Advancing Vasculature Segmentation in 3D Human Tissue using Deep Learning: Insights from SenNet + HOA Dataset

1 Introduction

The Vasculature Common Coordinate Framework (VCCF) provides a critical reference system for understanding cellular interactions within tissues by focusing on blood vessels. This framework is vital for enhancing our knowledge of various physiological functions and diseases. However, manually labeling vascular structures is time-consuming and prone to inconsistencies. Advanced deep learning techniques can be used to automate the process of vascular segmentation.

The SenNet + HOA - Hacking the Human Vasculature in 3D comes from a Kaggle competition. This competition, run by the Common Fund's Cellular Senescence Network (SenNet) Program and the Human Organ Atlas (HOA), is dedicated to advancing the segmentation of blood vessels in human tissues, with a particular focus on the kidneys. [3]

The goal of this work is to greatly improve the accuracy and efficiency of vascular segmentation in human kidney tissues by using cutting-edge image segmentation techniques to the SenNet + HOA - Hacking the Human Vasculature in 3D dataset.

Image segmentation is a core challenge in computer vision, playing a vital role in many visual understanding systems. This process involves dividing images or video frames into different segments or objects, and is integral to a wide array of applications.

The task of image segmentation can be approached in various ways, including semantic segmentation, which assigns a category label to each pixel; instance segmentation, which identifies and outlines each distinct object; and panoptic segmentation, which combines elements of both semantic and instance segmentation.

Historically, a variety of algorithms have been developed to tackle image segmentation, starting from basic methods like thresholding and region-growing, to more sophisticated techniques like k-means clustering, watershed methods, and graph cuts. Advances in technology have seen the rise of deep learning, which has revolutionized image segmentation. Deep learning models have significantly outperformed earlier methods, leading to major improvements in accuracy and setting new benchmarks in the field, marking a significant shift in the methodologies employed.[5]

For the task of this project, semantic image segmentation methods utilizing deep learning techniques are utilized for binary segmentation, as the provided images and masks in the dataset are grayscale. The models trained for this task are variations of the following architectures adapted for the task:

- **ResNet (Residual Network)**: This convolutional neural network utilizes skip connections and residual learning principles, allowing for the training of significantly deeper networks.[2]
- **U-Net**: A widely recognized architecture in the field of medical image analysis, U-Net features a U-shaped design consisting of an encoder and a decoder connected by a bridge. This fully convolutional network is particularly tailored for semantic segmentation tasks.[7]
- **Vision Transformers (ViT)**: Building on the success of Transformers in large language models, the ViT was introduced by researchers at Google. This model eschews traditional convolutional layers, instead applying the Transformer architecture with self-attention mechanisms directly to sequences of image patches, representing a novel approach in image analysis.[1]

Each architecture brings unique strengths to the table— ResNet's deep learning efficiency, U-Net's specialized design for medical imaging, and ViT's innovative use of Transformers. The objective of this study is to find the most suitable model for the blood vessel segmentation task. This project promises to have significant practical impacts, particularly in enhancing diagnostic processes and improving patient outcomes.

2 Literature Review

Segmentation plays a crucial role in the analysis of natural and medical images, aiding in scene understanding, image-guided interventions, radiotherapy, and enhanced radiological diagnostics. This technique is defined as the division of an image into distinct non-overlapping regions that collectively cover the entire image. Image segmentation has seen significant advances through deep learning across various medical imaging modalities like X-ray, MRI, PET, CT, and ultrasound. Recent research has concentrated on improving deep learning architectures to address issues like gradient problems, optimize model size for efficiency, and enhance performance through new optimization techniques. Although earlier efforts utilized shallow networks and superpixel maps for segmentation, current approaches focus on end-to-end trainable deep neural networks. Innovations in neural architecture, such as different layer types, depths, and configurations, have driven these advancements.[4] Notably, encoder-decoder networks like U-Net, residual-based models like ResNet and attention-based models like ViT have been pivotal.

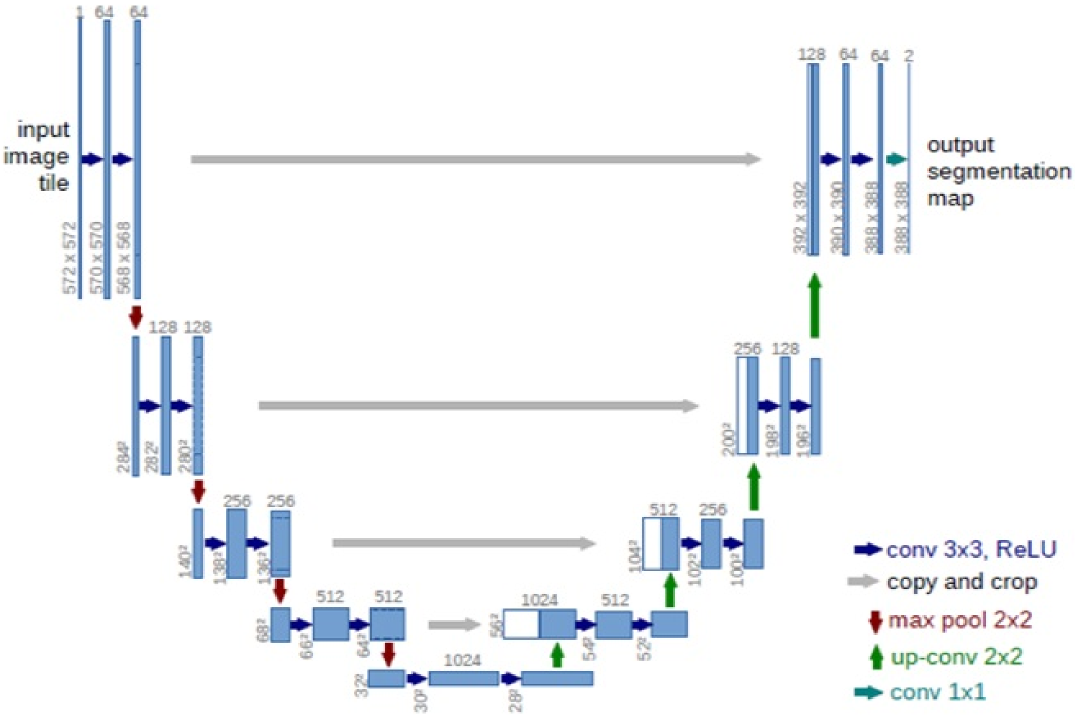


Figure 1: Architecture of U-Net.[7]

