AIN 413 – Machine Learning For Healthcare

Course Project Report

Burak Kurt

2200765010

Evaluating the Impact of Data Augmentation on U-Net Models for Breast Cancer Segmentation

Abstract

This report investigates the application of U-Net architectures for breast cancer tissue segmentation using the Breast Cancer Semantic Segmentation (BCSS) dataset. The study evaluates the performance of custom-trained and pretrained U-Net models under different data augmentation scenarios. The models were assessed using the Jaccard score, calculated independently for each class (background, tumor, and stroma). Results indicate that non-augmented models generally perform better, with data augmentation enhancing performance only for the stroma class. Key challenges identified include the limited number of training examples and computational constraints. Future work should focus on increasing data diversity, extending training duration, and employing more complex models to improve segmentation accuracy.

1 - Introduction

Breast cancer is one of the most prevalent cancers worldwide, and accurate tissue segmentation is crucial for effective diagnosis and treatment planning. Semantic segmentation techniques, particularly those based on deep learning, have shown promise in automating this process. Among these techniques, the U-Net architecture has gained popularity due to its encoder-decoder structure with skip connections, which allows for precise localization and efficient training even with limited data.

This report explores the application of U-Net models for breast cancer tissue segmentation using the BCSS dataset, derived from The Cancer Genome Atlas (TCGA). The study compares the performance of a custom-trained U-Net model and a pretrained U-Net model, assessing the impact of data augmentation on segmentation accuracy. The goal is to determine the effectiveness of these models and identify areas for improvement to enhance their performance in real-world clinical settings. The methodology, experimental setup, results, and discussions presented in this report aim to provide a comprehensive understanding of the challenges and potential solutions in this domain.

2 – Methodology

2.1 Dataset

The dataset used for this project is the Breast Cancer Semantic Segmentation (BCSS) dataset, derived from The Cancer Genome Atlas (TCGA). It includes over 20,000 segmentation annotations of breast cancer tissue regions, resized to two formats: 224x224 and 512x512 pixels per image. The resizing to 224x224 pixels aims to enhance computational efficiency, while the 512x512 pixel format retains greater detail and accuracy. Annotations were created collaboratively by pathologists, residents, and medical students using the Digital Slide Archive.

While 512x512 images have much more detail and more classes in it, 224x224 format is used because of computational limitations. In 224x224 format, there are three unique classes: outside_roi (background), tumor and stroma. Distribution of the classes in training set is shown in Fig1.

Mask images (ground truth) are in 224x224x3 format where the last channel indicates the pixel's actual class. For example, at pixel [x,y], its valuıe could be [1,1,1] where 1 belongs to "tumor" class. One hot encoding is used for conversion of pixel values into vectors. After preprocessing, each pixel has value of [1,0,0] , [0,1,0] or [0,0,1], depending on the class they belong.

In many cases, datasets for medical imaging tasks like this one are limited in size and variability, which can hinder the ability of deep learning models to generalize well to unseen data. Data augmentation is a technique used to artificially increase the size and diversity of the training dataset by applying various transformations to the original images. Augmentation were applied randomly to the training images, creating a dataset with a mix of original and augmented images. This approach aimed to increase the diversity of the training data and improve the model's generalization ability.

As a final step of data preprocessing, all image pixel values were normalized to the range [0, 1] by dividing by 255.

