Automated Polyp Segmentation in Gastrointestinal Tract Images: A Comparative Study of U-Net and Residual-Unet Architectures

Nicolas Friley
nfriley@stanford.edu

Abstract - This study presents a comparative analysis of the U-Net and Residual-Unet (Res-Unet) architectures for the task of automated polyp segmentation in gastrointestinal tract images. The models were evaluated on the Kvasir-SEG dataset, a challenging medical imaging dataset containing 1000 polyp images and their corresponding segmentation mask. The quantitative evaluation showed that the Res-Unet models generally outperformed the standard U-Net models on the validation sets in terms of the Dice coefficient, a common metric for segmentation performance. The most performant model was the Res-Unet with Sigmoid activation and Binary Cross-Entropy (BCE) loss, which achieved the highest Dice coefficient on the validation set. However, the overall segmentation performance was limited, with the best Dice coefficient on the test set struggling to reach above 0.6. The analysis revealed that this was likely due to the small size of the dataset and the imbalanced representation of polyps with varying sizes and levels of obstruction. To address these limitations, the author recommends either merging the Kvasir-SEG dataset with additional similar datasets or performing extensive data augmentation. Additionally, experimenting with hybrid loss functions that combine pixel-level classification and region-level segmentation objectives could lead to more robust and clinically relevant segmentation models. These improvements could make the automated polyp segmentation systems more valuable for medical image analysis and decision support in healthcare applications.

---

1 – INTRODUCTION

The field of medical image analysis has seen significant advancements in recent years, thanks to the sudden progress in deep learning techniques, particularly for tasks such as MR images denoising, automatic disease classification, automatic labelling, etc. Another key application of deep learning in the medical field is the task of image segmentation, where the goal is to partition an image into relevant regions or objects. In the context of medical imaging, accurate segmentation of anatomical structures and pathological elements can have a big impact on disease detection, diagnosis, and treatment planning, such as surgery preparation.

One of the most common deep learning architectures used for segmentation tasks on medical images is the U-Net architecture, which is particularly well-suited for the segmentation of complex medical images, such as those obtained from endoscopic procedures [1]. One modified version of the U-Net, the Residual-UNet architecture, incorporates residual connections into the U-Net structure and has also shown promising results in medical image segmentation tasks [2].

Instead of focusing on segmenting images from MR or CT scanners, we chose to work on a dataset containing images from the gastrointestinal (GI) tract, containing anatomical landmarks and clinically significant elements called polyps. A polyp is an abnormal growth that varies in shape and size and develops on the lining of the colon or the rectum and protrudes from their inner lining (fig. 1). They develop due to an overgrowth of cells in the intestinal lining, with factors like age, diet and genetics increasing their risk of formation. Polyps are important because they can be precursors to colorectal cancer [3]. Screening tests like colonoscopies can detect polyps early. If found, polyps are usually removed to prevent them from potentially developing into colorectal cancer over time. Identifying and removing polyps early is a key strategy to prevent colorectal cancer, which is one of the leading causes of cancer-related deaths [4].

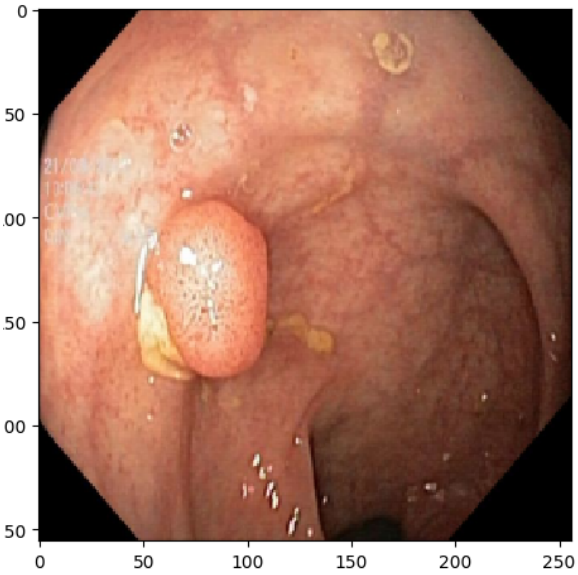

Fig. 1: Sample image of a polyp from Kvasir dataset

The use of automatic segmentation of pathological elements in images captured with endoscopic equipment could be particularly important for remote hospitals with little to no access to specialists in specific fields [5]. These hospitals or clinics could benefit from the performance of a model trained on careful manual segmentation from experts in the field, as it could provide valuable insights and support for early disease detection and diagnosis, ultimately improving patient outcomes in underserved regions.