2.1 UNet

U-Net was first proposed in 2015 by a group of researchers at University of Freiburg, Germany particularly focusing on medical image segmentation and using IOU as the main metric. The network architecture *(Figure 1)* features a U-shaped design with a contracting path on the left and an expansive path on the right. The contracting path follows a conventional convolutional network structure, employing two 3x3 convolutions per step, each followed by a ReLU activation and a 2x2 max pooling with a stride of 2 for down sampling, doubling the feature channels at each step. The expansive path includes up sampling of the feature map, a 2x2 convolution to reduce feature channels, followed by a concatenation with the matching cropped feature map from the contracting path, and two subsequent 3x3 convolutions with ReLU. The final layer uses a 1x1 convolution to classify each 64-component feature vector into the desired classes. The original architecture has 23 convolutional layers and was designed to enable seamless tiling of the output segmentation map, critical for processing large images. The architecture lacks fully connected layers and emphasizes valid convolutions, allowing for segmentation of arbitrarily large images through an overlap-tile strategy that compensates for missing context at the image borders by mirroring the input. This ensures high resolution without being constrained by GPU memory limitations [7].

2.2 ResNet

ResNet was also proposed in 2015 by researchers at Microsoft as a way to ease the training of deep neural networks. They introduced a framework for residual learning involving layers to learn residual functions with reference to the inputs instead of learning from scratch. These residual networks were tested up to 152 layers deep on the ImageNet dataset—eight times deeper than VGG networks—demonstrating not only ease of training but also improved accuracy due to their depth. The profound depth of these networks has been pivotal in visual recognition tasks, leading to a 28In the residual learning framework, instead of the traditional approach of expecting each stack of layers to directly approximate a desired function, they approximated a residual function. Specifically, for a desired function H(x), the layers are set to approximate the function F(x) = H(x) - x. This change transforms the original function into F(x) + x. They proposed that optimizing for this residual function is simpler than optimizing for the original function directly. Particularly, if the ideal function is an identity mapping, it's easier to drive the residual toward zero than to approximate an identity mapping through several nonlinear layers.

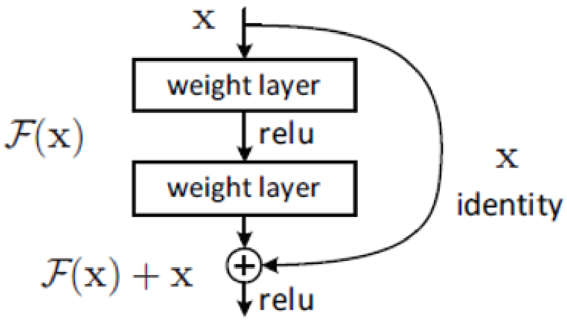

x
weight layer
$\mathcal{F}(\mathrm{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathrm{x}) + \mathrm{x}$
relu

Figure 2: Building block of residual learning.[2]

This concept is operational in neural networks through "shortcut connections" that skip one or more layers, as depicted in *Figure 2*. These shortcuts perform an identity mapping, and their outputs are added to those of the stacked layers, without introducing additional parameters or increasing computational complexity. The entire network is trainable end-to-end using standard techniques like stochastic gradient descent (SGD) with back propagation and can be implemented with commonly used frameworks like Caffe, without requiring changes to the underlying solvers. In their research, the authors enhanced a basic network by integrating shortcut connections, which transformed it into a residual network. These shortcuts, employed when dimensions are consistent, involve either padding with zeros or using 1x1 convolutions for dimension matching. This setup allows shortcuts across varying feature map sizes with a stride of 2.[2]

2.3 Vision Transformers (ViT)

Vision Transformers (ViT) was developed in 2021 by researchers at Google. The team experimented with applying a standard Transformer directly to images with minimal modifications by dividing the images into patches, treating them similarly to word tokens in NLP. This method involved embedding these patches and training the model for image classification in a supervised manner. They observed that while Transformers trained on mid-sized datasets like ImageNet without strong regularization achieved lower accuracies than comparable ResNets, the results significantly improved on larger datasets. For example, the Vision Transformer (ViT) excelled when pre-trained on larger datasets like ImageNet-21k or the in-house JFT-300M dataset, matching or surpassing the state of the art on multiple benchmarks. [1]

The Vision Transformer's design was noted for having less image-specific inductive bias compared to CNNs, relying on learned global self-attention mechanisms instead of built-in translation equivariance and locality. The hybrid architecture approach also allowed the integration of CNN-generated feature maps into the Transformer framework. This led to an enhanced ability of the model to handle images of varying resolutions during fine-tuning by interpolating pre-trained position embeddings based on their original location in the image. This innovative approach marks a significant shift in handling image recognition tasks, bridging methodologies from NLP to computer vision.[1]

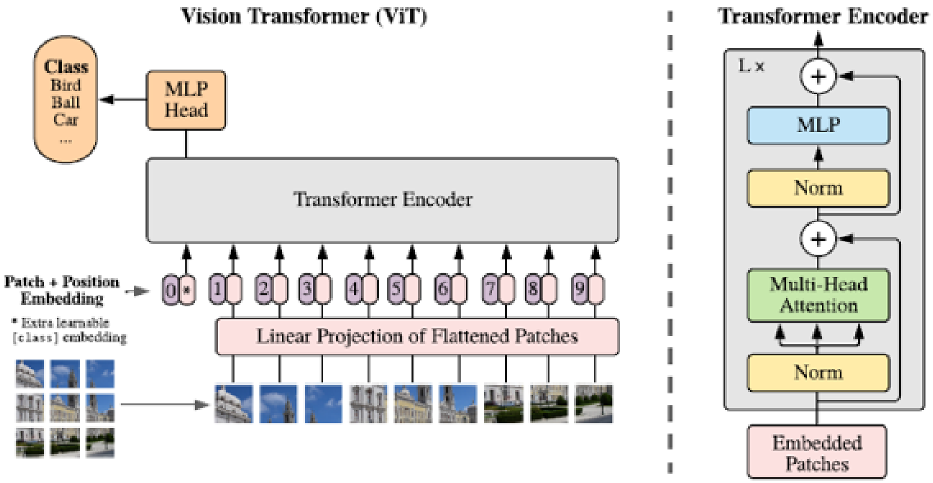


Figure 3: Overview of vision transformer architecture.[1]

3 Method

The SenNet + HOA - Hacking the Human Vasculature in 3D dataset is collected from Kaggle which was published as a research code competition. The main objective of the competition was to segment blood vessels. In this work, the same objective was adopted. The dataset for this competition includes high-resolution 3D images of various kidneys, accompanied by 3D segmentation masks that highlight their vascular structures. These kidney images were captured using Hierarchical Phase-Contrast Tomography (HiP-CT), an imaging method that provides high-resolution three-dimensional data of ex vivo organs, ranging from 1.4 micrometers to 50 micrometers in resolution.

