Machine Learning for Abdominal Aortic Aneurysm Characterization from Standard-Of-Care Computed Tomography Angiography Images

by

Anish Salvi

BS in Bioengineering, University of Pittsburgh, 2021

Submitted to the Graduate Faculty of the

Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This thesis was presented

by

Anish Salvi

It was defended on

November 7, 2022

and approved by

Ender Finol, PhD, Professor, Department of Mechanical Engineering, University of Texas at San Antonio

Sanjeev Shroff, PhD, Professor, Department of Bioengineering

Thesis Advisor: Prahlad G. Menon, PhD, Assistant Professor, Department of Bioengineering

Machine Learning for Abdominal Aortic Aneurysm Characterization from Standard-Of-Care Computed Tomography Angiography Images

Anish Salvi, MS

University of Pittsburgh, 2022

Abdominal aortic aneurysms (AAAs) are dilations in the descending aorta which can result in internal bleeding when ruptured, leading to hospitalization or death. AAAs are commonly asymptomatic and discovered by happenstance during imaging tests, including computed tomography (CT) and its blood vessel enhancing counterpart computed tomography angiography (CTA). However, a past evaluation indicates that radiologists correctly identified and referred to monitoring only 32%, or 43 of 133 AAAs, from 3292 CTs. AAAs with larger diameters (> 5 cm) are recommended for elective repair; however, < 5 cm AAAs may have a rupture rate as high as 23%. Utilizing diameter as a one size fits all approach fails to consider intraluminal thrombus (ILT) and calcifications, clinically relevant attributes associated with elevated rupture risk. While a prior study indicated that type I and III endoleaks, linked with incorrect graft positioning during elective repair, had an incidence of only 6.4%, these complications require urgent medical attention. Surgical planning may benefit from greater understanding of AAA geometry. There remains a critical need for the automated discovery, visualization, and elective repair indication of AAAs. Having explored the novel field at the intersection of state-of-the-art machine learning and standard-of-care medical imaging, we develop deep learning models to aid the detection, segmentation, and classification of AAAs based on pre-operative CTA characteristics while observing the frame-of-reference. We describe computational methods which include bounding box localization as a precursor to high-resolution segmentation, patch-based segmentation of

medical image sub-volumes, image transformers that identify AAA severity, and a vision transformer that provides heatmaps indicative of AAA severity prediction. We find that 1) our memory-efficient bounding box method outperforms conventional neural network based AAA lumen segmentation, 2) patch-based AAA wall segmentation has improved performance as compared to our memory efficient computational pipeline for asymptomatic cases, 3) image transformers approach and even beat the accuracy achieved by rudimentary classifiers (i.e., differentiating between asymptomatic v. symptomatic AAAs) when leveraging embeddings derived from class specific segmentation models, and 4) vision transformers not only predict AAA severity accurately, but localize the disease by its anatomical basis. In sum, we make key contributions to scientific literature concerning medical imaging and machine learning through our computational methods of AAA interpretation.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Deep learning algorithms underpinned on convolutional neural networks (CNNs) and vision transformers (ViTs) demonstrate significant promise for medical imaging based interpretation of abdominal aortic aneurysms (AAAs).

1.1 Clinical Imaging & Aneurysms

AAAs are dilations in the descending aorta typified by abrupt abdominal pain. AAA rupture can cause internal bleeding, leading to hospitalization or death. In fact, reported ruptured AAAs accounted for ~28,200 deaths and reported unruptured AAAs accounted for ~15,400 deaths from 2009 to 2019 [1]. There remains a critical need for the automated discovery, visualization, and elective repair indication of AAAs to assist with incidental identification, surveillance, and surgical planning. Despite being the 15th leading cause of death, most AAAs are asymptomatic and only discovered by happenstance during imaging tests, including computed tomography (CT) and its blood vessel enhancing counterpart computed tomography angiography (CTA), for other afflictions [2]. Unfortunately, screening evaluations for AAAs indicate inconsistent performance by radiologists. One study indicated that radiologists reported only 65%, or 122 of 187 AAAs, from 3246 CTs [3]; however, another report stated that radiologists correctly identified and referred to monitoring only 32%, or 43 of 133 AAAs, from 3292 CTs [4]. These imaging professionals struggled to detect AAAs with shorter diameters [4].

Surgical decisions are primarily informed by aneurysm diameter where smaller AAAs (< 5 cm) require monitoring every 6-12 months while larger AAAs (> 5 cm) are recommended for elective repair [2]. However, an autopsy analysis indicated that < 5 cm AAAs had a 13% rupture rate while > 5 cm AAAs had a 40% rupture rate [5]; Though a separate study suggested that the rupture rate of < 5 cm AAAs may be as high as 23% [6]. Utilizing diameter as a one size fits all approach fails to consider additional clinically relevant attributes, leading to faulty rupture risk stratification. More recently, intraluminal thrombus (ILT) volumes and calcification presence have been associated with elevated rupture risk [7, 8]. Further, the geometry of the nonuniform AAA wall can inform internal bleeding risk (i.e., where a thicker wall implies tissue inflammation and a thinner wall is more prone to rupture after exposure to high stress) [9, 10]. Allowing clinical staff to consider image features of the aneurysm wall, ILT, and calcifications could help ensure optimal treatments.

Clinicians require CTA imaging to determine AAA size, location, and geometry for endovascular aneurysm repair. While a study reported that type I and III endoleaks, linked with incorrect graft positioning, had an incidence of only 6.4%, these complications require urgent medical attention [11, 12]. Delineating AAA boundaries could allow surgeons to better plan stent graft placement and design by providing measurements of the aortic neck, aneurysm morphology, and iliac artery anatomy [13]. Specifically, segmentation of AAA associated anatomy can inform the stent's start and end points as well as its diameter.

1.2 Biomechanical Evaluation of Disease

Understanding the forces at play regarding the AAA anatomy gives insight into the underlying mechanisms of disease, especially for identifying rupture risk stratification parameters

outside of AAA diameter. After conducting force and deformation experiments to AAA wall specimens, researchers observed the wall strength of larger AAAs is unrelated to diameter, noting that thickness and stiffness could be better indicators of rupture [14]. Performing finite element analysis (FEA) via CTs, investigators concluded that peak AAA wall stress in symptomatic, ruptured AAAs is greater than the same for their asymptomatic, intact counterparts [15]. Understanding of stress can help relate the ILT to AAA wall. The findings observed from AAA specimens subject to various cell chemistry and tensile strength studies drew a connection between regional wall weakness and thicker ILT related hypoxia [16].

Volumetric segmentation of the AAA serves as a starting point for biomechanical evaluations. AAA segmentation related AAA asymmetry to posterior wall stress, implying surgeons consider asymmetry when formulating elective repair decisions [17]. However, the incorporation of AAA related biomechanical software, such as FEA, into the clinical workflow has been slow due to expenses and requisite expert knowledge [18].

Artificial intelligence tools that automate segmentation of diseased anatomy for downstream calculation of biomechanical properties have been gaining traction in recent years. Authors reported a framework which performed 1) 2D CNN segmentation of the AAA, 2) reconstitution of predictions into 3D geometries, and 3) prediction of wall stress via regression of surface geometries [18]. However, our machine learning research does not stop at the intermediate step of segmentation for downstream mechanical computation of the AAA endpoint (i.e., level of severity, rupture risk indication, etc.). Instead, we apply the model's intuitive understanding of the image volumes themselves to identify the same types of endpoints directly. While biomechanical evaluation explains the underlying forces behind disease progression and aggravation from a mechanical perspective, we provide interpretable visualization regarding disease classification

from an imaging standpoint. Nevertheless, understanding the conclusions of long-established biomechanical research helps guide the model development process.

1.3 State of Machine Learning

Deep learning refers to a subclass of machine learning algorithms capable of automatic feature extraction and interpretation tasks (i.e., detection, segmentation, classification, etc.). We present several different perspectives on AAA interpretation as compared to the current literature on vascular imaging related quantification techniques. For example, a prior paper reported a CNN pipeline which produced the 3D aorta following concatenation of 2D aortic segmentations [19]. However, since it has been indicated that 3D models can learn more organized and precise patterns within volumetric data as compared to their 2D and 2.5D counterparts, we pursue 3D architectures for our studies [20, 21]. Previous research described segmenting the AAA from image volumes fit to the AAA region of interest itself [22]. However, we improve upon this shortcoming by formulating a sequential inference pipeline which performs bounding box localization prior to high-resolution segmentation, accounting for how acquisition methods can vary across clinical imaging settings. A previous manuscript proposed 3D segmentation of the AAA wall through variable neighborhood search, alternating the search between two different search spaces, voxel intensity and voxel gradient [23]. To the best of our knowledge, we are the first in the literature to describe automatic 3D segmentation of the AAA wall from the standpoint of CNNs. While initially limited to segmentation at lower resolutions by computational constraints, we pursue patch-based AAA wall segmentation which involves splitting patient CTAs into image sub-volumes that are subsequently inferred upon by the model prior to being reconstituted into the final prediction. Expanding upon

the previous work, we utilize the encoding path of a pretrained 3D segmentation model as an image transformer to facilitate neural embedding, or latent representation, based classification of AAA severity [24]. Our approach considers all image features (i.e., including the AAA lumen, wall, ILT, and any surrounding calcifications), unlike prior methods, such as a deep belief network which measured AAA severity based off aneurysm diameter [25]. Building upon both the patch and embedding based strategies we formerly used for segmentation and classification, we implement a 3D ViT capable of AAA severity classification [26]. Thus, we provide a novel contribution to the literature by providing patient and class specific explanations that illustrate the anatomical basis for the 3D model's decision-making, lending itself to medical imaging detection capabilities. Note, that in the case of image segmentation experiments, we evaluate performance using ground truths derived from and consistent with the original patient CTAs to better demonstrate algorithm performance.

1.4 Key Contributions

1) We demonstrate our image transformers can approach and even beat the accuracy achieved by rudimentary image classifiers or CNNs. Our study emphasizes that embeddings from class specific segmentation models carry meaningful information which relate anatomical structure to disease severity. Our findings contribute to both the machine learning and medical imaging communities for improved disease identification.
2) Our memory efficient method of bounding box localization as a precursor to high resolution segmentation of AAA lumen outperforms the conventional CNN approach to the same. We apply a novel technology to a common issue faced by developers in the

machine learning field, overcoming computational constraints of the graphics processing unit (GPU).