2 – BACKGROUND

2.1 – Computer vision with medical images

Computer vision, a subfield of artificial intelligence, has proven to be a powerful tool for analyzing and interpreting medical images. The application of computer vision techniques to medical imaging data has revolutionized various aspects of healthcare, including disease diagnosis, surgery planning, image denoising and automated labeling. By leveraging deep learning models, computer vision systems can automatically extract valuable information from medical images, such as the presence of pathological structures, the segmentation of anatomical regions, and the quantification of specific features, volumes, and flow rates [6].

One of the main advantages of computer vision in the medical domain is to be able to handle large volumes of images quickly and efficiently. With the progress in medical imaging technologies, such as computed tomography and magnetic resonance imaging, the amount of available imaging data has increased exponentially. Computer vision algorithms can process these large datasets, identify patterns, and provide consistent and accurate analyses, which can assist healthcare professionals in making more informed decisions.

2.2 – Image segmentation

Image segmentation is an important task in computer vision, particularly in the context of medical image analysis. It splits an image into different regions representing a specific structure of interest. It can be used for tasks such as organ delimitation, lesion detection, and tumor characterization [6].

2.3 – U-Net Architecture

The U-Net architecture is a commonly used convolutional neural network model for image segmentation, particularly in the field of medical image analysis. It was developed by researchers at the University of Freiburg and consists of an encoder and a decoder. The U-Net architecture has been widely adopted and used for different medical imaging tasks, demonstrating its effectiveness in segmenting a wide range of anatomical structures and pathological regions.

2.4 – Res-Unet Architecture

The Res-Unet architecture is a variant of the U-Net model that incorporates residual learning, a technique that attempts to improve the performance and stability of deep neural networks. Residual learning was introduced in the ResNet architecture to try and address the problem of vanishing or exploding gradients that can occur in very deep neural networks [9]. In the context of medical image segmentation, the Res-Unet architecture combines the advantages of the U-Net structure with the benefits of residual learning. The integration of residual learning into the U-Net architecture has been shown to enhance the model's performance,

particularly in challenging medical imaging tasks, such as the segmentation of complex anatomical structures.

3 – DATA

3.1 – Dataset details

For this project's task of medical image segmentation, we focus on a dataset known as Kvasir, which was part of a Kaggle challenge on "Multi-Class Image-Dataset for Computer Aided Gastrointestinal Disease Detection". This dataset provides a valuable resource for researchers interested in developing automated systems for the detection and classification of various GI tract conditions [7].

This dataset is a comprehensive collection of images from the GI tract, collected using endoscopic equipment in gastroenterology departments of hospitals in Norway. The images have been carefully annotated and verified by medical experts, including experienced endoscopists from the Cancer Registry of Norway (CRN). The CRN is an independent institution under the Oslo University Hospital Trust, responsible for the national cancer screening programs and conducting research to prevent cancer deaths through early detection.

The Kvasir-SEG dataset is a subset of the Kvasir dataset, specifically focused on the task of polyp segmentation. It contains 1000 polyp images and their corresponding ground truth segmentation masks. The image resolutions in the Kvasir-SEG dataset range from 332x487 to 1920x1072 pixels. We used this subset of the Kvasir dataset for the scope of this segmentation project.

The Kvasir-SEG dataset presents a unique challenge due to the complexity of segmenting areas of different shape and size, sometimes disrupted by cropped squares masking patient-identifiable confidential information. Overcoming these challenges is crucial for the successful deployment of automated segmentation systems in real-world settings.