3.1 File information and competition requirements:

The competition dataset is structured as follows: train/kidney name/images: This directory contains TIFF scans from various kidney datasets. Each file in this directory is a 2D slice of a 3D kidney volume, taken along the z-axis. The slices are ordered from top to bottom and should be stacked depth-wise for analysis. train/kidney name/labels: This directory includes TIFF files of blood vessel segmentation masks corresponding to the images in the same dataset. The specific dataset folders include:

kidney_1_dense: This contains the complete 3D image of a right kidney at 50um resolution, with a comprehensive segmentation of the arterial vascular tree extending to two generations from the glomeruli (capillary bed), captured using beamline BM05.

kidney_ 1_ voi: A high-resolution subset of the kidney_1_dense dataset, imaged at 5.2um resolution.

kidney_2: Features the entire kidney from a different donor at 50um resolution, with segmentation covering approximately 65% of the structure.

kidney_3_dense: Consists of 500 slices from a portion of a kidney at 50.16um resolution using BM05, with dense segmentation. It does not contain any images; it only contains a label folder.

kidney_3_sparse: Contains all the remaining segmentation masks and images for kidney_3, with about 85% of the kidney sparsely segmented.

test/kidney_name/images: This directory holds the TIFF scans for the test set, which may feature different beamlines or resolutions compared to those in the training set. The datasets named here are kidney_5 and kidney_6. This directory only contains test images but no label images [3].

The original requirement of the competition was to submit RLE (Run-length encoding) which locates foreground objects in segmented images [3].

3.2 Adopted Methodology

To conduct the study, data was curated from four distinct folders provided under the train folder: kidney_ 1_dense, kidney_2, kidney_1_voi, and kidney_3_sparse. These directories provided a substantial compilation of image/mask pairings, totaling 2279, 2217, 1397, and 1035 pairs, respectively. Notably, the kidney_3_dense folder was excluded as it contained only labels with the corresponding images already incorporated within the kidney_3_sparse folder. The amassed dataset comprised 6928 grayscale image/label pairs for analysis.

Given that the test folders were devoid of labeled images, a portion of the training set was allocated to serve as a test set. This was essential to facilitate a robust evaluation, allowing to compare the predicted masks against true annotations.

For the assessment of segmentation performance, a suite of metrics was included. The Dice Similarity Coefficient (DSC) and Intersection-Over-Union (IOU) are used as the primary indicators. The DSC, also known as the F1 score, offers an insightful measure of the alignment between the predicted object boundaries and the actual ones, with scores ranging from 0 to 1. It is prevalently used in medical image segmentation and captures a balance between precision and recall. The IOU further complements the evaluation by stringently penalizing discrepancies in segmentation. Along with Dice score and IoU, accuracy, precision, recall, and f1 score were also measured. These carefully selected metrics provide a comprehensive and rigorous framework for assessing the performance of the proposed segmentation methodologies.[6]

$$\text{IoU} = \frac{TP}{TP + FP + FN} = \frac{\text{area of overlap}}{\text{Area of Union}} \tag{1}$$

The Dice Similarity Coefficient (DSC) is given by:

$$\text{DSC} = \frac{2 \cdot TP}{TP + FP + FN} = \frac{2 \cdot (\text{Area of overlap})}{\text{Total Area}} \tag{2}$$

4 Experimental Setup

In this investigation, the dataset was initially structured as a data frame to streamline the manipulation process. The pre-processing of images and corresponding masks involved the following:

- Reading the images in grayscale.
- Resizing them to a uniform dimension of 256x256 pixels.
- Transforming them into numpy arrays.
- Scaling the pixel values by dividing them by 255.
- Dimension expansion to conform to the input requirements of deep learning models.
- For the masks, a threshold was set at 0.5 to ensure precise segmentation, considering that the provided masks were in binary format (black and white)

After pre-processing, the images and masks were systematically segregated into separate datasets for the imaging and masking components.

For model evaluation, 10% of the data was partitioned to create a test set, utilizing the train-test split function from the sci-kit-learn library. This resulted in a total of 6235 image-mask pairs allocated for training and 693 pairs designated for testing. Given the ample volume of data available for training, any augmentation techniques were forgone in the pre-processing stage.

The data processing protocol was meticulously standardized across the board for all models trained, ensuring uniformity in the dataset provided to each.

4.1 U-Net based Architecture

The model based on the U-Net architecture included several blocks. The core component of the model is given below:

Convolutional Blocks: Performs feature extraction through repeated 3x3 convolutions, each followed by batch normalization and ReLU activation, enhancing training stability and incorporating non-linearity.

Encoder (Downsampling Path): Captures context and reduces spatial dimensions via max pooling, while deepening feature representation to encode higher-level semantic information.

Decoder (Upsampling Path): Recovers spatial dimensions and detail through transposed convolutions and concatenations with corresponding encoder outputs (skip connections), essential for precise pixel-level predictions.

Bottleneck: Located at the deepest part of the network, bridging the encoder and decoder. It processes the most abstract representations, containing two convolutional layers without any pooling.

Output Layer: Maps the features back to the desired output dimensions using a 1x1 convolution.

Activation: Utilizes a sigmoid function for binary outputs and softmax for multi-class segmentation tasks.

Compilation: For optimizing the model Adam optimizer was used with a learning rate of.0001, and binary cross entropy was used as a loss function. A validation dataset was created using a 0.2 split.

Training: The model showed a substantial loss at the 4th epoch and was finally trained for 4 epochs with the train data separated before with a validation split of 0.2%.