3) Patch based AAA wall segmentation improves upon our memory efficient computational pipeline regarding asymptomatic AAA cases. We apply a previously established technology to a novel biomedical imaging problem and fine tune the model to account for the large class imbalance present within the medical image volumes. Class imbalances are a common issue in the deep learning field (i.e., the anatomical volume of interest occupies significantly fewer voxels as compared to that of the whole patient medical image volume).
4) We engineer a ViT which not only accurately predicts AAA severity but also provides heatmaps that highlight the anatomical basis for its predictions in a class specific manner while respecting the DICOM format. The novel 3D model lends itself to detection capabilities via weakly supervised learning. Interpreting how models formulate their classification decisions, especially for novel state of the art transformers, is a nascent field within deep learning. Understanding and verifying that the model's decision is based upon the structure, or image characteristics of disease is especially relevant to the medical imaging community. The computational approach we describe generates class specific heatmaps (i.e., can incorporate more than two classes) and could be applied to other diseases which are more widespread throughout the medical image as compared to the AAA.

1.5 Peer Reviewed Publications

1) The sections pertaining to AAA-UNet are embodied in Salvi, A., Finol, E., & Menon, P. G. (2021, November). Convolutional Neural Network based Segmentation of Abdominal Aortic Aneurysms. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 2629-2632). IEEE.
2) The sections describing AAA Image Transformer are represented in Salvi, A., Finol, E., & Menon, P. G. (2022, April). Image transformers with regional attention for classification of aneurysm rupture risk without explicit segmentation. In Medical Imaging 2022: Biomedical Applications in Molecular, Structural, and Functional Imaging (Vol. 12036, pp. 278-286). SPIE.

2.0 Hypothesis

Having explored the novel field at the intersection of state-of-the-art machine learning and standard-of-care medical imaging, we develop deep learning models to aid the detection, segmentation, and classification of AAAs based on pre-operative CTA characteristics while observing the inherent DICOM format. The latent space of CNNs engineered for image segmentation encodes anatomical information indicative of not only AAA geometry, but severity. As such, we perform accurate segmentation of the AAA lumen using the de facto CNN approach to serve as a baseline. We note that the standard CNN, referred to as AAA-UNet, produces segmentations with ill-defined boundaries associated with poor resolution. In consequence, we develop a novel and memory efficient method which consists of first localizing the AAA to a bounding box prior to high resolution segmentation of the AAA lumen. The described sequential inference pipeline, named BB-AAA-UNet outperforms AAA-UNet for the same volumetric segmentation task. Segmentation of small objects remains a challenge due to the foreground v. background class imbalance. However, the AAA wall, a relatively thin object present within a much larger medical image, is a key volumetric feature related to AAA rupture. Thus, we reapply BB-AAA-UNet for successful segmentation of the AAA wall. While BB-AAA-UNet successfully localized the AAA, it was determined that preserving resolution during training is critical for elevating AAA wall segmentation accuracy. Therefore, we describe a patch-based segmentation pipeline, Patch Segmentation UNet which involves 1) converting the 3D medical image volume to sub-volumes, 2) performing segmentation on each sub-volume, and 3) reconstructing each segmentation sub-volume. By adapting and fine tuning this previously established technology, we innovate a model that performs more accurately than our previously described memory efficient

pipeline. The embeddings derived from segmentation models indicate structural disease through regional attention. As such, image transformers improve CNN classification accuracy for describing asymptomatic v. symptomatic AAAs as compared to the de facto approach. Established methods require measurement of aneurysm diameter, ILT volume, and AAA mechanics, etc. using computational tools. Volumetric segmentation by deep learning algorithms helps automate computation for such measurements. In their hidden layers, models engineer features which differentiate AAAs. Interpretation of these features avoids the need for manual measurement as an input to statistical modeling. Inspired by Patch Segmentation UNet and AAA Image Transformers, we develop a novel 3D medical image ViT which incorporates patch embedding of images at higher resolutions, multilayer perceptron classification, and self-attention. We surmise that AAA-VIT provides accurate AAA severity classification and AAA localization by class specific heatmaps. Unlike the image transformers, the ViT does not require AAA annotation for training. Further, the model itself is interpretable, providing an anatomical explanation for reaching its conclusion regarding a patient image.

3.0 Methods

We conduct several deep learning studies to improve the clinical discovery and understanding of AAAs starting from pre-operative CTAs. Note, each model is fit to the training data (learning volumetric patterns) whereas the test set quantifies model performance on previously unseen data. Note, all images displayed in the methods section are the CTA images and labels, not model predictions.

3.1 Datasets

The two datasets used for the following course of studies are referred to as AAA-DICOM (54 patients) and AAA-Wall (123 patients), both were provided courtesy of researchers from University of Texas at San Antonio.

3.1.1 AAA DICOM Dataset

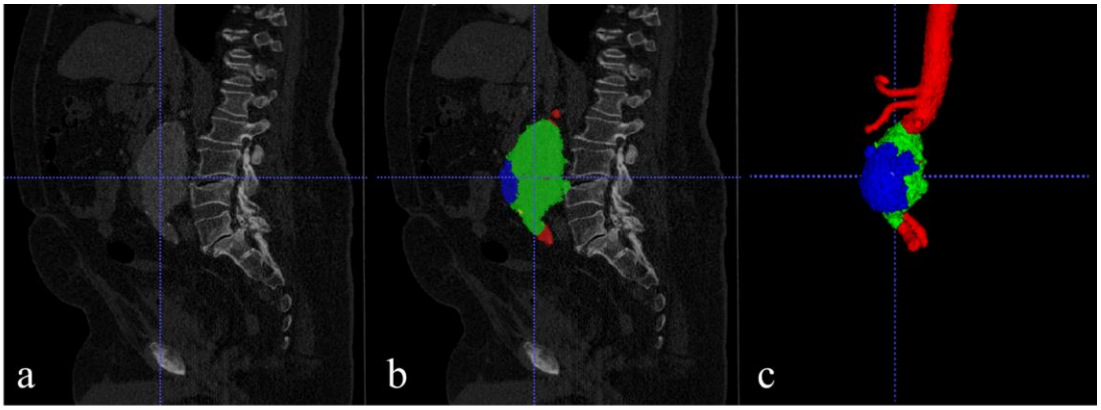

Figure 1: AAA DICOM Dataset

Displays a) the computed tomography angiography (CTA) of a patient recommended for elective repair; b) the abdominal aortic aneurysm (AAA) annotated for the aorta in red, lumen in green, intraluminal thrombus (ILT) in blue, and calcifications in yellow; and c) the previous in 3D. The AAA-DICOM dataset had an original pixel size is 512 x 512 where the z-dimension varied. The voxel spacing varied per patient as well.

The first dataset, referred to as AAA DICOM, consisted of 54 CTAs presenting for elective repair or surveillance acquired from two separate hospitals which are abbreviated as NMH and AGH (Figure 1). Each segmentation was prepared by leveraging ITK-SNAP's voxel contour annotation tool to label the aorta, AAA lumen, ILT (if present), and calcifications (if present) [27]. Segmentation of each volume was performed by the author, leveraging knowledge of AAA image characteristics (i.e., the brightness of the contrast enhanced lumen, the high intensity of calcification voxels, approximate location of the darker ILT, etc.). Following a stratified split according to hospital and treatment indication (i.e. elective repair or surveillance), the training set consisted of 24 patients (13 NMH & 11 AGH; 9 elective-repair & 15 surveillance) while the test set consisted of 30 patients (15 NMH & 15 AGH; 12 elective-repair & 18 surveillance).

3.1.2 AAA Wall Dataset

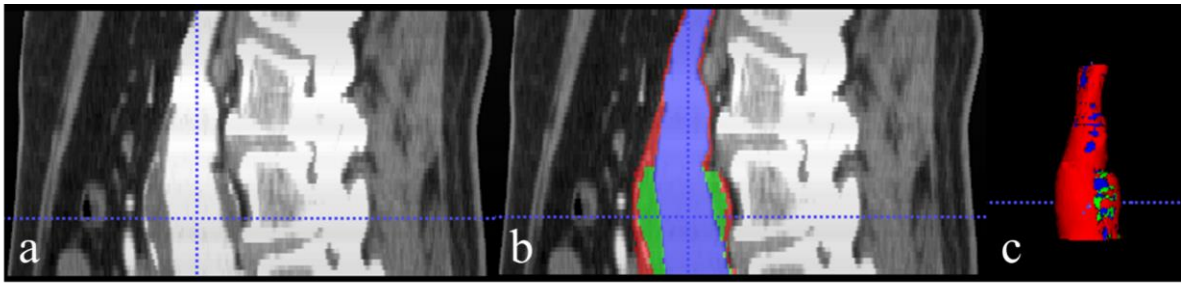

Figure 2: AAA Wall Dataset

Displays a) an asymptomatic patient's computed tomography angiography (CTA) image with; b) the labelled CTA annotated as abdominal aortic aneurysm (AAA) lumen in blue, intraluminal thrombus (ILT) in green, and AAA wall in red; and c) the previous in 3D. The original voxel size each image is 512 x 512 where the z-dimension varies. The voxel spacing is set to 0.9375 x 0.9375 x 3.5 millimeters.

The second dataset, referred to as AAA Wall, consisted of 123 patient-specific CTAs presenting for asymptomatic and symptomatic cases, the latter consisting of a majority unruptured and minority ruptured walls. Each CTA had been labelled for the AAA lumen, ILT, and AAA wall. Note, the AAA-Wall had been fit to the AAA region of interest on the z-axis, unlike AAA-DICOM (Figure 2). Following a stratified split according to symptom presence and rupture status, the training set consisted of 73 patients (38 Asymptomatic & 35 Symptomatic with 30 Unruptured & 5 Ruptured Walls) while the test set consisted of 50 patients (25 Asymptomatic and 25 Symptomatic with 21 Unruptured & 4 Ruptured Walls).