3.2 – Preprocessing

The images are resized to a fixed size of 256x256 pixels. For the input images, the original 3-channel color images are kept as-is, maintaining the RGB color representation. However, for the segmentation masks used as our labels, the original RGB masks are converted to grayscale in order to work with a single-channel binary mask. The grayscale mask is then expanded to have a single channel dimension, resulting in a 3D tensor with shape (256, 256, 1). The segmentation mask data is stored as Boolean values, where True represents the foreground (the polyp) and False represents the background (no presence of polyp).

4 – PROPOSED METHOD

4.1 – Architectures

4.1.1 – Unet

The U-Net architecture is characterized by its encoder-decoder structure and the use of skip connections, which is commonly represented in a U-shape to easily visualize the interaction of the skip connections between the encoder and the decoder (fig. 2). This infrastructure allows the model to combine local and global information to produce accurate segmentation results [8].

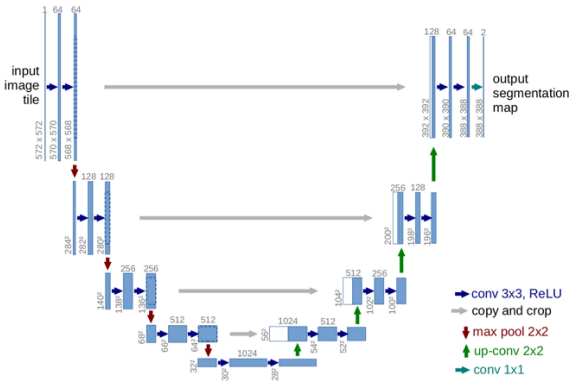

Fig. 2: Layout of the U-Net architecture [8]

The encoder part of the model represents the contracting path. It is responsible for extracting relevant features from the input image. It

consists of a series of convolutional layers, max-pooling layers, and dropout regularization. The convolutional layers learn to extract features, and the pooling layers reduce the spatial dimensions of the feature maps to try and capture the hierarchical representation of the input. The model we used has 5 encoder blocks.

The bridge layer connects the encoder and the decoder branches of the network. This set of convolutional layers strive to capture the most abstract features from the input and applies dropout regularization to prevent the model from overfitting.

The decoder part of the model represents the expanding path. It reconstructs the segmentation map from the features extracted by the encoder. It is made of a series of up-sampling blocks with convolutional layers and dropout regularization. The up-sampling increases the spatial dimensions of the feature maps. The convolutional layers learn to combine the up-sampled feature maps with the skip connections to produce the final segmentation map. Our model has 5 decoder blocks.

The skip connections concatenate the feature maps from the encoder layers with the corresponding feature maps in the decoder layers, which allows the model to retain important spatial information that could have been lost during the downsampling process otherwise. Our model has 5 skip connections.

4.1.2 – Residual-Unet (Res-Unet)

This architecture is very similar to the classic implementation of the U-Net architecture. The main difference is the addition of residual connections in the encoder and decoder branches (fig. 3). The residual connections create shortcut connections that bypass the convolutional layers within the encoder and decoder blocks and add the input of the block directly to the output of that same block. This improves the gradient flow by providing an alternate path for the gradients to flow in case the model faces an issue of vanishing gradient [9]. As a result, these residual connections help increase the capability of the model to propagate important features through the network, which is supposed to help improve the performance for our task of image segmentation.