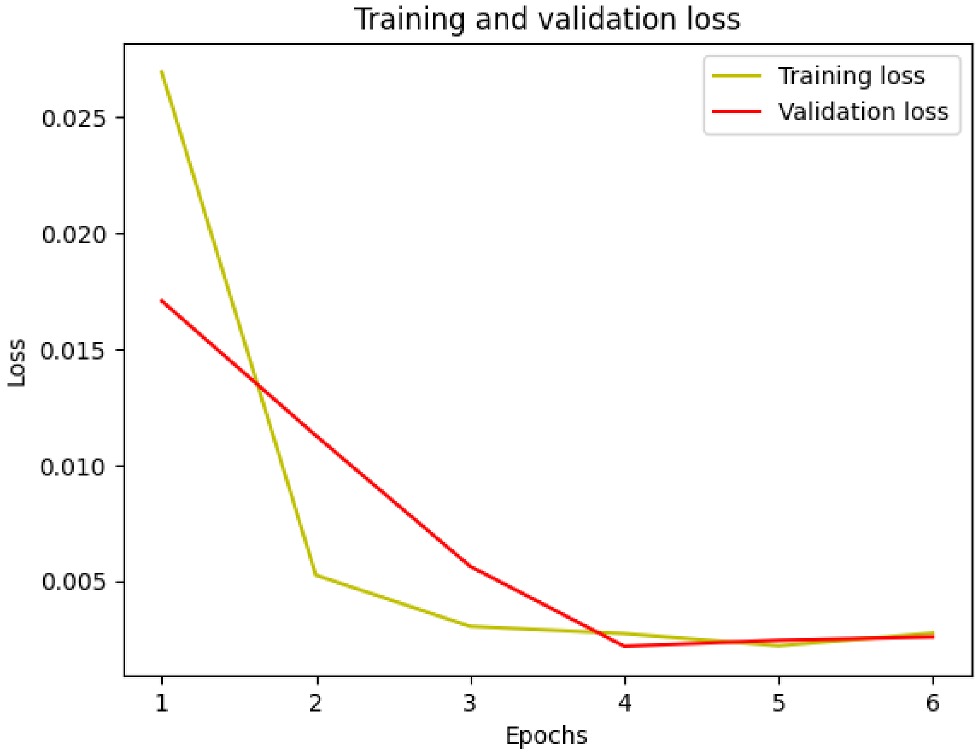


Figure 4: Training and validation loss curve for U-Net-based architecture

4.2 ResNet based Architecture

This is a convolutional neural network (CNN) based on the ResNet architecture, which is widely used for image recognition tasks. It includes residual blocks that allow for training deeper networks by addressing problems like vanishing gradients. Following is a breakdown of its architecture:

Input Layer: It takes an input shape. The input shape was 256,256,1.

Initial Convolution: Before entering the residual blocks, the model applies a convolutional layer with a 7x7 kernel, a stride of 2, and padding to maintain the spatial dimensions. This layer also includes L1 and L2 regularization to prevent overfitting.

Batch Normalization and Activation: Follows the initial convolution to standardize the outputs and apply a ReLU activation function for non-linearity.

Residual Blocks: The core of the network consists of a series of residual blocks. Each block applies two convolutional layers with a 3x3 kernel, each followed by batch normalization. ReLU activation is used after the first batch normalization, and dropout is applied to help with regularization.

Implements a skip connection that adds the input of the residual block to its output, which helps in training deeper networks by allowing gradients to flow directly through the network.

The number of these blocks is configurable with the variable num_blocks.

Upsampling: After processing through the residual blocks, the network uses a transposed convolution layer to increase the spatial dimensions of the output.

Output Layer: Finally, a 1x1 convolution is applied to reduce the number of output channels to 1, suitable for binary segmentation tasks. It uses a sigmoid activation function to output probabilities indicating the presence of a class at each pixel.

Compilation: The model is compiled with the Adam optimizer and binary cross-entropy loss, suitable for binary classification tasks. Accuracy is tracked as a metric.

Training: This model was also trained for 4 epochs with the same training and validation data with a 0.2 validation split.

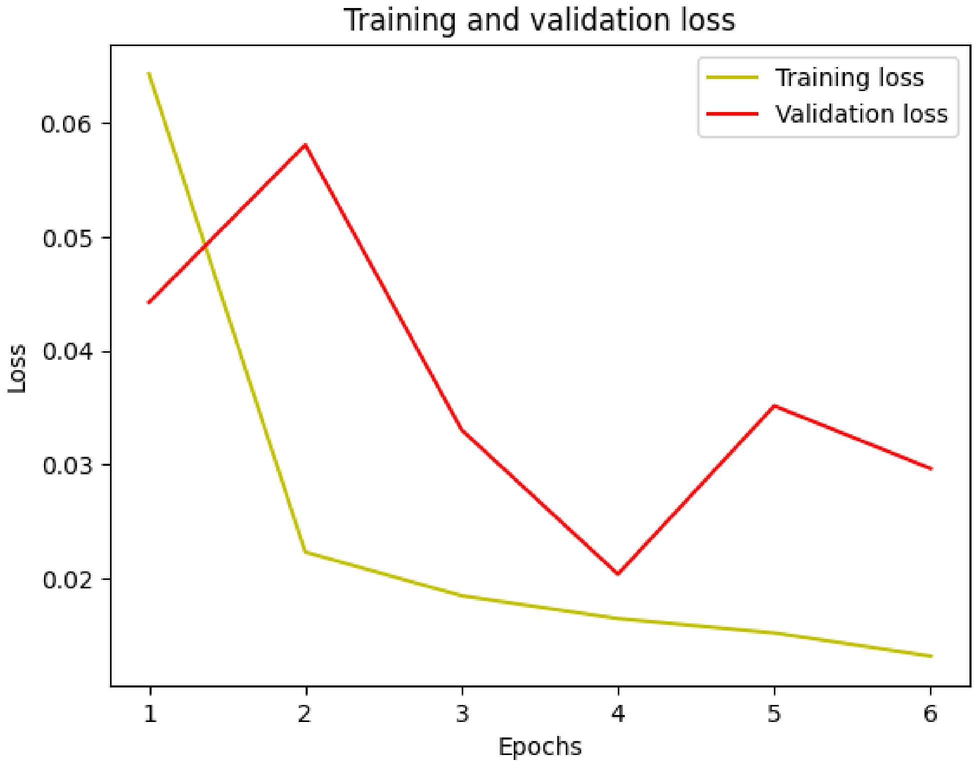


Figure 5: Training and validation loss curve for ResNet-based architecture

4.3 Vision Transformer Based Model

This model is comprised of a mixture of transformers and CNN architecture utilizing multi-head attention. The model included:

Input Layer: Accepts a single-channel (grayscale) image with dimensions of 256x256x1 pixels. However, the original model was trained on 224x224x3 shaped color images.