3.2 AAA-UNet: Baseline Aneurysm Segmentation

3.2.1 Medical Image & Data Preprocessing

The undermentioned study involved AAA-DICOM, exclusively. Each image and annotation pair was resampled to 64x64x64 voxels via b-spline and nearest-neighbor interpolation, respectively. Subsequently, each image was z-normalized while each annotation was binarized to include the AAA lumen, exclusively, to serve as model inputs.

3.2.2 Model Development

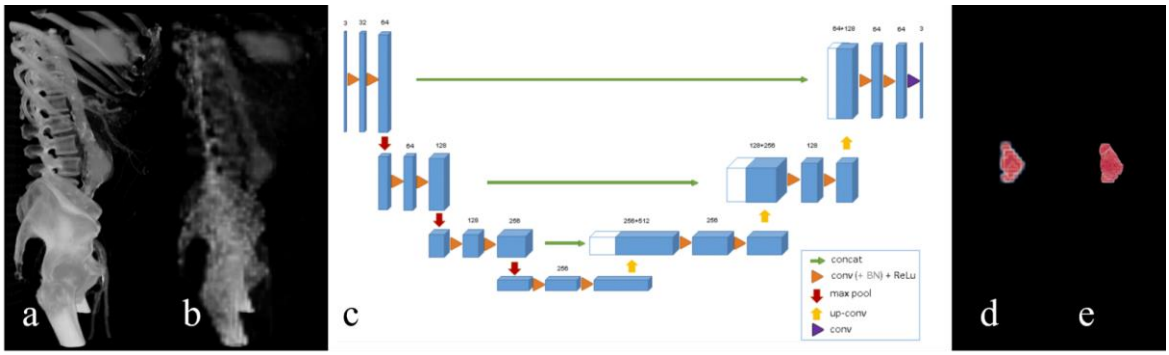

Figure 3: AAA-UNet

Describes the AAA-UNet methodology where a) represents the original computed tomography angiography (CTA) image; b) represents the previous resampled to 64x64x64 voxels and z-normalized; c) displays the 3D UNet architecture [28]; d) illustrates the abdominal aortic aneurysm (AAA) lumen volume in 64x64x64 voxels; and e) depicts the previous, consistent with the original CTA.

The UNet applies a series of convolutional layers to a medical image, culminating in a feature map after which a sequence of deconvolutional layers output a prediction of equal size to the input [28]. Each voxel of the output receives a likelihood of representing the volume of interest

(i.e., the AAA lumen). We first develop AAA-UNet, a single model, which typifies the approach to automatic image segmentation leveraged by conventional deep learning models, thereby providing baseline results for later comparison (Figure 3).

$$DSC = \frac{2 * (P \cap Y)}{P + Y} \quad \textbf{3-1}$$

Where the Dice Similarity Coefficient (DSC) represents the twice the intersection of the voxels between the Prediction (P) and Ground Truth (Y) divided by total number of voxels attributed to P and Y. Note, that P represents the model predicted segmentation while Y represents the ground truth annotation of the same.

$$DL = 1 - DSC \quad \textbf{3-2}$$

Where the Dice Loss (DL) is the complement of the DSC. The model is iteratively fit to minimize the DL.

AAA-UNet was trained upon the 64x64x64 z-normalized CTAs with the AAA lumen serving as the target volume and was fit according to the Dice loss (Equations 3-1 and 3-2). The model employs a base filter 64, block number 4, and utilized max-pooling as the down-sampling operation. As opposed to the original UNet which implemented the rectified linear unit (ReLU) as the activation function, we apply the leaky ReLU (LeakyReLU). Training parameters were specified as batch size 1, learning rate 1e-4, epochs 100, and optimizer AdamW with weight decay 1e-2 [29]. To prevent the model from overfitting (i.e., memorizing the training data as opposed to learning the AAA's volumetric features which are applicable to unseen data), image augmentation consisting of random rotations, flipping, reflection, and noise was performed during the training process to increase data variability.

3.2.3 Performance Evaluation

The sigmoid function was applied to each of AAA-UNet's predictions. By preserving the inherent meta information (i.e., real-world spacing, origin, size) associated with each patient's medical image via the SimpleITK library, we resample the model predicted segmentations via b-spline interpolation to ensure they are clinically consistent with their patient-specific CTA counterparts [30]. After Otsu thresholding to form binary representations of the AAA lumen, we evaluate model segmentation accuracy, in terms of DSC, with ground truths consistent with the original frame of reference as obtained during standard-of-care imaging (Equation 3-1) [31].

3.3 BB-AAA-UNet: Memory Efficient High-Resolution Segmentation with Prior Aneurysm Localization

3.3.1 Medical Image & Data Preprocessing

As above, the following study used the AAA-DICOM dataset with several of the model inputs from the previous exercise being used here as well. For the first stage of training, each image and annotation pair was resampled to 64x64x64 voxels before each CTA was z-normalized while each annotation was binarized to the AAA lumen, ILT, and calcifications. For the second stage of training, each CTA was cropped to the region-of-interest inclusive of the AAA lumen, ILT, and calcifications with same cropping policy applied to each corresponding patient annotation. Further, each region-of-interest CTA was resampled to 64x64x64 voxels before being z-normalized while each cropped annotation was binarized to the AAA lumen, exclusively, and resampled to the same

dimensions. As before, each image and label were interpolated via b-spline and nearest neighbor, respectively.

3.3.2 Model Development

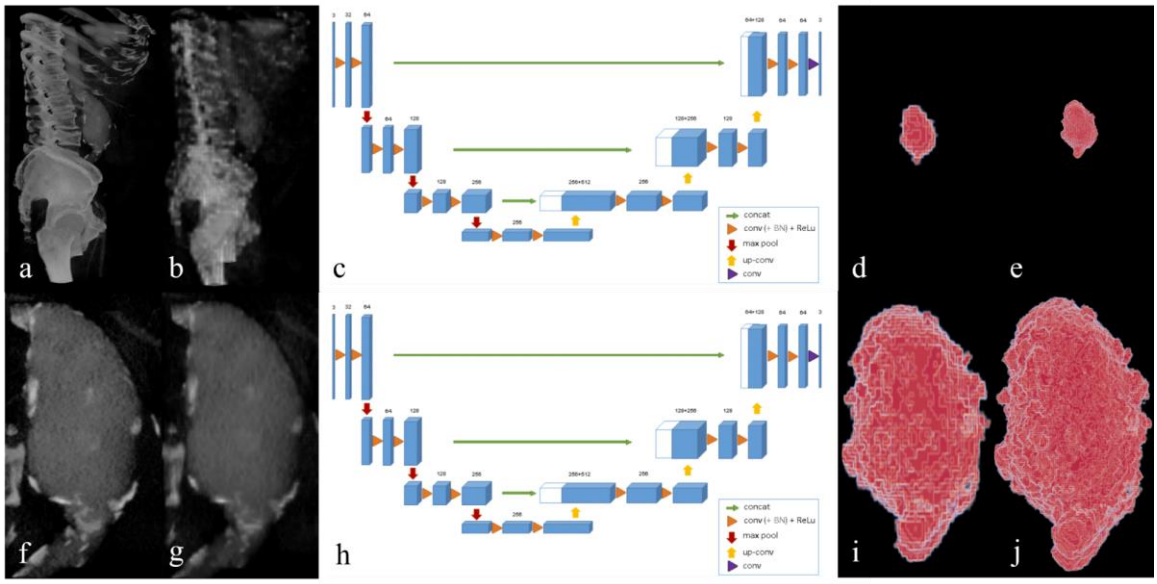

Figure 4: BB-AAA-UNet

Describes the BB-AAA-UNet methodology for bounding box localization prior to high resolution segmentation. The first stage where a) represents the original computed tomography angiography (CTA); b) depicts the previous resampled 64x64x64 and z-normalized; c) is the 3D UNet used for localization; d) is the 64x64x64 volume inclusive of the abdominal aortic aneurysm (AAA), intraluminal thrombus (ILT), and surrounding calcifications; and e) is the previous with the same frame of reference as the original CTA. The second stage where f) represents the region-of-interest CTA; g) is the previous cropped to 64x64x64 voxels and z-normalized; h) is the 3D UNet used for high-resolution segmentation; i) is the 64x64x64 AAA lumen volume; and j) represents the previous in same frame of reference as the original region-of-interest CTA.

Image segmentation at higher resolutions often suffers from computational overhead (i.e., the memory constraints of the compute environment), preventing the development of segmentation

pipelines which can capture the finer details of the anatomical region of interest, especially along the boundaries. As such, we proposed our own invention, BB-AAA-UNet, sequential models which fix a bounding box to the region of interest prior to detailed segmentation of the AAA lumen (Figure 4). BB-AAA-UNet's inference pipeline occurs in two stages, functioning within a low-cost development environment.

The localization model deployed in the first stage was trained on the 64x64x64 z-normalized CTA with the AAA lumen, ILT, and calcifications serving as the target. The high-resolution segmentation model deployed in the second stage was trained on the cropped 64x64x64 z-normalized CTAs with the AAA lumen serving as the target. For training, identical hyperparameters, loss function, and augmentation policies as the AAA-UNet study were used. Note, the model is not trained on its own predictions.

3.3.3 Performance Evaluation

BB-AAA-UNet's inference occurs sequentially, resembling the training procedure. In the first stage, the localization model infers the AAA lumen, ILT, and calcifications from the 64x64x64 z-normalized CTA. Following sigmoid function application, the prediction is resampled to the original patient CTA volume by b-spline interpolation before being Otsu thresholded to form the rudimentary AAA segmentation. By applying a bounding box fit to this rudimentary region-of-interest, the same bounded region is extracted from the input CTA resampled to 64x64x64 voxels. After z-normalization of the region-of-interest CTA, the high-resolution segmentation model infers the AAA lumen. The sigmoid function is applied to the resulting secondary prediction after which b-spline interpolation and Otsu thresholding ensure the AAA lumen is consistent with its region-of-interest CTA counterpart. This final prediction (i.e., confined within the bounding

box) is reinserted into the patient-specific standard-of-care, original CTA's frame of reference. As such, AAA lumen segmentation performance comparisons, in terms of DSC, can be made between AAA-UNet and BB-AAA-UNet (Equation 3-1).