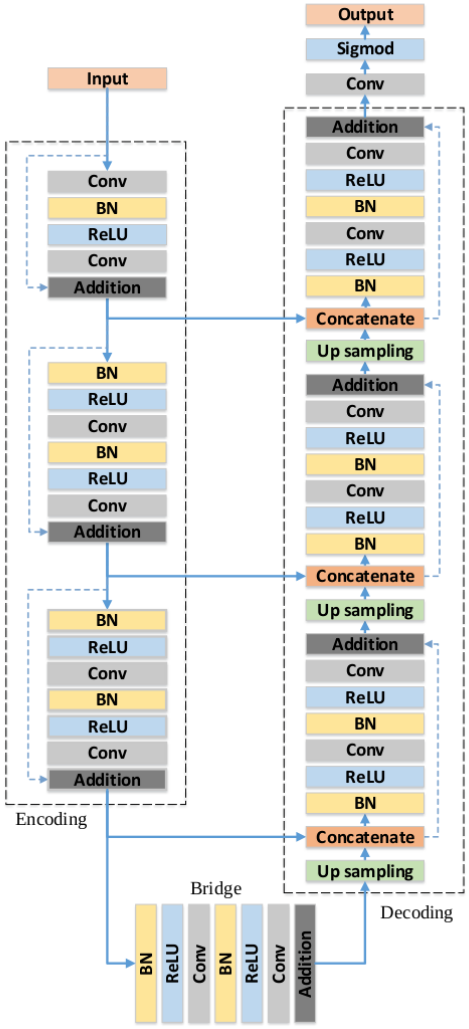

Fig. 3: Layout of the Res-Unet architecture [10]

4.2 – Loss Functions

In the scope of this project, we experimented with two different loss functions. The Binary Cross Entropy (BCE) loss, and the Soft Dice Loss.

BCE loss is a loss function that is well-suited for segmentation tasks because it directly optimizes the pixel-wise classification accuracy. Essentially, it encourages the model to correctly predict the class label (polyp or background) for each pixel, which is crucial for medical image segmentation task where we attempt to precisely define lesions or pathological elements.

The other loss function we considered was the Soft Dice Loss function. This loss function is specifically designed to optimize the Dice Coefficient, which is our key metric for this project. By directly maximizing the overlap between the predicted segmentation and the ground truth, the Soft Dice loss tries to ensure that the model prioritizes the accurate segmentation of the clinically relevant regions.

4.3 – Metrics

Accuracy is a common metric that measures the proportion of correctly classified pixels (or voxels for 3D imaging such as MRI) out of the total number of pixels in the image. In the context of medical image segmentation, accuracy represents the overall agreement between the predicted segmentation mask and the ground truth segmentation.

However, the Dice coefficient is a metric that is more robust and specifically designed for tasks such as segmentation [11]. This is why we will base our quantitative analysis on Dice coefficient comparisons. The Dice coefficient measures the overlap between the predicted segmentation and the ground truth segmentation. It is defined as follow:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

The Dice Coefficient ranges from 0 to 1, where 1 indicates perfect overlap, and 0 indicates no overlap. The Dice Coefficient is particularly useful in medical image segmentation tasks because it focuses on the accuracy of the segmentation of the target regions, rather than the overall image accuracy. This makes the Dice Coefficient more sensitive to the performance of the model on the clinically relevant regions, which is the primary concern for our application.

4.4 Output Layer Activation Function

For this project, we experimented with 3 different output layer activation functions: the sigmoid function, the arctangent function, and the cumulative distribution function of the normal distribution (CDF).

The sigmoid function is a commonly used activation function that maps any input value to a value between 0 and 1. It is often used for binary classification tasks, where the output represents the probability of an input belonging to one of two classes, in our case: polyp or background. The sigmoid function is defined as:

$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

The sigmoid function has a smooth, S-shaped curve, making it suitable for modeling probabilities or introducing non-linearity in neural networks.

The arctangent function maps any input value to a value between -π/2 and π/2 (approximately 1.57 and 1.57). This activation function is often used when the output needs to be bounded within a specific range, such as when predicting angular quantities or circular data [12]. We defined the arctangent function as follow:

$$atan\_act(x) = \varepsilon + (1\text{-}2 \times \varepsilon) \times \frac{0.5 + atan(x)}{\pi}$$

The small constant epsilon is added to the function for numerical stability.

Finally, the normal CDF maps any input value to a value between 0 and 1, representing the probability that a random variable drawn from a standard normal distribution (with mean 0 and standard deviation 1) is less than or equal to the input value. This activation function can be useful when the output needs to represent a probability or a value that is naturally bounded between 0 and 1 [13]. We defined the CDF as:

$$CDF(x) = (0.5 \times \varepsilon) \times erf(\frac{x}{\sqrt{2}}) + 0.5$$

The eft function is the error function.