Patch Extraction: The Patches layer divides each image into 16x16 pixel patches, reducing the image into a sequence of smaller, manageable data pieces.

Patch Encoding: A Patch Encoder layer linearly transforms each patch into a 64-dimensional space and adds positional embeddings to retain the order information of the patches.

Transformer Layers: Comprising eight transformer blocks, each block includes a layer normalization, a multi-head attention mechanism with four heads, and a feed-forward network (MLP). These layers are designed to capture complex dependencies between different image regions.

Upscaling via Transposed Convolutions: Following the transformer blocks, a series of transposed convolution layers gradually increase the spatial resolution of the feature maps. The number of filters is halved sequentially through the layers to construct the final image segmentation map.

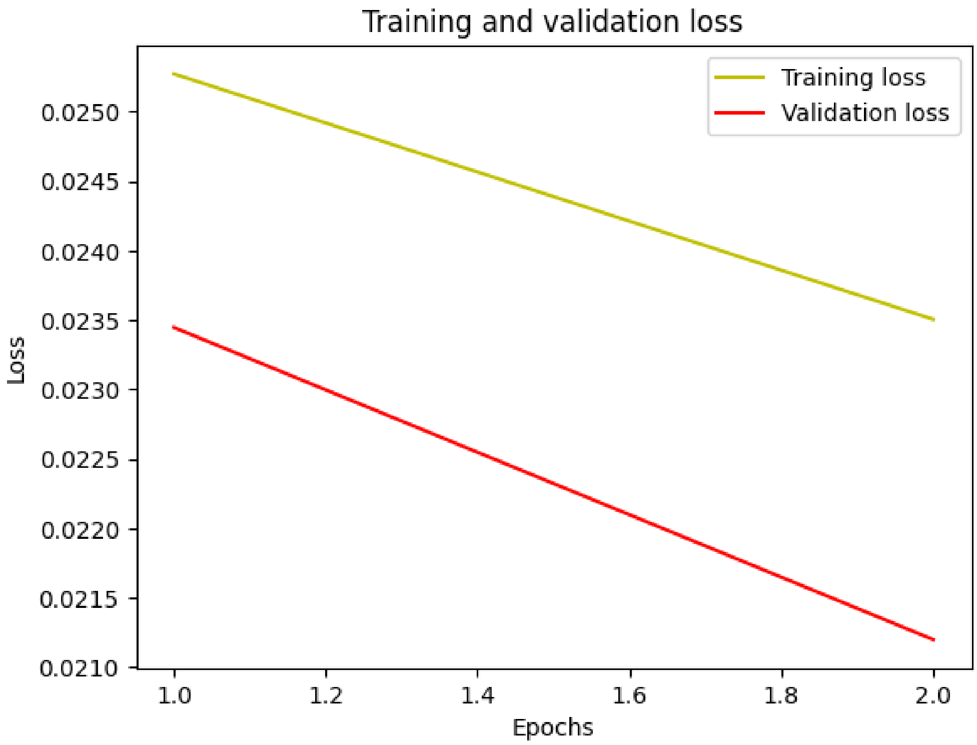


Figure 6: Training and validation loss curve for ViT-based architecture

Output Layer: A 1x1 convolution with sigmoid activation produces a final output, a binary mask that indicates the presence of the target class at each pixel.

Hyperparameters Hyperparameters used for this model are:

Patch Size: Chosen to effectively balance detail against computational efficiency.

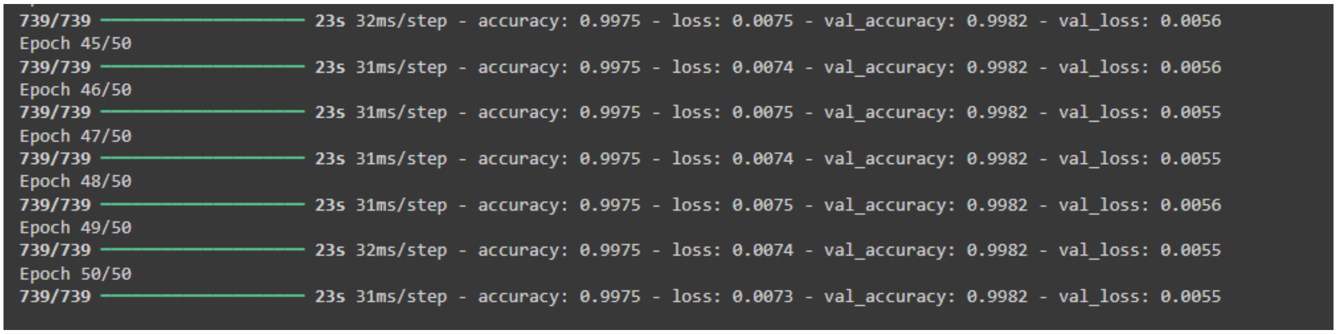
Number of Transformer Heads: Set to four, allowing the model to attend to various features at each position.

Learning Rate and Weight Decay: Adjusted to prevent overfitting while ensuring sufficient training speed

Compilation: The model is compiled with the Adam optimizer and binary cross-entropy loss, suitable for binary classification tasks. Accuracy is tracked as a metric, and a learning rate of 0.001 and weight decay of 0.0001 were used.

Training: The model was trained for 50 epochs with a 0.2 validation split. Number of epochs was selected at 50 because there was not significant change in loss/accuracy after 45 epochs.