We implemented the Anderson-Darling test to determine the distribution (i.e., normal v. non-normal) of the test set AAA lumen DSCs of AAA-UNet v. BB-AAA-UNet. After, we applied the appropriate test for statistical significance, t-test for normal distributions or Mann-Whitney U and Wilcoxon rank-sum for non-normal distributions. For all tests, a = 5e-2.

3.4 BB-AAA-UNet: As Applied to Aneurysm Wall Segmentation

3.4.1 Medical Image & Data Preprocessing

The following study used the AAA Wall dataset. For the first stage of training, each image and annotation pair was resampled to 64x64x64 voxels before each CTA was z-normalized and each annotation was binarized to the AAA lumen, ILT, and wall. For the second stage of training, each CTA was cropped to the region-of-interest inclusive of the AAA lumen, ILT, and wall with same cropping policy applied to each corresponding patient annotation. Further, each region-of-interest CTA was resampled to 64x64x64 voxels before being z-normalized while each cropped annotation was binarized to the AAA wall and resampled to the same dimensions, exclusively. As before, each image and label were interpolated via b-spline and nearest neighbor, respectively.

3.4.2 Model Development

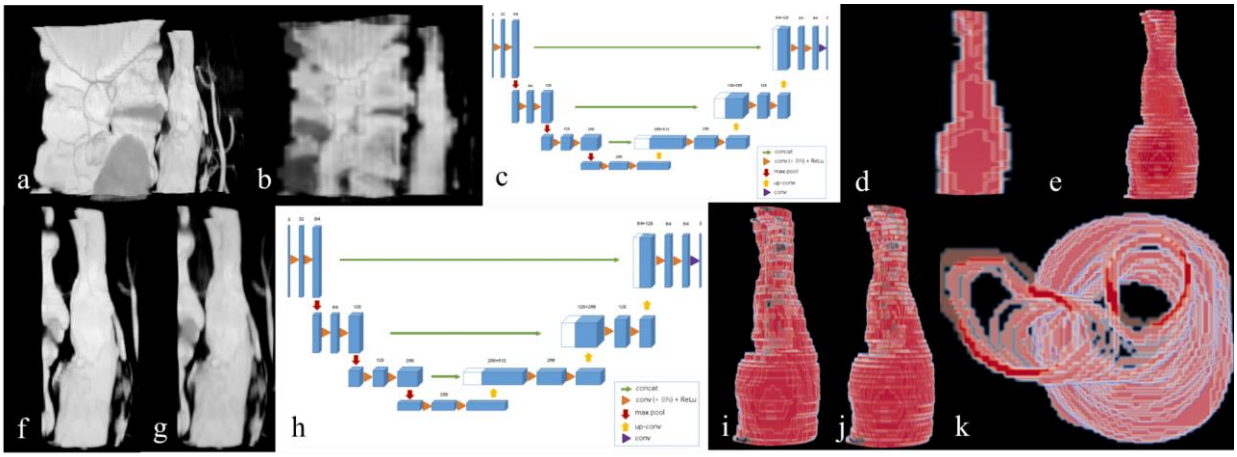

Figure 5: BB-AAA-UNet for AAA Wall Segmentation

Describes the BB-AAA-UNet methodology as applied to AAA wall segmentation. The first stage where a) represents the original computed tomography angiography (CTA); b) depicts the previous resampled 64x64x64 and z-normalized; c) is the 3D UNet used for localization; d) is the 64x64x64 volume inclusive of the abdominal aortic aneurysm (AAA) lumen, intraluminal thrombus (ILT), and surrounding AAA wall; and e) is the previous with the same frame of reference as the original CTA. The second stage where f) represents the region-of-interest CTA; g) is the previous cropped to 64x64x64 voxels and z-normalized; h) is the 3D UNet used for high resolution segmentation; i) is the 64x64x64 AAA wall volume; j) represents the previous with the same frame of reference as the original region-of-interest CTA; and k) is the AAA wall from the top view.

As compared to the AAA lumen, segmentation of the AAA wall is a more difficult task owing to the few numbers of voxels occupied by the anatomy within a medical image; the problem is compounded when considering image resampling to lower resolutions can lose relevant information. Thus, predicting this small foreground target within a large background remains a challenging endeavor (due to the foreground-background class imbalance).

Given BB-AAA-UNet's success at AAA lumen segmentation, we applied the same for AAA wall segmentation (Figure 5). The localization model deployed in the first stage was trained on the 64x64x64 z-normalized CTA with the AAA lumen, ILT, and wall serving as the target. The high-resolution segmentation model deployed in the second stage was trained on the cropped 64x64x64 z-normalized CTAs with the AAA wall serving as the target. For training, the same hyperparameters, loss function, and augmentation policies as the AAA-UNet and BB-AAA-UNet studies were conserved. Owing to the dataset size, the number of epochs trained were 55 for the first stage localization model and 80 for the second stage high-resolution segmentation model.

3.4.3 Performance Evaluation

As before, the model performs inference in two parts. In the first stage, the localization model infers the AAA lumen, ILT, and wall from the 64x64x64 z-normalized CTA. Following sigmoid function application, the prediction is resampled to the original patient CTA volume by b-spline interpolation before being Otsu thresholded to form the rudimentary AAA segmentation. By applying a bounding box fit to this rudimentary region-of-interest, the same bounded region is extracted from the input CTA resampled to 64x64x64 voxels. After z-normalization of the region-of-interest CTA, the high-resolution segmentation model infers the AAA wall. The sigmoid function is applied to the resulting secondary prediction after which b-spline interpolation and Otsu thresholding ensure the AAA wall is consistent with its region-of-interest CTA counterpart. This final prediction (i.e., confined within the bounding box) is reinserted into the patient-specific, original CTA's frame of reference. DSC was used to evaluate AAA wall segmentation performance (Equation 3-1).

3.5 Patch Segmentation UNet: Prediction of Aneurysm Wall by Medical Image Sub-volumes

3.5.1 Medical Image & Data Preprocessing

The following study used the AAA Wall dataset. Each image and annotation pair was resampled to 256x256x64. As before, each CTA was z-normalized and annotation was binarized to include the AAA wall, exclusively.

3.5.2 Model Development

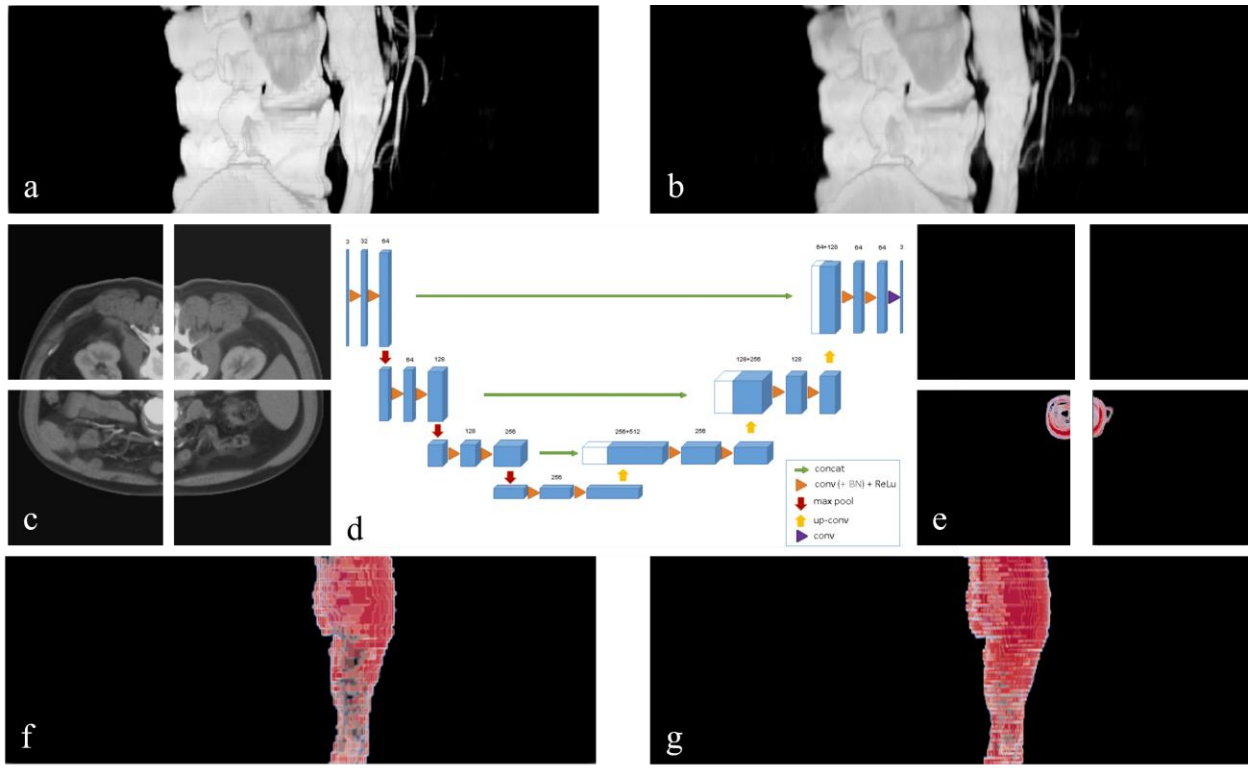

Figure 6: Patch Segmentation UNet

Depicts the Patch Segmentation UNet methodology where a) represents the original computed tomography angiography (CTA) image; b) represents the previous resampled to 256x256x64 and z-normalized; c) depicts the previous split into four patches each of size 128x128x64; d) indicates the 3D UNet used for subvolume segmentation; e) displays the AAA wall volume at size 128x128x64 amongst four patches; f) represents the reconstituted AAA wall volume at 256x256x64; and g) is the previous at the identical frame of reference as the original CTA.

Under the assumption that segmentation at higher resolutions would elevate the accuracy associated with AAA wall segmentation, we pursue a patch-based segmentation pipeline which operates upon medical image sub-volumes (Figure 6) [32]. We finetune the training and model parameters for our specific biomedical imaging endeavor.

Training parameters were specified as batch size 2, learning rate 1e-5, epochs 27, and AdamW optimizer with weight decay 1e-5. As part of the training process, 6 samples of size 128x128x64 sub-volumes were randomly taken from the input 256x256x64 CTAs with uniform probability. The model parameters were specified as 4 encoding blocks, linear up-sampling, and the parametric rectified linear unit (PReLU) as the activation function. The model was trained with the 128x128x64 z-normalized CTA sub-volumes serving as the input and the image background (0) and AAA wall (1) serving as the mutually exclusive targets. Data augmentation consisted of random flipping, noise, and intensity image transformations.