4.5 - Learning Rate

The choice of learning rate was made after performing hyperparameter optimization on 5 epochs. We used the maximum dice coefficient obtained for all 7 models used, to identify which learning rate to use for each model combination. Every model saw better performance with a learning rate of 5e-5.

4.6 - Early Stopping

Using the Unet architecture with certain combinations of loss functions and output activation function would see the loss function increasing somewhere around 50 epochs (fig. 4). For this reason, we decided to implement early stopping and evaluate the models' performance at 50 epochs.

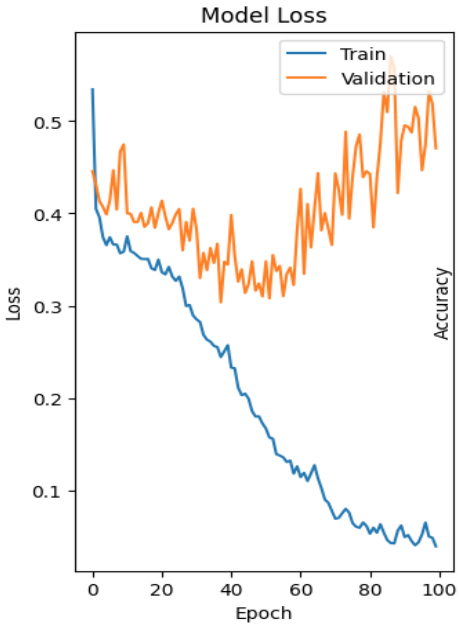

Fig. 4: Training and validation loss of the U-Net model over 100 epochs

4.7 - Models Presentation

To evaluate the performance of different architectural choices for polyp segmentation, we implemented and trained 7 different models. The models varied in their architecture, output activation function and loss function used during training. The first 3 models were based on the U-Net convolutional neural network architecture. The remaining 4 models were build using the Residual U-Net architecture. The 7 different models are defined below (table 1).

1	U-Net	Sigmoid	BCE
2	U-Net	Normal CDF	BCE
3	U-Net	Sigmoid	Soft Dice
4	Res-Unet	Sigmoid	BCE
5	Res-Unet	Normal CDF	BCE
6	Res-Unet	Sigmoid	Soft Dice
7	Res-Unet	Arctangent	Soft Dice

Table 1: Models presentation (Left to right: model number, model architecture, output layer activation function, loss function)

5 - RESULTS

5.1 - Quantitative evaluation

5.1.1 - Loss

The validation loss for two models (4 and 5) starts going back up around the 20th epoch, indicating that these models are starting overfitting (fig. 6). This is likely due to the small size of the dataset (only 1000 examples) combined with the complexity of the models. As noted earlier, we had already chosen to implement early stopping after noticing that most models saw a similar increase in loss after the 50th epoch. The small dataset size coupled with the models complexity seems to be leading to overfitting, which may lead the models to struggle to generalize well beyond the training data.

