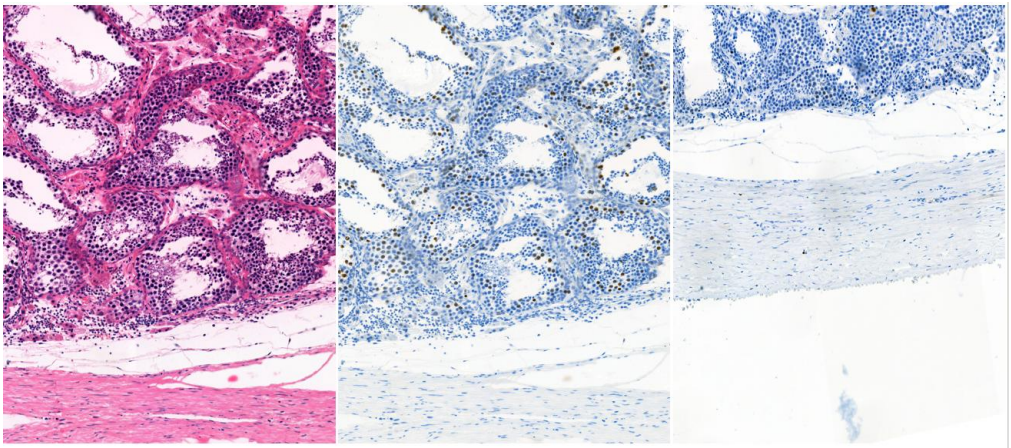
2. Alignment with ground truth of IHC: Direct segmentation on H&E lacks the direct correlation with such ground truth, making validation and comparison difficult.

We explored cycleGAN, a GAN-based model that does not require paired images for training, making it well-suited for our task. CycleGAN operates by translating images between two domains using a combination of loss functions:

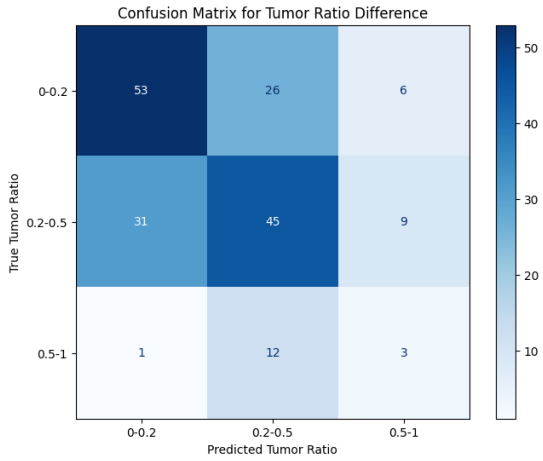
1. Cycle-Consistency Loss to maintain the overall structure and integrity of the image.
2. Identity Loss to preserves key features of the input.
3. GAN Loss

These features make CycleGAN particularly effective in our case where paired datasets are unavailable.

Here we cloned the repository https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix and trained 389 images from both groups. Limited to the T4 GPU offered by google colab, with 15GB memory, we cannot change the crop size for training. Therefore, different resizing settings were tried for the optimal results. Load size of 4096 without resizing ended up with the highest **accuracy of 0.6** and lowest **MAE of 0.13**.



H&E real IHC fake IHC real



Confusion matrix with 2048 pixels

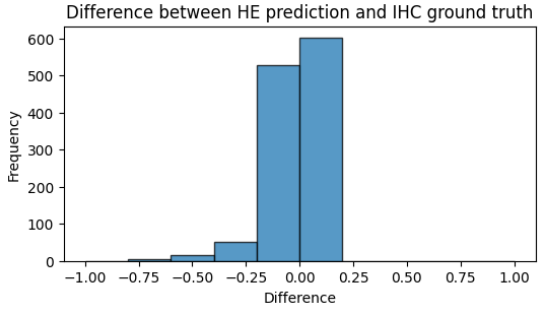
To improve accuracy further, we next compare the results to the original paper's ResNet18 model and Vit.