Given the chance that the image sub-volume contained no instance of the AAA wall, we adopt the Generalized Dice Loss (GDL) in lieu of the DL implemented for fitting prior models. Mathematically, the GDL applies the weight of the inverse of the class label volume squared when computing the losses associated with each class, background and the AAA wall (i.e., a class weighted average of the loss) [33]. In effect, the contribution towards the loss of an incorrectly

predicted AAA wall is much greater than the same for the background, despite the AAA volume being much smaller than the background. In sum, the GDL is designed to reduce the class imbalance posed by multiclass segmentation.

3.5.3 Performance Evaluation

As part of model inference, the 256x256x64 z-normalized CTA is split into patches of 128x128x64 sub-volumes. The model's inference pipeline predicts the probability map of each sub-volume being associated with either the background or AAA wall, though following the application of the channel-wise softmax function and reconstitution of sub-volumes, these predictions are later binarized to include the AAA wall, exclusively. These predictions of size 256x256x64 voxels are resampled to the original patient's CTA frame of reference using nearest neighbor interpolation and later evaluated for segmentation accuracy (Equation 3-1).

We implemented the Anderson-Darling test to determine the distribution of the test set asymptomatic AAA wall DSCs of BB-AAA-UNet v. Patch Segmentation UNet. After, we applied the appropriate test for statistical significance, t-test for normal distributions or Mann-Whitney U and Wilcoxon rank-sum for non-normal distributions. The process was repeated for symptomatic cases. For all tests, a = 5e-2.

3.6 AAA Image Transformers: Classifying Medical Images by Aneurysm Severity with Latent Representations

3.6.1 Medical Image & Data Preprocessing

The following study used the AAA Wall dataset. For the first stage of training, each image and annotation pair was resampled to 64x64x64 voxels before each CTA was z-normalized and each annotation was binarized to the AAA lumen, ILT, and wall. Patients with asymptomatic cases were labelled as the negative class while the symptomatic cases were labelled as the positive class. Thus, the presence of AAA symptoms served as an estimate for AAA severity (i.e., positive cases being more severe). Severity is a proxy to rupture risk, a limitation to our data gathering efforts.

3.6.2 Model Development

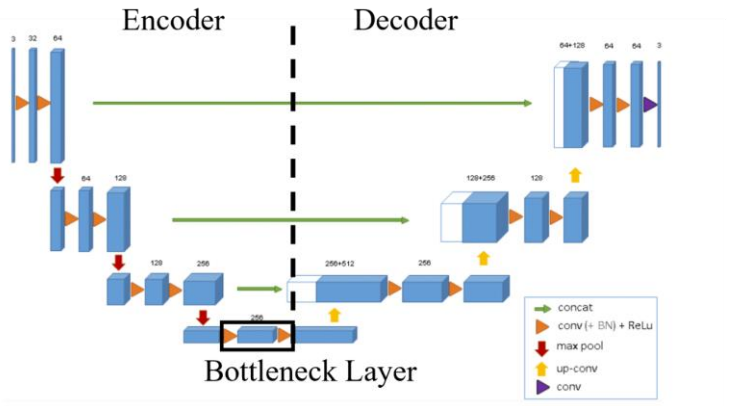

Figure 7: CNN Encoder & Decoder

The 3D UNet consists of contracting path (encoder) followed by an expanding path (decoder). The 3D UNet's bottleneck layer (feature map) can contain meaningful information on the region the model was trained to segment.

A pretrained 3D UNet's contracting path can serve as an image transformer, converting CTAs into their corresponding neural embeddings, or numerical vectors which contain feature representations on the region the model was trained to segment, or the AAA (Figure 7).

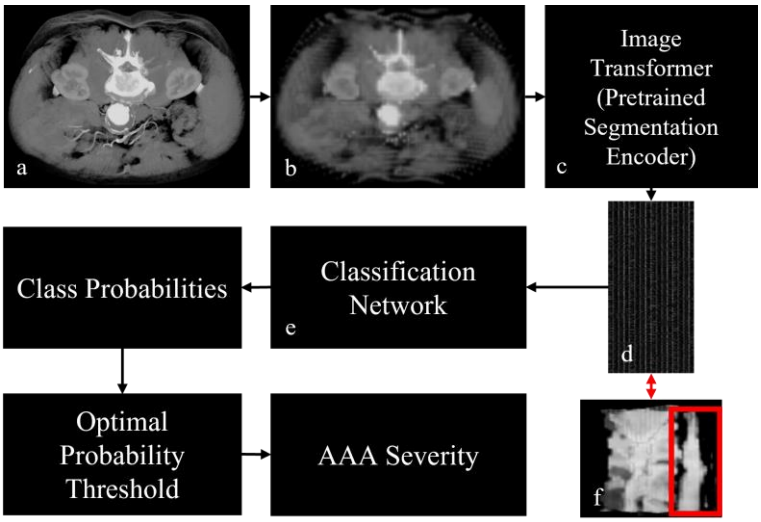

Figure 8: Representation Learning

Describes the representation learning workflow where a) displays the original patient computed tomography angiography (CTA); b) is the same resampled to 64x64x64 voxels and z-normalized; c) is the pretrained encoding or contracting path of the 3D UNet; d) the latent representation or neural embedding of the input CTA; e) depicts the neural network engineered for volumetric classification tasks. Following the optimization of probability thresholds governing the model prediction of positive v. negative classes, the final asymptomatic vs symptomatic predictions are given. Note, f) represents how the embedding serves as a latent representation of the abdominal aortic aneurysm (AAA).

We adopt a representation learning pipeline which involves converting CTAs to latent representations prior to neural net classification (Figure 8). The contracting path, or encoder, is extracted from a localization model. This pretrained encoder is applied to each 64x64x64 z-normalized CTA, serving as an image transformer which generates the corresponding patient-

specific 2048x4x4x4 neural embeddings from the bottleneck layer. The belief is that the transformed feature representation of the anatomical region the segmentation model was trained to identify contains characteristics which differentiate between asymptomatic and symptomatic AAAs. Each of the patient-specific embeddings are reshaped to 64x64x32 and z-normalized.

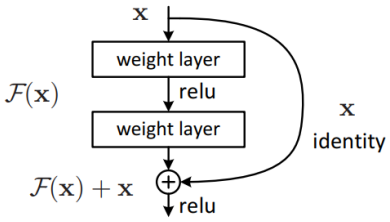

Figure 9: Skip Connection

The characteristic building block (skip connection) of a 3D ResNet designed for training deeper neural networks, or more layers. Adapted from Deep Residual Learning for Image Recognition [34].

Several 3D residual neural networks (ResNets) underpinned on skip connections are trained to identify these neural embeddings as either the positive or negative class, fit according to the cross-entropy loss (Figure 9) [34, 35, 36].

We first perform a baseline study by classifying the 64x64x64 z-normalized CTA images directly using the 3D ResNet. Training parameters for the image classifier were specified as batch size 4, learning rate 1e-4, and epochs 150.

We then conduct an image transformer study leveraging the encoding path of a class agnostic segmentation model derived from the first stage localizer as described in the BB-AAA-UNet when applied for AAA wall segmentation, which predicts AAA lumen, ILT, and wall. The training parameters of the embedding classifier were batch size 4, learning rate 5e-7, and epochs 300.

The remaining two image transformer studies utilize the encoding paths of class specific segmentation models. The asymptomatic and symptomatic UNets were trained upon the 38 asymptomatic and 35 symptomatic cases from the training set, respectively and exclusively. Identical training and model parameters to the first stage localizer as described for AAA wall segmentation of BB-AAA-UNet were applied though the number of epochs trained was 50. For the embeddings derived from the asymptomatic segmentation encoder, the training parameters of the embedding classifier were set to batch size 4, learning rate 1e-06, and epochs 300. For the embeddings derived from the symptomatic segmentation encoder, training parameters of the embedding classifier were specified as batch size 4, learning rate 1e-06, and epochs 300. In sum, we present one baseline and three image transformer studies.

3.6.3 Performance Evaluation

The softmax function was applied to the 3D ResNet's predicted, unnormalized logits to ensure the probability of the negative class was p and the same for the positive class was $1 - p$ per patient-specific embedding. The area under the receiver operating characteristic curve (AUC) is a threshold agnostic measurement of classification performance between the probability of the positive class and the class labels. As influenced by the AUC, we search for the optimal probability threshold (OPT) of p by which to define the threshold for the positive and negative predictions as informed by the training set and later applied to the test set. The OPT is determined by the top accuracy achieved by setting the threshold for the positive class. The precision, negative predictive value, recall, specificity, and F1 are reported as well. These metrics fully capture class specific variation in performance, ensuring both an accurate and balanced model.

3.7 AAA-ViT: Moving Towards Detection with Classification of Aneurysm Severity with Anatomical Explanation

3.7.1 Medical Image & Data Preprocessing

The following study involved the AAA Wall dataset. Each patient CTA was resampled to 128x128x64 voxels prior to z-normalization. Identical to the AAA Image Transformer study, patients with asymptomatic cases were labelled as the negative class while the symptomatic cases were labelled as the positive class under the assumption that symptoms imply greater disease severity.