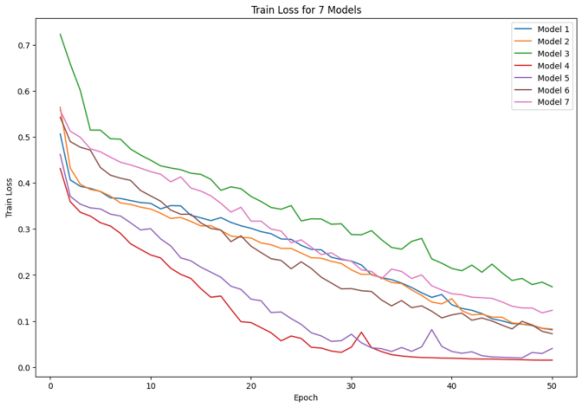

Fig. 5: Training losses of all 7 models over 50 epochs

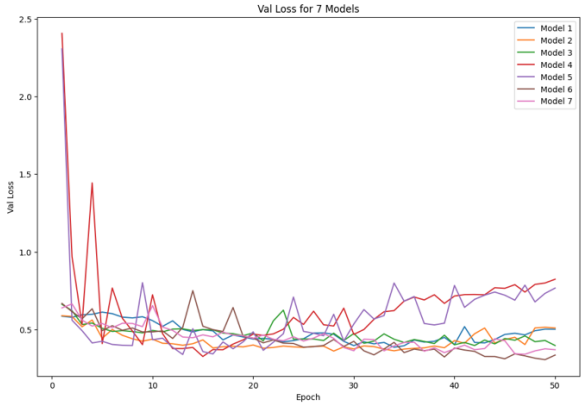

Fig. 6: Validation losses of all 7 models over 50 epochs

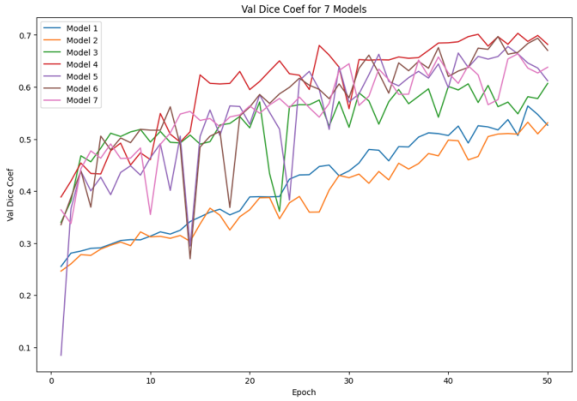

Fig. 8: Dice coefficients for all 7 models on the validation set over 50 epochs

5.1.2 - Dice coefficient

The 4 Res-Unet models' Dice coefficient increase more rapidly than the 3 U-Net models' Dice coefficient, on both the training and validation sets (fig. 7 & 8). Although the Dice coefficient generally increases with the number of epochs, the values show a lot of variability every 5 epochs. Therefore, the final Dice coefficient values collected should be treated as approximate indicators of the models' performance (table 2).

After 50 epochs of training, 3 of the Res-Unet models outperform all U-Net models while the 3rd U-Net model maintain a similar performance to that of model 5 (which is a the lowest performing Res-Unet model on the validation set).

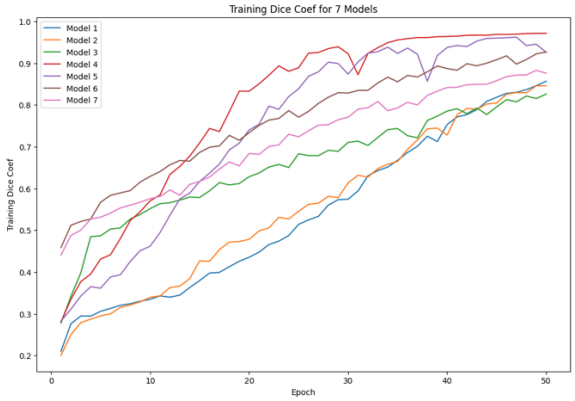

Fig. 7: Dice coefficients for all 7 models on the training set over 50 epochs

The most performant model appears to be model 4, which is the Res-Unet with the Sigmoid activation function and the BCE loss (table 2). This suggests that the Res-Unet architecture, combined with the appropriate loss function and activation, can lead to improved segmentation performance compared to the standard U-Net model, even with the limited dataset size.

Model #	Validation Dice Coefficient
1	0.5282
2	0.5553
3	0.6132
4	0.6817
5	0.6118
6	0.6702
7	0.6379

Table 2: Final Dice coefficient on the validation set for all 7 models after 50 epochs of training

5.1.3 - Performance on test set

Models 1, 3 and 5 perform on the test set as well as they did on the validation set, indicating good generalization. However, models 2, 4, 6 and 7 perform slightly less well on the test set, suggesting they have more difficulty generalizing beyond the validation set (table 3). This is likely due to the small dataset size and potential distribution shifts between the training, validation, and test sets, which would have been made more prominent given the small size of the dataset. Among the 3 best performing models (3, 5 and 6), two of them are the U-Net and Res-Unet, both with Sigmoid activation and Soft Dice Loss

function, suggesting this to be a strong combination of design choices on this dataset. regardless of the architecture choice.