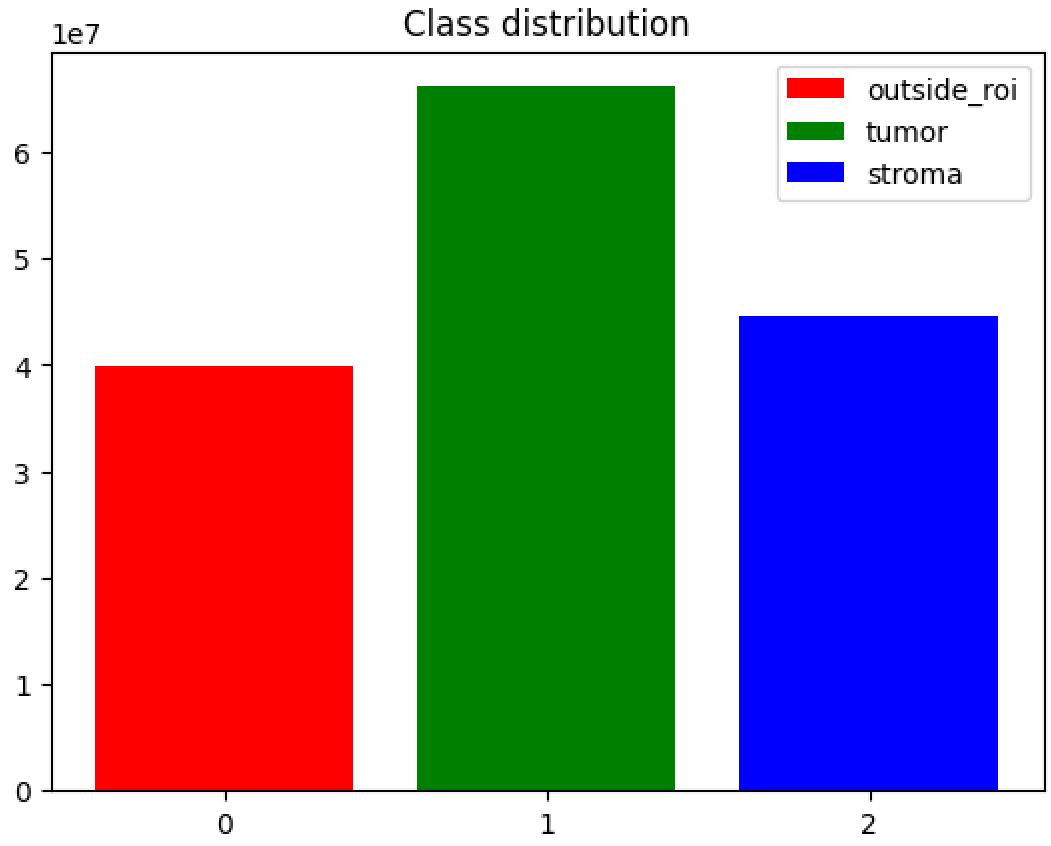


Fig 1 Class Distribution of Training Set

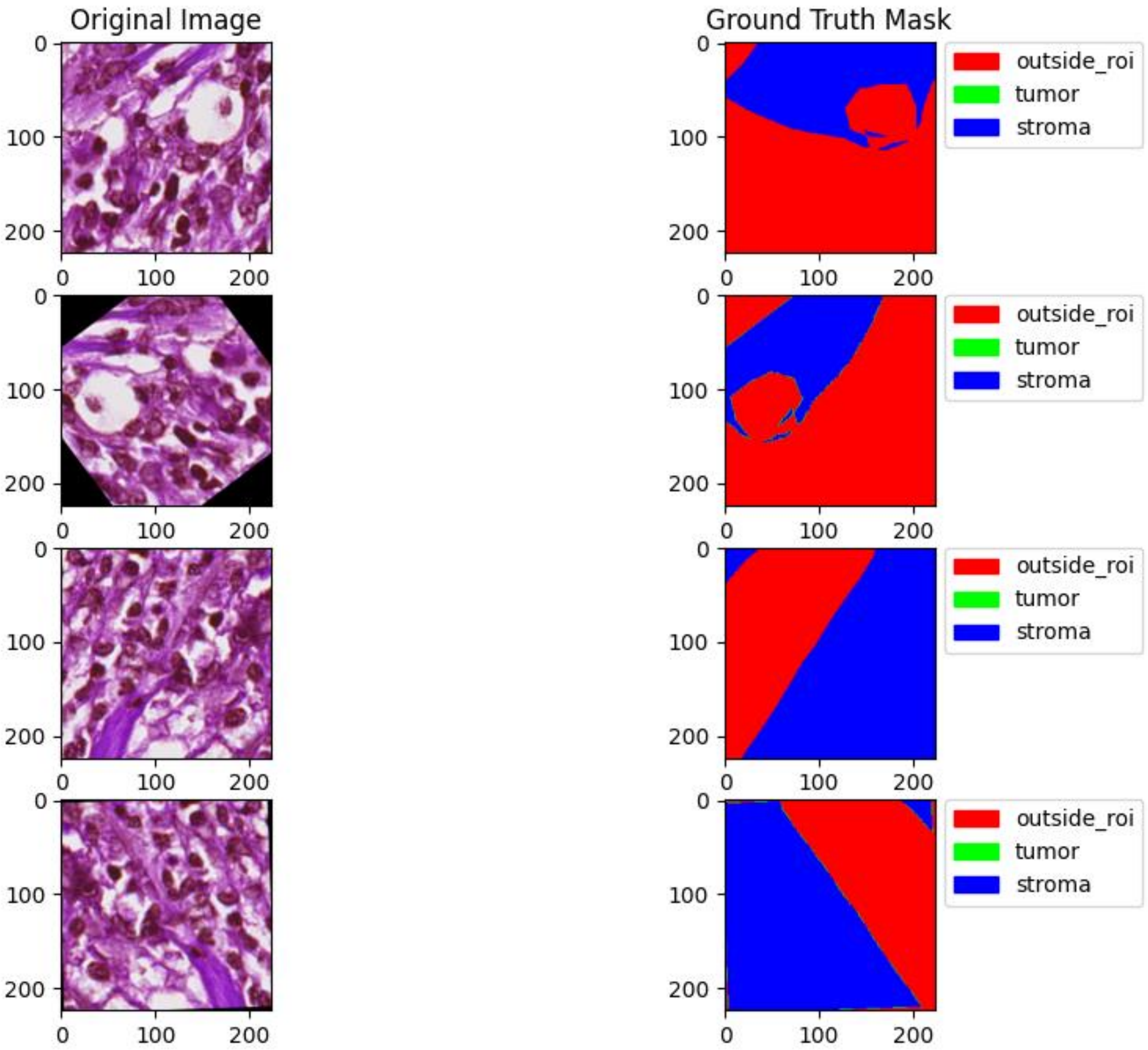


Fig 2 Example Images from Training Set (With Data Augmentation)

2.2 Model

A U-Net architecture was employed for pixel-level classification, implemented using PyTorch. The U-Net consists of an encoder-decoder structure with skip connections, which ensures the combination of low-level and high-level features.

The encoder part of the U-Net comprises a series of convolutional blocks (DoubleConv), each followed by a max-pooling layer to down-sample the feature maps. The decoder upsamples the feature maps using transposed convolutions and concatenates them with the corresponding feature maps from the encoder via skip connections. Further convolutional layers refine the upsampled feature maps to produce the final segmentation map. The bottleneck connects the encoder and decoder, applying convolutions to the lowest resolution feature maps before upsampling. A final 1x1 convolution reduces the number of output channels to the number of classes. A softmax function is applied to obtain the probability distribution for each class.