All the models were tested on the unseen test data separated before. After prediction, the predicted arrays were thresholded at 0.5, the same way it was done to the ground truth label at the pre-processing step



```
739/739 ━━━━━━━━━━━━━━━━━━━━ 23s 32ms/step - accuracy: 0.9975 - loss: 0.0075 - val_accuracy: 0.9982 - val_loss: 0.0056
Epoch 45/50
739/739 ━━━━━━━━━━━━━━━━━━━━ 23s 31ms/step - accuracy: 0.9975 - loss: 0.0074 - val_accuracy: 0.9982 - val_loss: 0.0056
Epoch 46/50
739/739 ━━━━━━━━━━━━━━━━━━━━ 23s 31ms/step - accuracy: 0.9975 - loss: 0.0075 - val_accuracy: 0.9982 - val_loss: 0.0055
Epoch 47/50
739/739 ━━━━━━━━━━━━━━━━━━━━ 23s 31ms/step - accuracy: 0.9975 - loss: 0.0074 - val_accuracy: 0.9982 - val_loss: 0.0055
Epoch 48/50
739/739 ━━━━━━━━━━━━━━━━━━━━ 23s 31ms/step - accuracy: 0.9975 - loss: 0.0075 - val_accuracy: 0.9982 - val_loss: 0.0056
Epoch 49/50
739/739 ━━━━━━━━━━━━━━━━━━━━ 23s 32ms/step - accuracy: 0.9975 - loss: 0.0074 - val_accuracy: 0.9982 - val_loss: 0.0055
Epoch 50/50
739/739 ━━━━━━━━━━━━━━━━━━━━ 23s 31ms/step - accuracy: 0.9975 - loss: 0.0073 - val_accuracy: 0.9982 - val_loss: 0.0055
```

Figure 7: Insufficient change in loss and accuracy after 45 epochs

5 Results

This study evaluated the performance of three distinct types of deep learning models—U-Net, ResNet, and Vision Transformer (ViT)—across various metrics pertinent to image segmentation. These metrics include Precision, Recall, F1 Score, Pixel-wise Accuracy, Intersection over Union (IoU), and Dice Score. The results are summarized in *Table 1* and discussed below:

Metric	UNet	ResNet	ViT
Precision	0.9482	0.9252	0.9316
Recall	0.8199	0.540	0.6744
F1 Score	0.8794	0.5835	0.7824
Pixel-wise Accuracy	0.9988	0.9975	0.99797
IoU	0.7848	0.5268	0.6426
Dice Score	0.8794	0.5935	0.7824

Table 1: Performance comparison of the three models.

Intersection over Union (IoU) is a critical metric in segmentation, measuring the overlap between the predicted and actual segmentations. U-Net achieved the highest IoU at 0.7848, followed by ViT at 0.6426 and ResNet at 0.5268.

Dice Score, similar to IoU, measures the overlap between two samples. The results mirrored those of the IoU, with U-Net showing the highest score of 0.8794, ViT at 0.7824, and ResNet at 0.5935.

Pixel-wise Accuracy reflects the overall accuracy of pixel classification across the image. All models exhibited high pixel-wise accuracy, with U-Net at 0.9988, ViT at 0.99797, and ResNet at 0.9975.

Precision measures the accuracy of the positive predictions. The U-Net model outperformed the others with a precision of 0.9482, indicating its superior capability in correctly identifying relevant pixels as compared to ResNet and ViT. ResNet, with a precision of 0.9252, and ViT, at 0.9316, also demonstrated high accuracy, albeit slightly lower than U-Net.

Recall assesses the model's ability to identify all relevant instances. U-Net achieved a recall of 0.8199, significantly higher than ResNet's 0.5040 and ViT's 0.6744. This suggests that U-Net is more effective in capturing relevant pixels without missing as many as the other two models.

F1 Score is the harmonic mean of precision and recall, providing a balance between the two metrics. Here, U-Net also led with an F1 score of 0.8794, indicating a well-balanced performance between precision and recall. ViT followed with a score of 0.7824, while ResNet lagged at 0.5835, reflecting some challenges in balancing recall with precision.

In summary, all three models performed quite well in the binary vascular segmentation task in terms of accuracy and precision. However, in the case of dice coefficient and IoU, U-Net did significantly better. Overall U-Net gave the best performance over all the metrics followed by ViT. However, the ViT model was also computationally more expensive and required longer training hours to give similar results. **Based on the results, we can conclude that the U-Net based model is the most suitable for this binary vascular segmentation task on the SenNet + HOA - Hacking the Human Vasculature in 3D dataset.**

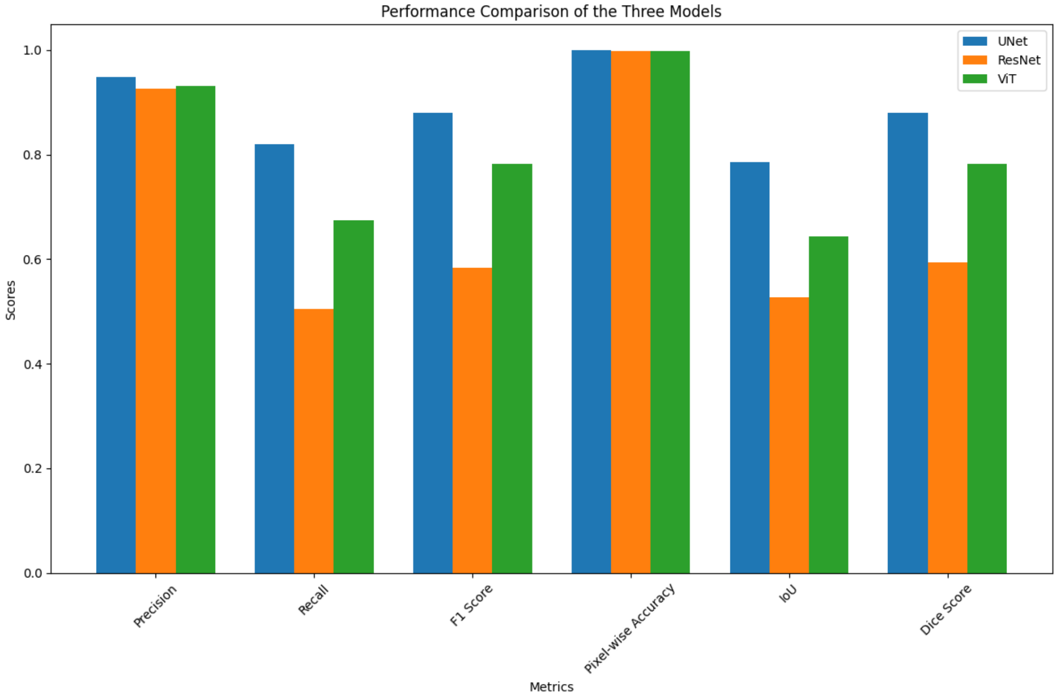


Figure 8: Comparison performance for U-Net, ResNet and ViT-based architectures for binary image segmentation