Model #	Test Dice Coefficient
1	0.5512
2	0.4530
3	0.5877
4	0.5474
5	0.5820
6	0.5823
7	0.5492

Table 3: Average Dice coefficient for all 7 models on the test set

5.2 - Qualitative evaluation

Overall, the qualitative performance of all models appears similar for the well predicted masks across all models, and similar for the wrongly predicted masks across all models. The well predicted masks show clear contours (fig. 9), sometimes with small areas being misclassified with low confidence as can be seen with parts that are less opaque than the predicted mask (fig. 10). The models excelled at capturing the contours of elliptical shapes that were entirely visible in the images, but struggled with partially obstructed polyps, where the obstruction was due to the cropping box in the lower left corner (fig. 11).

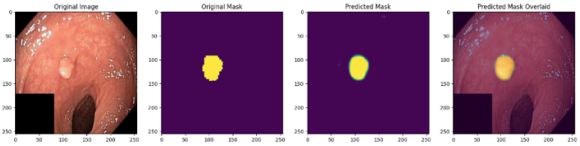

Fig. 9: Example of correctly predicted mask (Left to right: original image, original mask, predicted mask, original image overlaid with predicted mask)

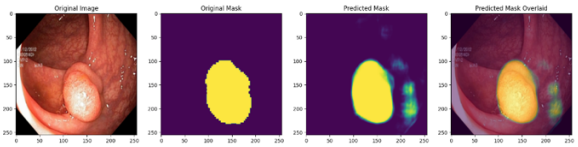

Fig. 10: Example of correctly predicted mask with small bleed out prediction to the right (Left to right: original image, original mask, predicted mask, original image overlaid with predicted mask)

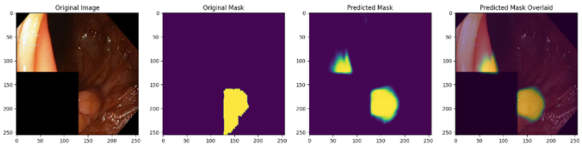

Fig. 11: Example of incorrect predicted mask due to cropping box (Left to right: original image, original mask, predicted mask, original image overlaid with predicted mask)

As for the wrongly predicted masks, they either overshot by representing a larger area than the actual polyp or undershot by only covering a small portion of the polyp. Another example of misclassification happened when the label segmentation was made of 2 or more separate shapes, in which case the models tended to build a mask for only one of the shapes, with some overshooting (fig. 12).

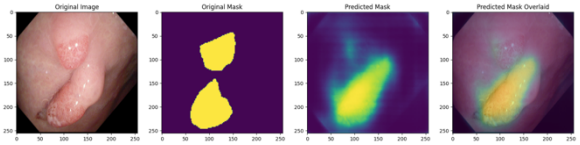

Fig. 12: Example of incorrect predicted mask due to label segmentation containing more than one shape (Left to right: original image, original mask, predicted mask, original image overlaid with predicted mask)

These errors could be due to class imbalance, not in the sense of "polyp" vs. "background", but rather as groups representing polyps with different scale factors and levels of obstruction. Some images had polyps taking up most of the image, while others had polyps that were only small circular shapes in the image. Others had partially obstructed polyps due to patient privacy protection cropping. If these special cases were underrepresented in the data, it could lead to a form of class imbalance that the models struggled to overcome.

The models' performance was also related to whether the polyp's contour was too close or even touching the edge of the image. In those cases, the models tended to undershoot or overshoot the segmentation (fig. 13).

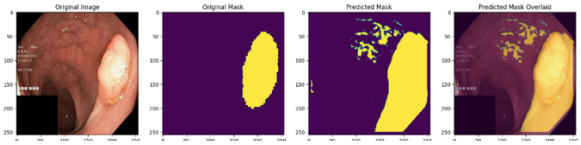

Fig. 13: Example of incorrect predicted mask due to polyp being very close to the edge of the image.