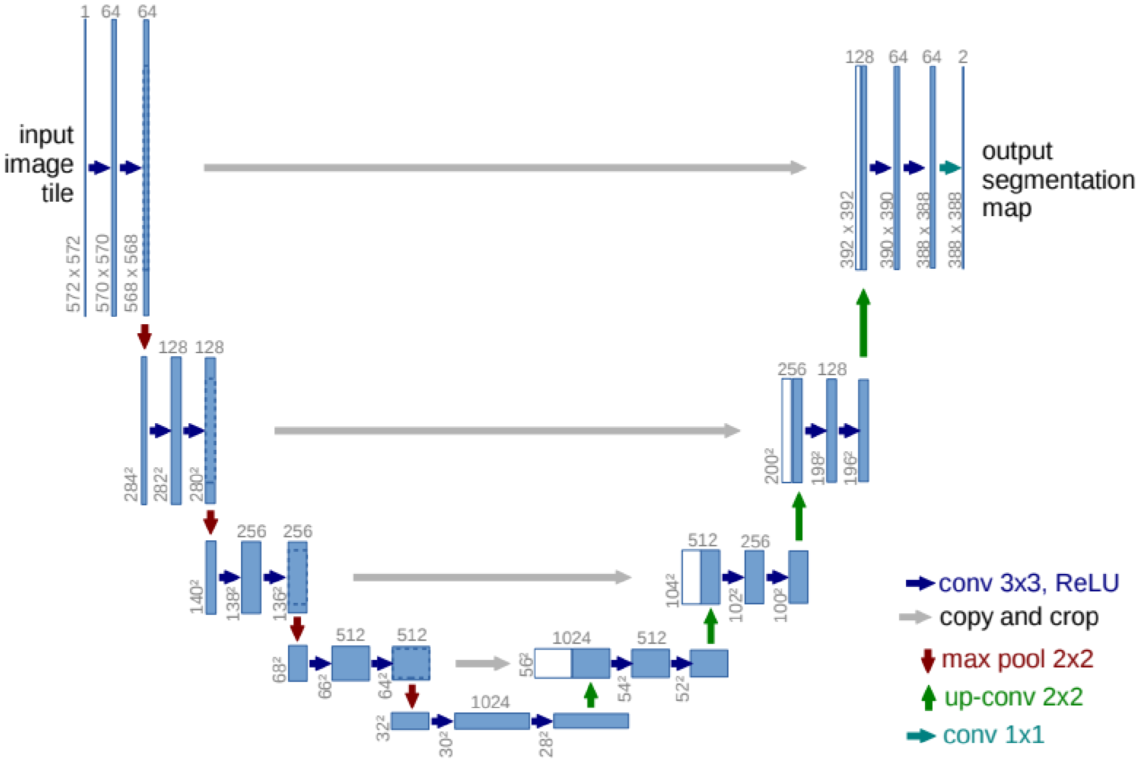


Fig 3 U-Net Architecture [1]

In addition to training a custom U-Net model, a pretrained U-Net model [2] was also trained on the dataset. Model is used for abnormality segmentation in brain MRI images. Since tasks are similar and there are enough data, it is suitable to train the whole network again, just using the initial weights from pretrained model. The aim was to compare the performance of the custom-trained model with that of the pretrained model to understand the benefits of transfer learning in this context.To be able to use that model in breast cancer data, output layer was changed and set as the number of the classes. The pretrained model underwent the same training and evaluation process as the custom model, ensuring a fair comparison between the two approaches. Details of the training process are discussed in Experiments part.

3 – Experiments

The primary objective of the experiments was to evaluate the performance of the U-Net architecture for breast cancer tissue segmentation under different data augmentation scenarios. Two main experimental setups were used to train and evaluate both a custom U-Net model and a pretrained U-Net model. Each setup utilized the Breast Cancer Semantic Segmentation (BCSS) dataset, resized to 224x224 pixels, and split into 5500 training images and 500 validation images.

In the first experimental setup, the training dataset comprised a mix of original and augmented images. Specifically, the training dataset included 2750 original images and 2750 augmented images, creating a total of 5500 training images. Augmentation techniques such as random rotations and flips were applied to the original images to generate the augmented images. This approach aimed to increase the diversity of the training data and potentially improve the model's generalization ability.

The second experimental setup used a non-augmented dataset where all 5500 training images were unique and original, with no augmentation applied. This setup aimed to evaluate the performance of the models when trained on a dataset with no additional artificial variability introduced through augmentation.

For both experimental setups, the same set of hyperparameters was used to ensure consistency and comparability between the results. The hyperparameters were as follows:

Learning Rate: 1e-5 (Custom Model) , 1e-4 (Pretrained Model)