3.7.2 Model Development

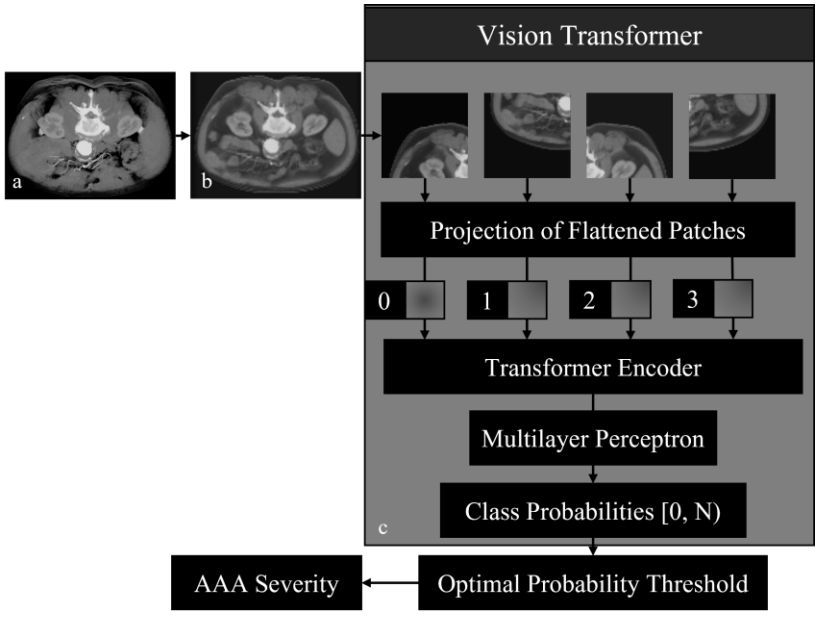

Figure 10: Vision Transformer

Represents the methodology and objective associated with AAA-ViT where a) represents the original patient computed tomography angiography (CTA); b) is the same resampled to 128x128x64 voxels and z-normalized;

c) represents the processes underlying the vision transformer (ViT). During the forward pass, the ViT projects a sequence of flattened patches following a 3D convolution across each portion of the input image. The resulting patch + positional embeddings are passed through a transformer encoder, before culminating in a multilayer perceptron which predicts the class probabilities [0, N) with N classes. Following the optimization of probability thresholds governing the model prediction of positive v. negative classes, the final asymptomatic vs symptomatic predictions are given.

Transformers have been predominantly applied to natural language processing (NLP) studies, though in more recent years, such algorithms have been embraced for vision tasks as well. Serving as both an extension and advancement of our previous work on image transformers, we engineer a 3D ViT which classifies the AAA severity from medical images. Authors had reported a 2D ViT which provides class specific visualization for attention models [37]. By altering the 2D convolutional operation confined within the model's patch embedding layer to 3D, the model has been modified to accept 3D medical image inputs. The ViT converts medical images of pre-specified sizes into tokens based upon pre-specified volume sizes (i.e., where each token represents an image patch). Following projection, the positional + patch embeddings are passed through a series of self-attention and transformer-encoder layers, culminating in a multilayer perceptron which determines the image's probability associated with belonging to [0, N) classes (Figure 10). As such, the embedding process and classification task are unified into a single, cohesive model unlike the image transformers we presented earlier.

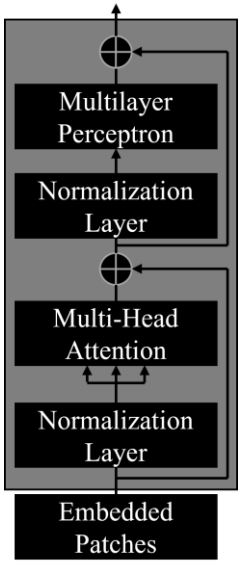

Figure 11: Transformer Encoder

Illustrates the transformer encoder architecture. The patch embeddings are passed through 1) a normalization layer, 2) a multi-head attention module, 3) before combining with the original input via skip connection, 4) after which the input passes through a secondary normalization layer, 5) as well as a multilayer perceptron, and 6) finally combining with the intermediate via the second skip connection. Vision transformers (ViTs) employ self-attention, a mechanism which relates embeddings with respect to each other.

The attention mechanism is based upon the notion that certain pieces of information are more instructive and relevant than other pieces of information [38]. In the field of NLP, attention mechanisms allow models to focus on specific sets of words within a sentence. However, in contrast to attention which focuses on which embeddings relate to the desired task, ViTs utilize self-attention which considers how embeddings relate with respect to each other (Figure 11). Leveraging the attention layers, we indicate the anatomical regions of the medical image the model concentrates upon when forming its prediction in a class specific manner.

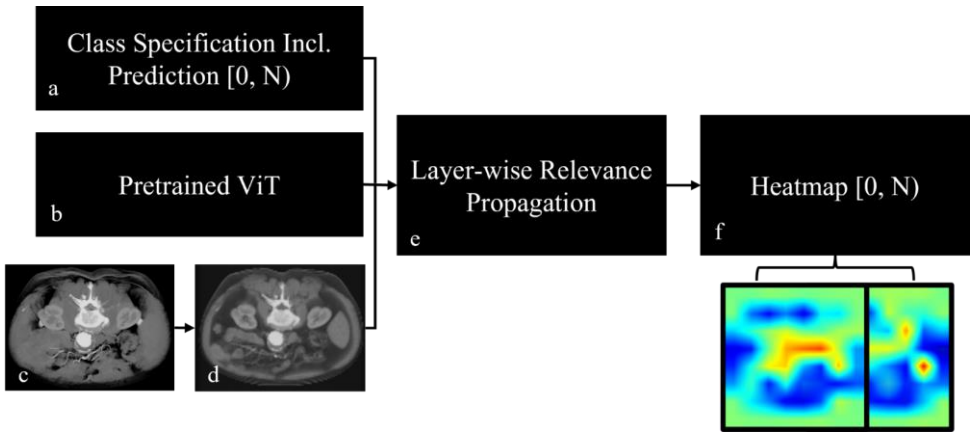

Figure 12: Class Specific Heatmaps

Details the heatmap generation procedure where a) represents the input class or model prediction from one of [0, N) classes; b) is the pretrained 3D ViT; c) is the original patient CTA; d) is the same resampled to 128x128x64 voxels; e) applies the layer-wise relevance propagation based upon the input; and f) indicates the class specified heatmap [0, N) which is generated. Note, however, that since the heatmaps are class-specific, multiple heatmaps of the same image can be generated regardless of the model's prediction.

Heatmap visualization can involve three separate aspects: 1) gradient-based propagation in which the gradient is calculated with respect to the previous input layer, a procedure known as backpropagation; 2) attribution-based propagation which considers the contribution of all previous layers, including the input; and 3) attention scores which are derived from the self-attention head. Gradient-based propagation are class insensitive and attention scores do not consider the information obtained throughout the model [37]. Layer-wise relevance propagation (LRP), an attribution-based technique, underpinned on the Deep Taylor Decomposition, propagates relevance with respect to the predicted class [39, 40]. The method described by the heatmap algorithm's original authors, and consequently used in our own study, implements a LRP inspired relevance to calculate the scores for each attention head of the transformer's layers. By considering the relevancy and gradient information while minimizing negative contributions, the scores are

combined across the attention graph. In sum, a class-specific visualization (i.e., not constrained to the predicted class) can be provided from the pretrained 3D ViT (Figure 12) [37].

AAA-ViT was trained upon the 128x128x64 z-normalized CTAs with the class labels serving as the target and was fit according to the cross-entropy loss. The model employed a patch size of 16x16x8, embedding dimension 256, depth 12, number of heads 8, multilayer perceptron ratio 8, dropout probability 10%, and attention dropout probability 10%. Model training parameters were specified as batch size 12, learning rate 5e-6, epochs 200, and optimizer AdamW with weight decay 5e-4. To prevent the model from overfitting, image augmentation consisted of random rotations, flipping, and noise.

3.7.3 Performance Evaluation

The softmax function was applied to AAA-ViT's predicted, unnormalized logits to ensure the probability of the negative class was p and the same for the positive class was $1 - p$ per patient-specific image. Similar to the AAA Image Transformer study, we compute the AUC between the probability of the positive class and the class labels, and later search for the OPT as defined by the training set, collecting the aforementioned performance metrics in the process.

For patient-specific heatmap generation, we provide the LRP based algorithm with the 128x128x64 z-normalized CTA, pretrained AAA-ViT, and specify the negative class label. The method produces several tokens which are reshaped based upon the model input dimensions and patch size, or 8x8x8 as specified in our study. These 8x8x8 tokens are then scaled to the model input dimensions of 128x128x64 using trilinear interpolation. Using b-spline interpolation as well as the information provided by the original patient CTA, the heatmap is set to the identical frame of reference as the original patient CTA, 512x512 where the z-dimension varied, before being

normalized between 0 and 1. This process is repeated for the positive class, generating two class-specific heatmaps per patient image. Note that by default, the heatmap of the model predicted class will be produced. The average precision (AP) score, a threshold agnostic measurement, is computed between the heatmaps and the labels inclusive of the AAA lumen, ILT, and wall to determine the anatomical regions the model focuses upon.

4.0 Results

Note, the DSC ranges from 0 to 1 where 1 is considered perfectly accurate. The whiskers of the boxplot indicate the distribution of the data, except for outliers identified via the inter-quartile range (i.e., Q1 – 1.5*IQR & Q3 + 1.5*IQR where Q1 is the first quartile, Q3 is the third quartile, and IQR is the interquartile range).

4.1 AAA-UNet: Baseline Aneurysm Segmentation

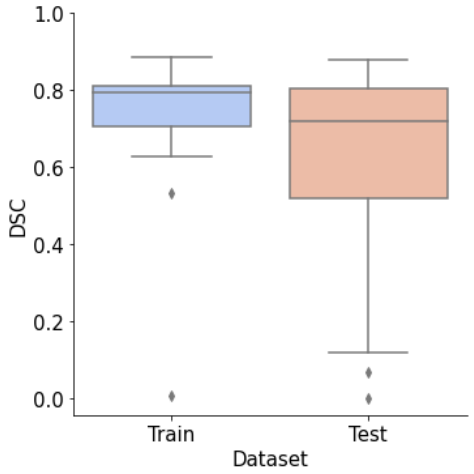

Figure 13: Train & Test AAA Lumen DSC for AAA-UNet

Displays the dice score coefficient (DSC) performance of AAA-UNet. The training set with 24 patients achieved mean 0.74, std 0.18, Q1 0.71, median 0.79, and Q3 0.81. The test set with 30 patients achieved mean 0.63, std 0.25, Q1 0.52, median 0.72, and Q3 0.81.

The model achieved mean 0.74 and median 0.79 DSCs for the training set while the same for the test set were 0.63 and 0.72, respectively (Figure 13).