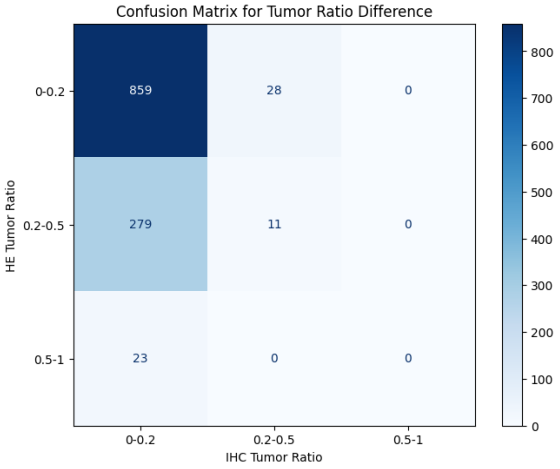
5. Ratio prediction with ResNet18 and ViT

First, we further tiled the paired images from 4098*4098 to 224*224 for typical ResNet and ViT inputs. After tiling 112 HE images from 206 pairs, we got patches from 49 slides. And to prevent data leakage, we split the samples into a 9:1 ratio, with 45 slides for training and 4 slides for testing. Training was sampled for 200, 200, 100 patches for bins of ratios of 0-0.2, 0.2-0.5, 0.5-1, aiming for a more balanced ratio. And testing patches were randomly sampled for 300 patches per slide.

5.1 ResNet18 results

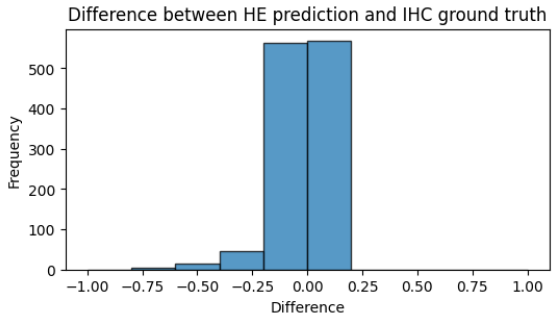
We followed the original paper and used 2 layers of fully connected networks of 512 nodes, 0.2 dropout rate, 0.9 momentum, and 0.0001 learning rate. And here, with over 18,000 training samples and 1,200 testing samples, we observed a drop of **MSE to 0.0129 and MAE to 0.078. Accuracy is 72.5%**.

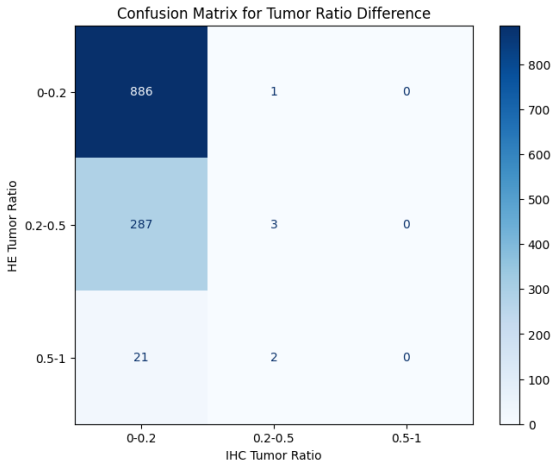




5.2 ViT B16 results

Here, we followed the previous setting, only switched the Relu activation with Gelu. This time, we observed a slight increase in **MSE (0.0126), MAE (0.0076), and Accuracy (74.1%).**





6. Conclusion

This project aims to demonstrate a practical pipeline for estimating tumor burden from H&E histology under weak supervision. By systematically comparing segmentation-based, generative, and discriminative

models, it clarifies the trade-offs between interpretability, robustness, and data requirements in computational pathology.

Here, we explored 4 different models: a simple pixel-level random forest segmentation with ilastik, a generative DL model of cycleGAN, a numeric prediction with ResNet18, and a numeric prediction with ViT B16. Since our input is skewed, the random forest segmentation is least trustworthy for categorization with random forest segmentation. ResNet and ViT both aimed for numeric prediction; therefore, the skewness is less severe as a problem, but it still affects the result. And cycleGAN is generative, so the skewness is the least of a problem.

From the accuracy and MAE readout from all models, ResNet and ViT both perform well, even without the pairing of inputs. This confirmed the original paper's choice of ResNet18 as the main model. However, MAE plateaued early during training (within 10 epochs for ViT), indicating pairing could still be the bottleneck here.

Future work will focus on improving H&E-IHC pairing at the patch level and exploring hybrid approaches that combine generative translation with direct regression. Such improvements are expected to further enhance quantitative accuracy while maintaining scalability for large histological datasets.

1. Petríková, D., Cimrák, I., Tobiášová, K., & Plank, L. (2024). Ki67 expression classification from HE images with semi-automated computer-generated annotations. In Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 1: BIOINFORMATICS (pp. 536-544). SciTePress. https://doi.org/10.5220/0012535900003657
2. Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2223-2232.