Batch Size: 16

Loss Function: CrossEntropyLoss

Optimizer: Adam

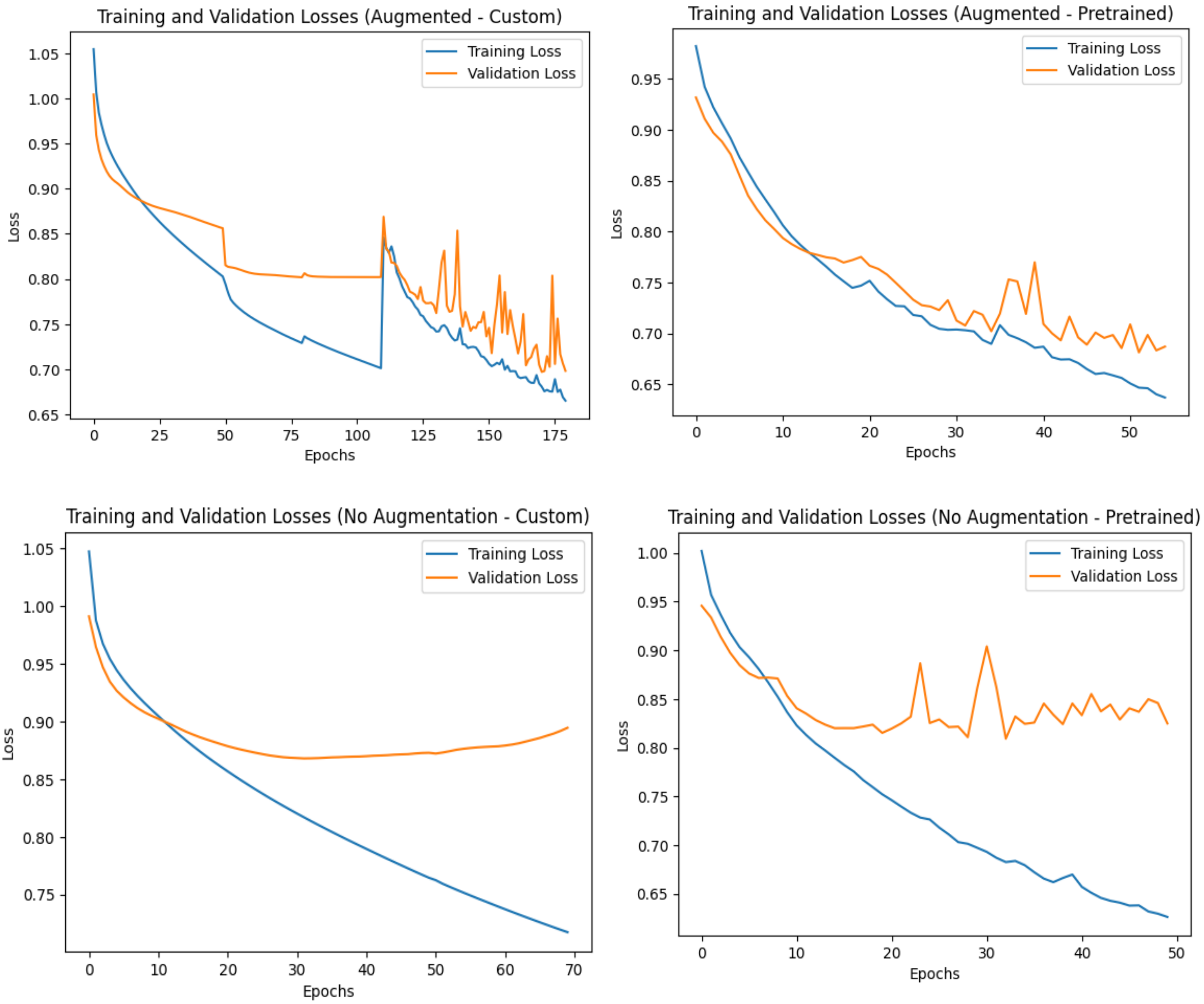
The learning rate of 1e-5 was chosen to allow the model to learn gradually, minimizing the risk of overshooting the optimal weights. A batch size of 16 was selected to balance between computational efficiency and the stability of gradient estimates. CrossEntropyLoss was used as the loss function due to its suitability for multi-class segmentation tasks, while the Adam optimizer was selected for its adaptive learning rate properties and efficiency in training deep neural networks.

In both experiments, the training procedure involved initializing the U-Net models (both custom and pretrained), the loss function, and the optimizer. The training dataset was then fed into the models in batches of 16 images. For each epoch, the models processed the training images, computed the loss using CrossEntropyLoss, and updated the weights using the Adam optimizer.

After each epoch, the models were evaluated on the validation set comprising 500 images. The validation loss was recorded to monitor the models' performance and detect any signs of overfitting or underfitting. Training and validation loss curves were plotted to visualize the learning progress over the epochs.

Table 1 Lowest validation losses for each model - data combination. Training losses obtained from epochs where validation loss is the lowest.

	Training Loss	Validation Loss
Custom – Augmented	0.66529	0.69830
Custom – No Augmentation	0.76458	0.87290
Pretrained – Augmented	0.63713	0.68719
Pretrained – No Augmentation	0.62627	0.82494



Graphs represents change in training and validation losses during training process. Custom-Augmented model takes too much epoch to converge compared to others. In No Augmentation models, after some steps, training loss is decreasing while validation loss is increases or stays still. That shows model is overfitting, so weights in the earlier epochs are used for final evaluation.