6 - Conclusion

In this experiment, we were able to demonstrate how Residual U-Net models outperformed regular U-Net models for a segmentation task on medical images. While the comparison allowed identifying the improved performance of the Res-Unet models over the U-Net models, the best Dice coefficients had trouble reaching above 0.6. The main issue was the small size of the dataset (only 1000 examples), and the imbalanced representation of cases where the polyps were either very large, very small, very close to the edge of the image, or partially obstructed.

To improve the results, we would recommend either merging this dataset with another similar dataset of segmented images from endoscopic procedures or performing extensive and careful data augmentation on the current dataset. By careful, we mean that only certain forms of image transformation would be considered valid given the nature of the dataset: horizontal and vertical flips, or combination of both would be acceptable. Slight changes in brightness intensity would also be acceptable. However, any form of cropping or rotation that isn't by a multiple of 90 degrees might result in the polyps being cropped out of the image, since many images have the polyps very close to the edge of the image, or appear already very zoomed in. Any transformation that would make the image quality slightly worse, such as blurs or too drastic a change in brightness intensity should be avoided, to avoid adding representational shift in the dataset, since most of the images appear very bright and clear. We also suggest manually separating the images by "classes" of polyp representation (size, obstruction level, proximity to the edge of the image, etc), and performing data augmentation in a way that resolves the possibility of class imbalance within the dataset. This careful data augmentation implementation would not only increase the number of examples but also help fix the class imbalance mentioned earlier.

Experimenting with different loss functions and output layer activations yielded noticeable differences, showing that tuning these design choices based on the dataset can be fruitful. One continuation of this work could be to implement a combination of the BCE loss and the Soft Dice loss as a hybrid loss function that encourages both accurate pixel-level classification and region-level segmentation. This complementary approach could lead to more robust and clinically relevant segmentation models, making them valuable tools for medical image analysis and decision support in healthcare applications.

7- Acknowledgements

We thank Nikhil Tomar for his example implementation of the Res-Unet architecture [14] and K. Pogorelov et al. for making the KVASIR dataset available to the public.

References

[1] N. Punn and S. Agarwal. Modality specific u-net variants for biomedical image segmentation: a survey, 2022.
[2] F. I. Diakogiannis et al. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 2020.
[3] Ali, Sharib, et al. "Assessing Generalisability of Deep Learning-Based Polyp Detection and Segmentation Methods through a Computer Vision Challenge." Scientific Reports, vol. 14, no. 1, 2024.
[4] Huck, Michael, and Jaime Bohl. "Colonic Polyps: Diagnosis and Surveillance." Clinics in Colon and Rectal Surgery, vol. 29, no. 04, 2016.
[5] Kazem shahmoradi, Mohammad, et al. "Evaluation of Colonoscopy Data for Colorectal Polyps and Associated Histopathological Findings." Annals of Medicine and Surgery, vol. 57, 2020.
[6] C. L. Chowdhary and D.P. Acharjya. Segmentation and feature extraction in medical imaging: A systematic review. Procedia Computer Science, 2020. International Conference on Computational Intelligence and Data Science.
[7] KVASIR Dataset for Segmentation. https://www.kaggle.com/datasets/abdallahwagih/kvasir-dataset-for-classification-and-segmentation

[8] O. Ronneberger, et al. Unet: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.
[9] K. He, et al. Deep residual learning for image recognition, 2015.
[10] Z. Zhang et al. Road Extraction by Deep Residual U-Net. IEEE Geoscience and Remote Sensing Letters, 2017.
[11] Müller, Dominik, et al. "Towards a Guideline for Evaluation Metrics in Medical Image Segmentation." *BMC Research Notes*, U.S. National Library of Medicine, 2022.
[12] J. Kamruzzaman, Arctangent activation function to accelerate backpropagation learning. IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences, 2002
[13] L. Nieradzik et al. Effect of the output activation function on the probabilities and errors in medical image segmentation, 2021.
[14] Nikhil Tomar's Github: nikhilroxtomar. https://github.com/nikhilroxtomar/Deep-Residual-Unet/