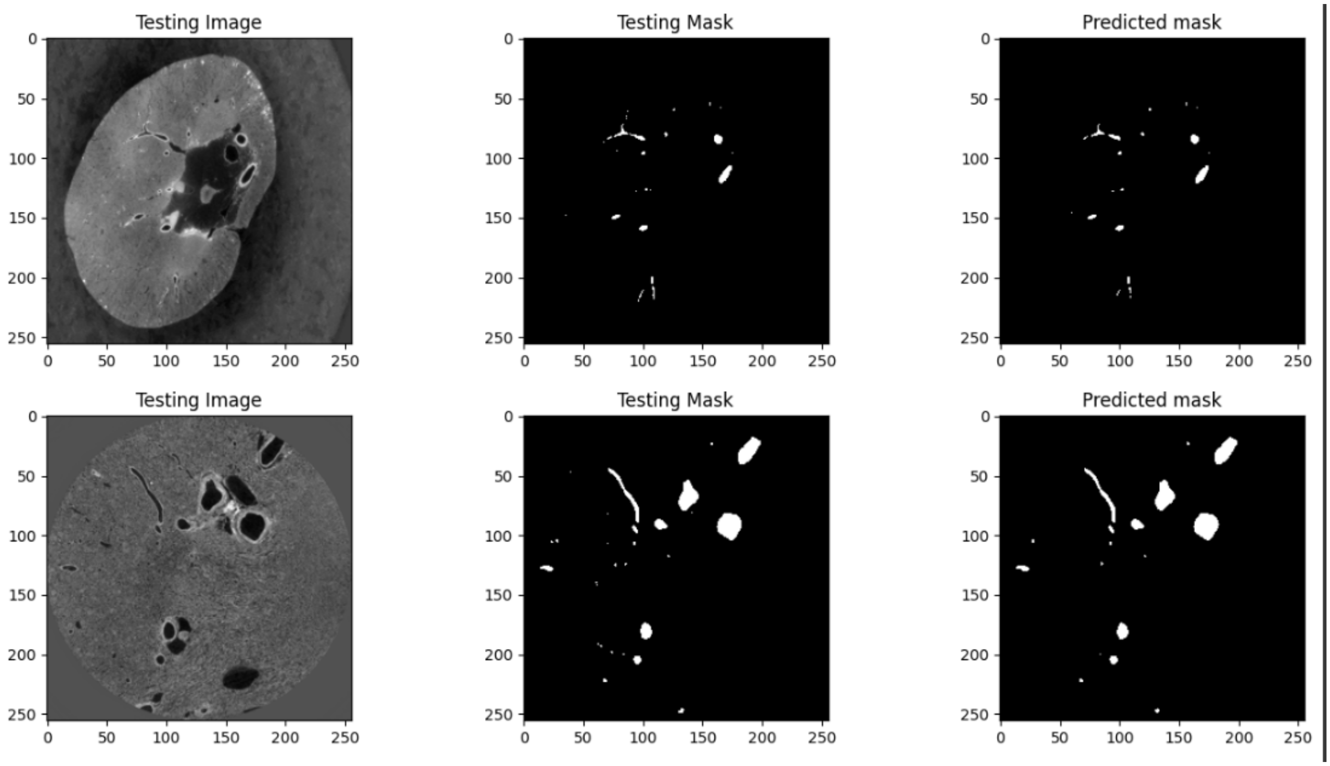


Figure 9: Prediction for two random test instances by UNet along with ground truth.

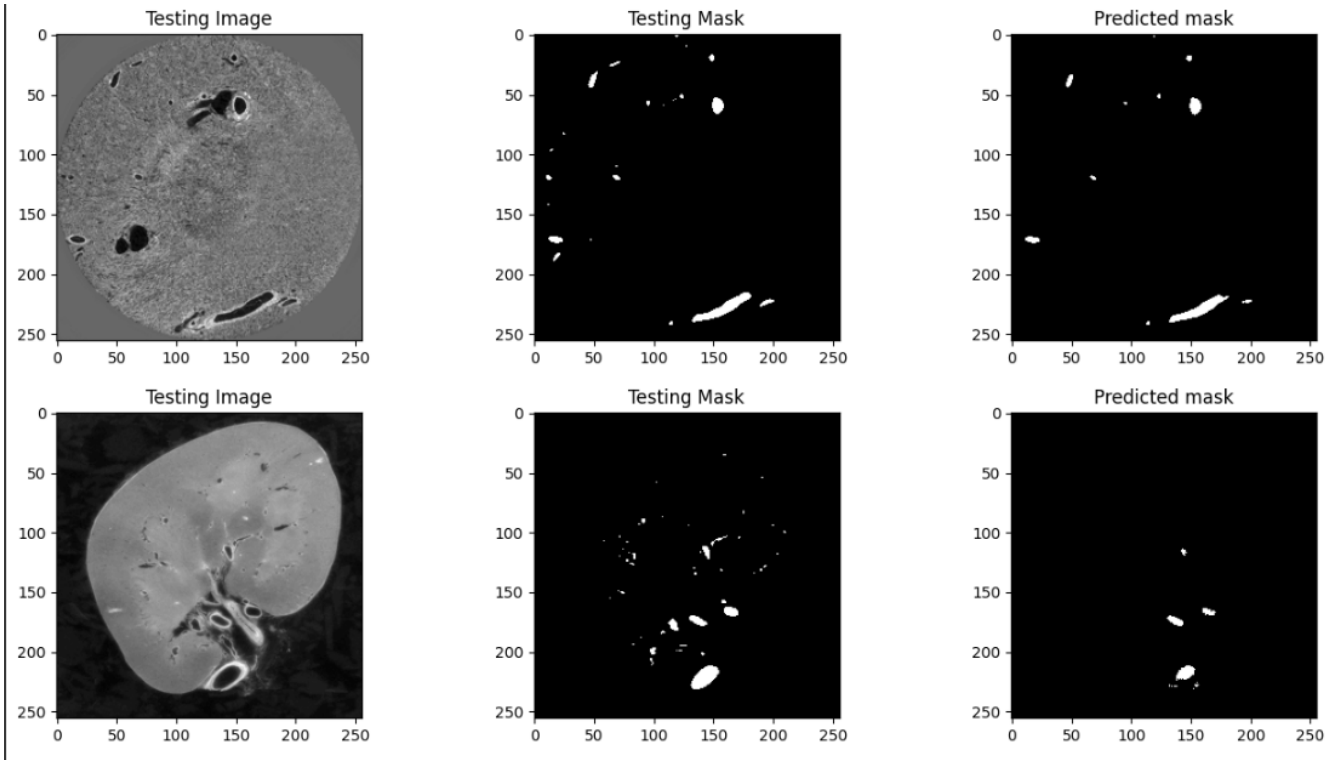


Figure 10: Prediction for two random test instances by ResNet along with ground truth.