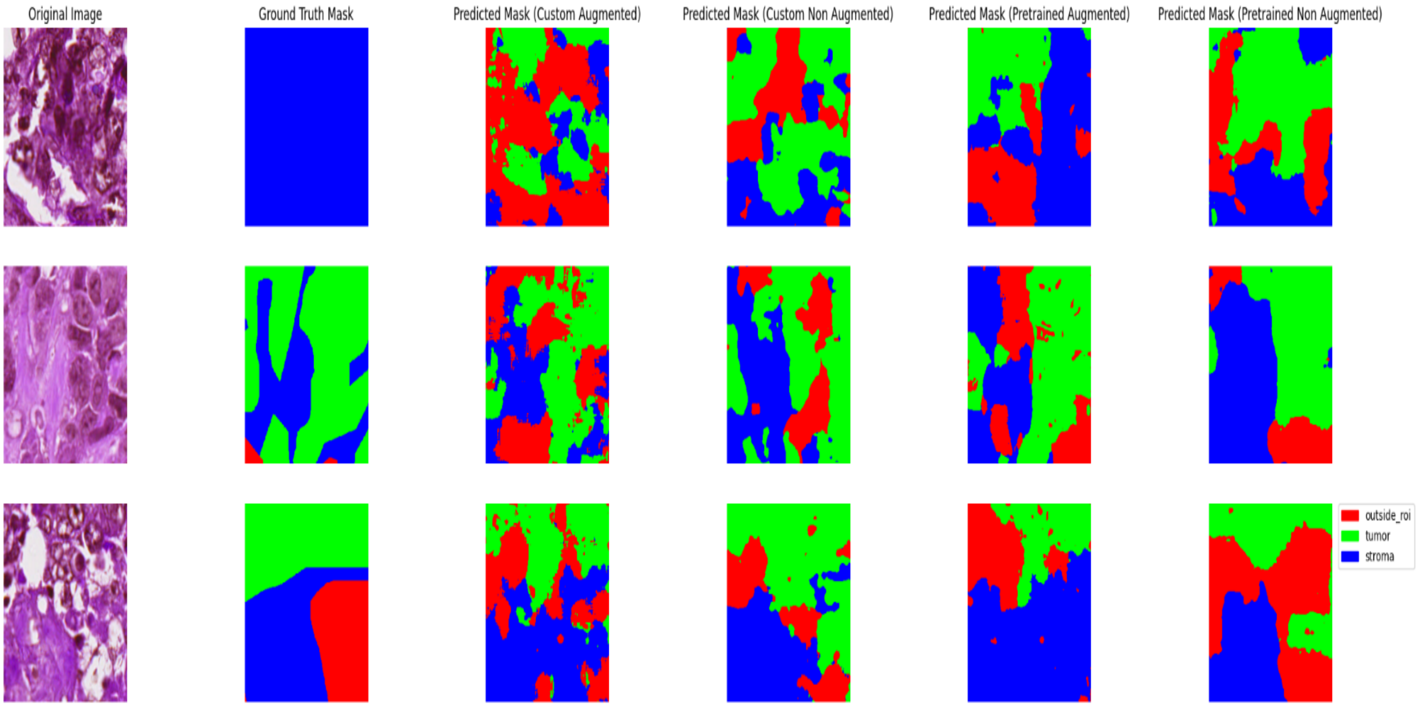

Original Image
Ground Truth Mask
Predicted Mask (Custom Augmented)
Predicted Mask (Custom Non Augmented)
Predicted Mask (Pretrained Augmented)
Predicted Mask (Pretrained Non Augmented)
outside_roi
tumor
stroma

Fig 4 Example of Training Images with Their Predictions.

4 – Results

The evaluation of the U-Net models for breast cancer tissue segmentation was carried out using the Jaccard score, a commonly used metric for assessing the accuracy of segmentation models. The Jaccard score, also known as the Intersection over Union (IoU), measures the similarity between the predicted segmentation and the ground truth. It is defined as the size of the intersection divided by the size of the union of the predicted and ground truth sets. The score ranges from 0 to 1, with 1 indicating perfect overlap between the predicted and true segmentation. For this project, the Jaccard score was computed independently for each class (background, tumor, and stroma) to provide a detailed evaluation of the model's performance. The results were obtained for both the custom U-Net model and the pretrained U-Net model, under two different data augmentation scenarios.

Fig 5 shows the Jaccard scores for each class (background, tumor, and stroma) for the custom and pretrained U-Net models under both augmented and non-augmented data scenarios.