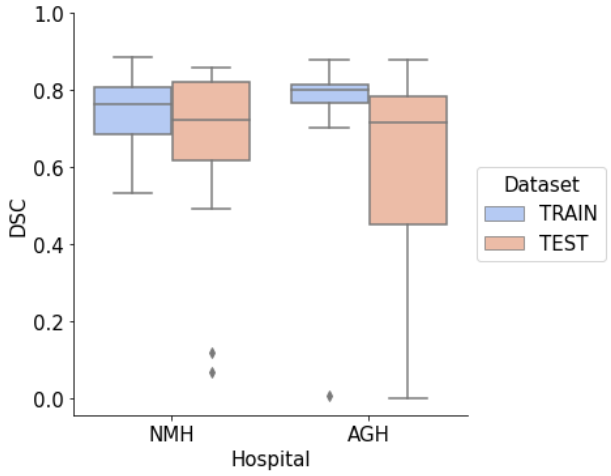

Figure 14: Train & Test AAA Lumen DSC by Hospital for AAA-UNet

Displays the dice score coefficient (DSC) performance of AAA-UNet for each hospital subgroup. For NMH, the training set with 13 patients achieved mean 0.74, std 0.10, Q1 0.69, median 0.76, and Q3 0.81 while the test set with 15 patients achieved mean 0.65, std 0.25, Q1 0.62, median 0.72, and Q3 0.82. For AGH, the training set with 11 patients achieved mean 0.73, std 0.24, Q1 0.77, median 0.80, and Q3 0.81 while the test set with 15 patients achieved mean 0.60, std 0.25, Q1 0.45, median 0.72, and Q3 0.78.

For NMH, the model achieved DSCs of mean 0.74 and median 0.76 for the training set while the same for the test set were 0.65 and 0.72, respectively. For AGH, the training set achieved DSCs of mean 0.73 and median 0.80 while the same for the test set were 0.60 and 0.72, respectively (Figure 14).

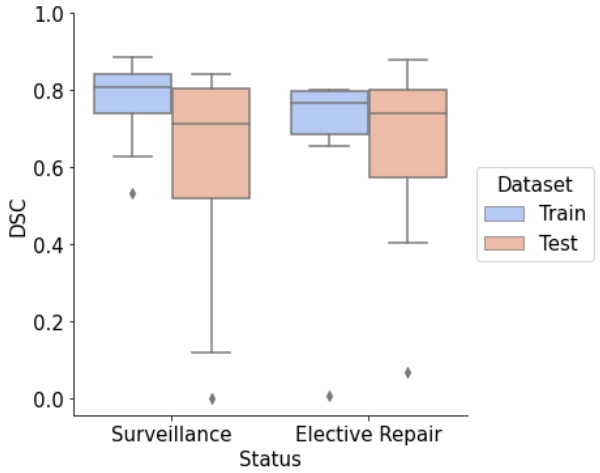

Figure 15: Train & Test AAA Lumen DSC by Treatment for AAA-UNet

Displays the dice score coefficient (DSC) performance of AAA-UNet for each treatment indication. For the surveillance subgroup, the training set with 15 patients achieved mean 0.78, std 0.10, Q1 0.74, median 0.81, and Q3 0.84 while the test set with 18 patients achieved mean 0.61, std 0.26, Q1 0.52, median 0.71, and Q3 0.81. For the elective repair subgroup, the training set with 9 patients achieved mean 0.67, std 0.25, Q1 0.69, median 0.77, and Q3 0.80 while the test set with 12 patients achieved mean 0.65, std 0.24, Q1 0.57, median 0.74, and Q3 0.80.

For surveillance, the model achieved DSCs of mean 0.78 and median 0.81 while the same for the test set were 0.61 and 0.71, respectively. For elective repair, the training set achieved mean 0.67 and median 0.77 while the same for the test set were 0.65 and 0.74, respectively (Figure 15).

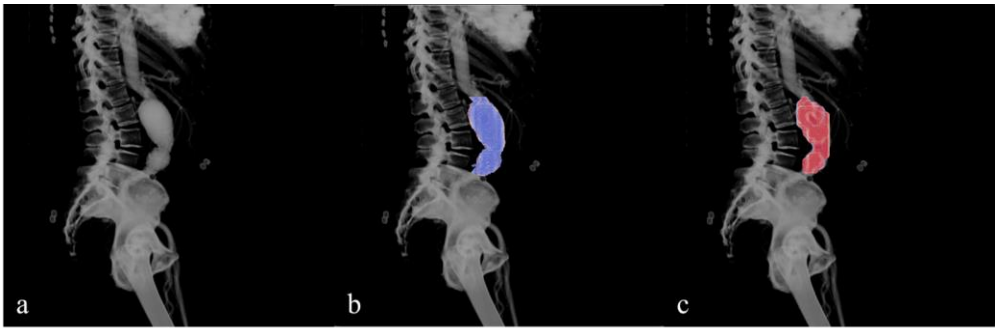

Figure 16: AAA Lumen Segmentation Result for AAA-UNet (Surveillance)

Illustrates the AAA-UNet test set performance where a) depicts a computed tomography angiography (CTA) image recommended for surveillance at NMH; b) depicts the ground truth abdominal aortic aneurysm (AAA) lumen label; and c) depicts the predicted segmentation of the previous which achieved dice score coefficient (DSC) of 0.84.

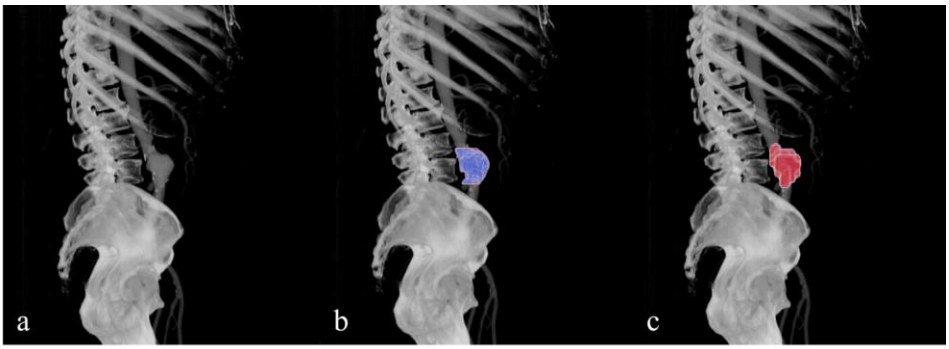

Figure 17: AAA Lumen Segmentation Result for AAA-UNet (Elective Repair)

Illustrates the AAA-UNet test set performance where a) depicts a computed tomography angiography (CTA) image recommended for elective repair at AGH; b) depicts the ground truth abdominal aortic aneurysm (AAA) lumen label; and c) depicts the predicted segmentation of the previous which achieved dice score coefficient (DSC) of 0.88.

The predicted segmentations appear to suffer from the effects of low image resolution and interpolation as seen by the ill-defined boundaries (Figures 16 and 17).

4.2 BB-AAA-UNet: Memory Efficient High-Resolution Segmentation with Prior Aneurysm Localization

Note, the DSC ranges from 0 to 1 where 1 is considered perfectly accurate. The whiskers of the boxplot indicate the distribution of the data, except for the outliers.

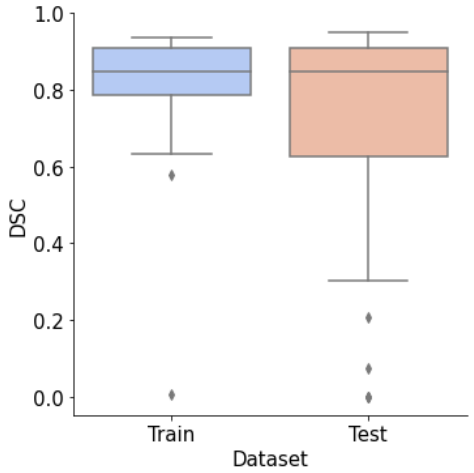

Figure 18: Train and Test AAA Lumen DSC for BB-AAA-UNet

Displays the dice score coefficient (DSC) performance of BB-AAA-UNet. The training set with 24 patients achieved mean 0.80, std 0.19, Q1 0.79, median 0.85, and Q3 0.91. The test set with 30 patients achieved mean 0.70, std 0.33, Q1 0.63, median 0.85, and Q3 0.91.

The model achieved DSCs of mean 0.80 and median 0.85 for the training set and the same for the test set were 0.70 and 0.85, respectively (Figure 18).

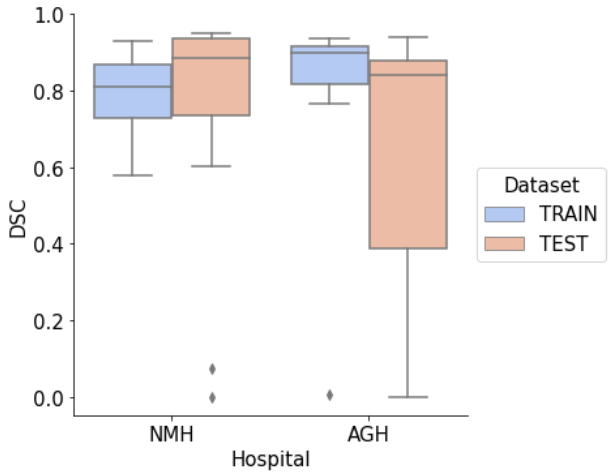

Figure 19: Train & Test AAA Lumen DSC by Hospital for BB-AAA-UNet

Displays the dice score coefficient (DSC) performance of BB-AAA-UNet for each hospital subgroup. For NMH, the training set with 13 patients achieved mean 0.80, std 0.11, Q1 0.73, median 0.81, and Q3 0.87 while the test set with 15 patients achieved mean 0.75, std 0.30, Q1 0.74, median 0.89, and Q3 0.94. For AGH, the training set with 11 patients achieved mean 0.80, std 0.27, Q1 0.82, median 0.90, and Q3 0.92 while the test set with 15 patients achieved mean 0.65, std 0.35, Q1 0.39, median 0.84, Q3 0.88.

For NMH, the model achieved DSCs of mean 0.80 and median 0.81 for the training set while the same for the test set were 0.75 and 0.89, respectively. For AGH, the model achieved DSCs of mean 0.80 and median 0.90 while the same for the test set were 0.65 and 0.84, respectively (Figure 19).