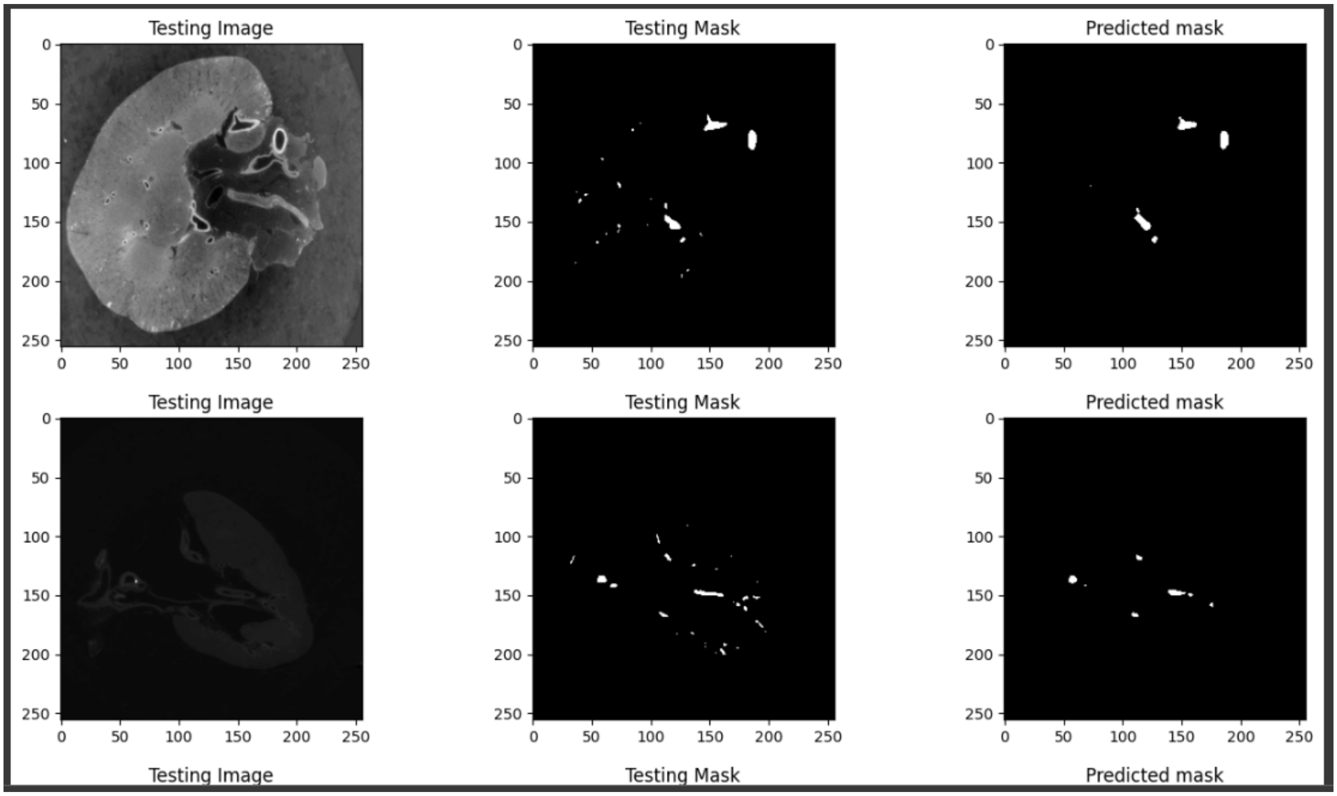


Figure 11: Prediction for two random test instances by ViT along with ground truth.

6 Conclusion

In this study, the performance of variations of three distinct models U-Net, ResNet, and Vision Transformer (ViT) was evaluated on the task of binary vascular segmentation using the SenNet + HOA - Hacking the Human Vasculature in 3D dataset. All three models demonstrated commendable precision and pixel-wise accuracy, indicating their effectiveness in classifying pixels correctly within the vascular structures.

U-Net consistently outperformed the other models across almost all metrics, particularly excelling in the Dice coefficient and Intersection over Union (IoU) metrics. These metrics are crucial for assessing the quality of segmentation, as they measure the overlap between the predicted and actual segmentations. It not only provided high accuracy and excellent overlap metrics but also did so with relatively lower computational demands compared to the ViT. These characteristics make U-Net particularly appealing for this segmentation task.

While the models demonstrated high accuracy metrics, their performance on the principal metrics such as IoU and the Dice coefficient varied significantly. This implies that there is still room for improvement in how these models handle the spatial coherence and continuity of anatomical structures.

Future work may explore more efficient transformer models or hybrid approaches that could potentially reduce the training demands while maintaining or enhancing performance. Further research could also focus on leveraging recent advancements in U-Net topologies such as V-Net, nn-UNet, generative models like GAN, and the latest Segment Anything Model (SAM) developed by Meta. Additionally, ensembling different models and integrating domain-specific expertise with multimodal imaging methods may provide deeper insights.

The exploration of 2.5D methods can also have interesting prospects which involve processing slices from consecutive layers to capture richer spatial context. This method could potentially bridge the gap between purely 2D and 3D methods, offering a deeper understanding of spatial relationships within the kidney volume, which could be particularly beneficial for complex anatomical structures like vasculature.

In conclusion, the U-Net model has proven highly effective and efficient compared to other ResNet and ViT for the task of binary vascular segmentation on the SenNet + HOA - Hacking the Human Vasculature in 3D dataset. There is ample opportunity for further advancements with continued research, which will eventually pave the way for breakthroughs in medical image processing.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[3] Yashvardhan Jain, Katy Borner, Claire Walsh, Nancy Ruschman, Peter D. Lee, Griffin M. Weber, Ryan Holbrook, and Addison Howard. Sennet + hoa - hacking the human vasculature in 3d., 2023. https://kaggle.com/competitions/blood-vessel-segmentation [Accessed: (25/4/2024)].

[4] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, December 2017.

[5] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey, 2020.

[6] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15, 12 2022.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.