For the Custom U-Net (Augmented), the tumor class has the highest Jaccard score, followed by stroma and background. Custom U-Net (Non Augmented) model follows same pattern with the augmented version but have higher scores for tumor and stroma classes compared to augmented version. Success of augmentation in background class could be related to more background pixel addition during augmentation. For the pretrained models, augmented version of pretrained model only has higher score in stroma class than non-augmented version. For tumor and background classes, non-augmented version outperforms. When all models are compared, non-augmented versions slightly perform better than setups with data augmentation.

Fig 6 compares the Jaccard scores for each class across all models. For the background (outside_roi) class, all models perform similar. For the tumor class, models with no augmentation perform better than models with data augmentation. For the stroma class, the pretrained U-Net with augmented data achieves the highest score, and the custom U-Net with no augmentation performs significantly better than with augmentation.

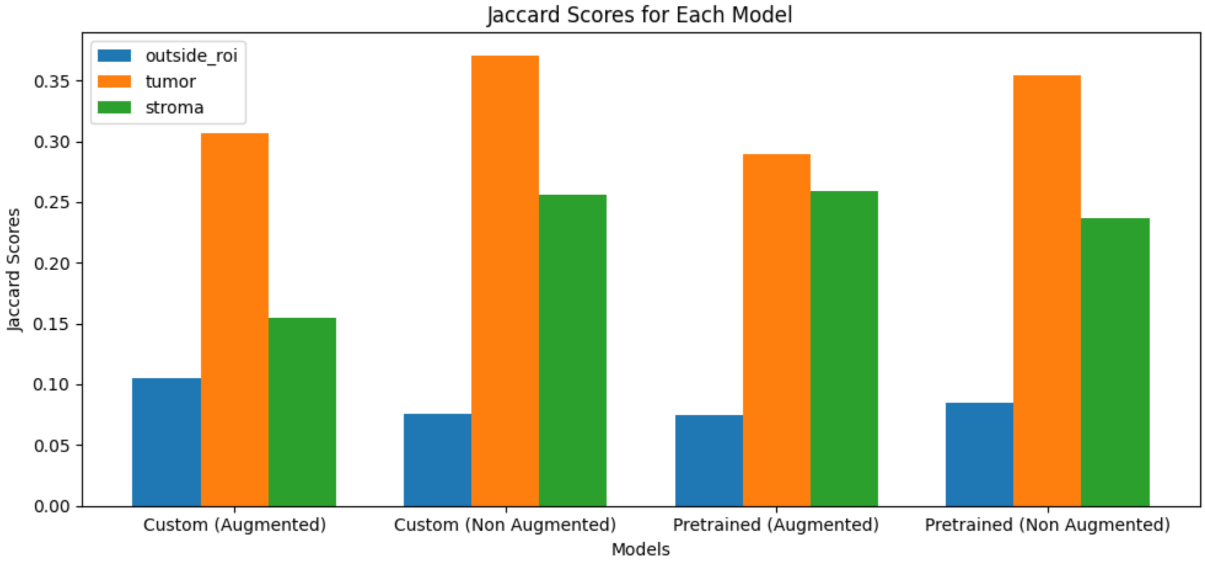


Fig 5 Jaccard Scores Based on Models

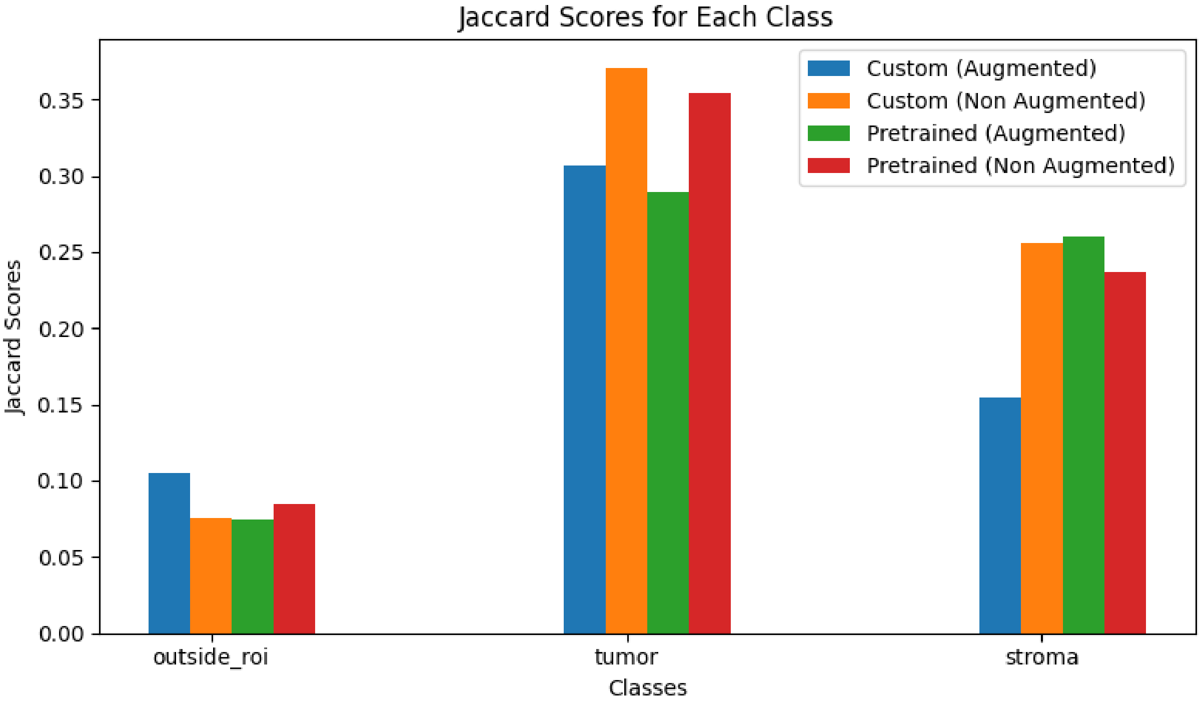


Fig 6 Jaccard Scores Based on Classes

The overall results indicate that data augmentation has a nuanced impact on the performance of U-Net models for breast cancer tissue segmentation. While data augmentation significantly improves the segmentation accuracy for the stroma class, it does not provide the same benefit for the tumor and background classes. In fact, for these classes, models trained without augmentation slightly outperform those trained with augmented data. This suggests that while augmentation can enhance model performance by increasing data diversity, it may also introduce noise that affects the segmentation accuracy for certain classes. The pretrained models generally perform better than the custom models, underscoring the value of transfer learning. However, the best performing model for each class varies depending on whether augmentation was used. Therefore, the choice of using data augmentation should be carefully considered based on the specific segmentation task and class characteristics.

5 - Discussion

The experimental results reveal several critical insights into the segmentation of breast cancer tissue using U-Net models under different training conditions. One of the key challenges identified was the presence of very low training examples, which inherently limits the ability of the models to generalize well. Given the dataset's size, with only 5500 training images, the diversity of the data might not be sufficient to capture the variability present in real-world scenarios. Choice of small batch sizes increase this problem because they may contain photos with only one class, which makes it more difficult for the model to learn during training rounds.

Moreover, the custom models were trained for fewer epochs than necessary due to computational limitations and time constraints. Extending the training duration would likely allow the custom models to achieve better convergence and performance. Additionally, the custom U-Net models, while functional, were relatively simple with smaller convolution layers due to the same computational constraints. Increasing the complexity of these models by adding more layers or larger convolutional filters could potentially enhance their learning capability and improve segmentation accuracy.

The impact of data augmentation was another crucial factor in this study. Although data augmentation is generally beneficial for increasing data diversity and preventing overfitting, it was implemented randomly in this project without accounting for class imbalance. It's possible that this randomly augmentation overrepresented classes and unintentionally included extra examples of already-dominant groups. As a result, the models trained with augmented data sometimes performed slightly worse than those trained without augmentation. This outcome suggests that more sophisticated augmentation techniques, which consider the class distribution, might be necessary to achieve optimal performance.

Future work should focus on several areas to enhance model performance. Firstly, increasing the dataset size through additional data collection or more effective data augmentation strategies that address class imbalance could provide the models with a richer learning environment. Secondly, training custom models for a longer duration and exploring more complex architectures could significantly improve their ability to learn and generalize from the data. Additionally, fine-tuning the augmentation process to ensure a balanced representation of all classes during training might mitigate the observed negative impacts of random augmentation.

6 – References

[1] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[2] https://github.com/mateuszbuda/brain-segmentation-pytorch