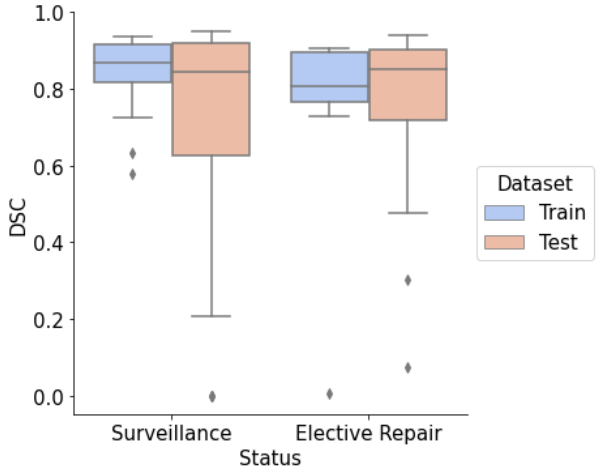

Figure 20: Train & Test AAA Lumen DSC by Treatment for BB-AAA-UNet

Displays the dice score coefficient (DSC) performance of BB-AAA-UNet for each treatment indication. For the surveillance subgroup, the training set with 15 patients achieved mean 0.84, std 0.11, Q1 0.82, median 0.87, and Q3 0.92 while the test set with 18 patients achieved mean 0.67, std 0.36, Q1 0.63, median 0.84, and Q3 0.92. For the elective repair subgroup, the training set with 9 patients achieved mean 0.73, std 0.28, Q1 0.77, median 0.81, and Q3 0.89 while the test set with 12 patients achieved mean 0.73, std 0.29, Q1 0.72, median 0.85, and Q3 0.90.

For surveillance, the model achieved DSCs of mean 0.84 and median 0.87 for the training set while the same for the test set were 0.67 and 0.84, respectively. For elective repair, the model achieved DSCs of mean 0.73 and median 0.81 while the same for the test were 0.73 and median 0.85, respectively (Figure 20).

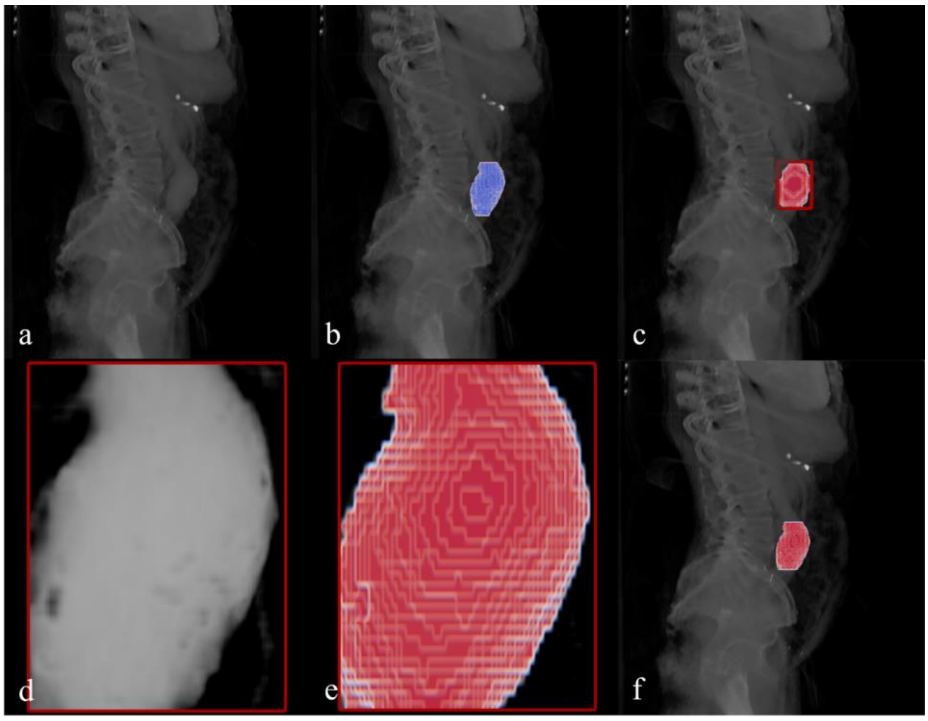

Figure 21: AAA Lumen Segmentation Result for BB-AAA-UNet (Surveillance)

Illustrates the BB-AAA-UNet test set performance where a) depicts the computed tomography angiography (CTA) image recommended for surveillance at NMH; b) depicts the ground truth abdominal aortic aneurysm (AAA) lumen label; c) depicts BB-AAA-UNet's localized predicted segmentation of the AAA lumen, intraluminal thrombus (ILT), and calcifications; d) depicts the bounding box of the previous predicted segmentation applied to the CTA; e) depicts BB-AAA-UNet's high resolution predicted segmentation of the AAA lumen; and f) depicts the re-insertion of the high-resolution segmentation into the original CTA frame of reference, achieving a dice score coefficient (DSC) of 0.95.

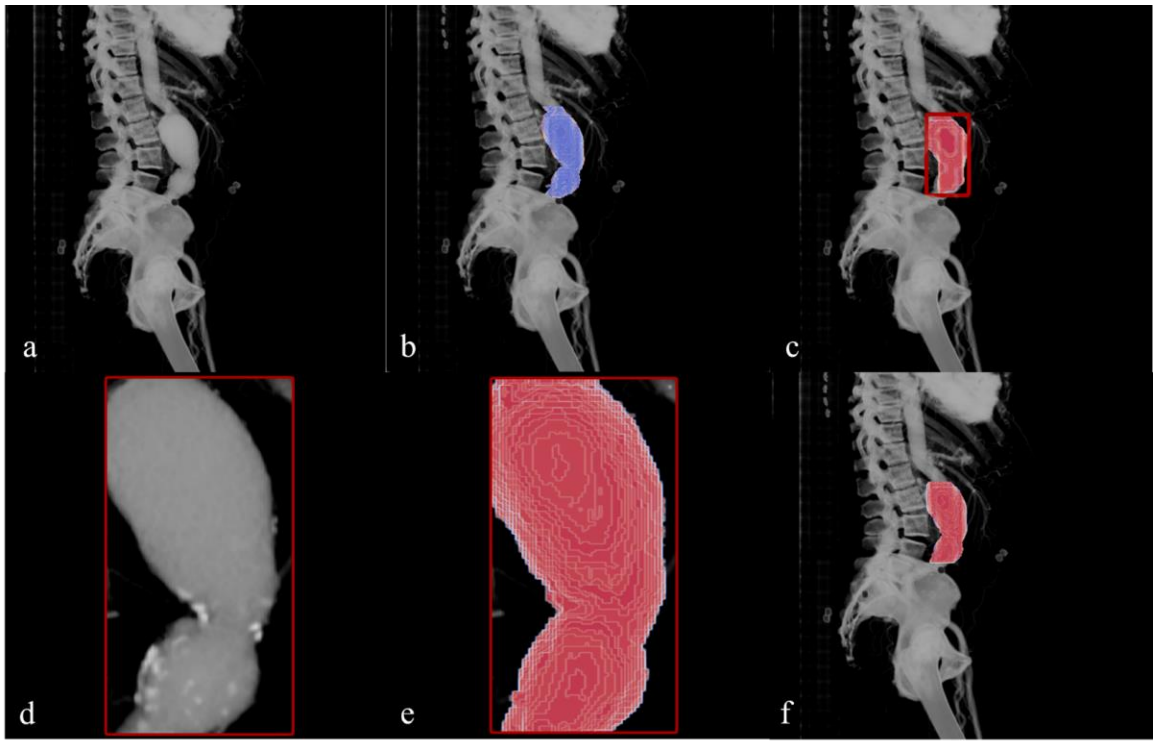

Figure 22: AAA Lumen Segmentation Result for BB-AAA-UNet (Elective Repair)

Illustrates the BB-AAA-UNet test set performance where a) depicts the computed tomography angiography (CTA) image recommended for elective repair at AGH; b) depicts the ground truth abdominal aortic aneurysm (AAA) lumen label; c) depicts BB-AAA-UNet's localized predicted segmentation of the AAA lumen, intraluminal thrombus (ILT), and calcifications; d) depicts the bounding box of the previous predicted segmentation applied to the CTA; e) depicts BB-AAA-UNet's high resolution predicted segmentation of the AAA lumen; and f) depicts the re-insertion of the high-resolution segmentation into the original CTA frame of reference, achieving a dice score coefficient (DSC) of 0.94.

The predicted segmentations are of higher quality owing to the mitigation of interpolation effects; however, small regions of the AAA lumen appear flattened due to the bounding box (Figures 22 and 23). For the Anderson Darling test, both statistics associated with AAA-UNet and BB-AAA-UNet test set DSCs were found to be above the respective critical values for a = 5e-2. . The same was true for the Mann-Whitney U and Wilcoxon rank-sum tests with *p-values* < 1e-2.

4.3 BB-AAA-UNet: As Applied to Aneurysm Wall Segmentation

Note, the DSC ranges from 0 to 1 where 1 is considered perfectly accurate. The whiskers of the boxplot indicate the distribution of the data, except for the outliers.

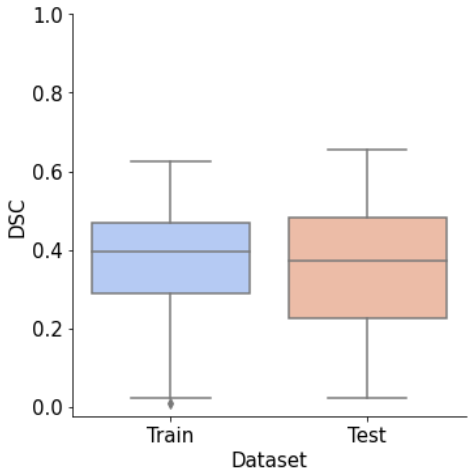

Figure 23: Train & Test AAA Wall DSC for BB-AAA-UNet

Displays the dice score coefficient (DSC) performance of BB-AAA-UNet. The training set with 73 patients achieved mean 0.37, std 0.15, Q1 0.29, median 0.39, and Q3 0.47. The test set with 50 patients achieved mean 0.35, std 0.16, Q1 0.23, median 0.37, and Q3 0.48.

The model achieved DSCs of mean 0.37 and median 0.39 for the training set and the same for the test set were 0.35 and 0.37, respectively (Figure 23).

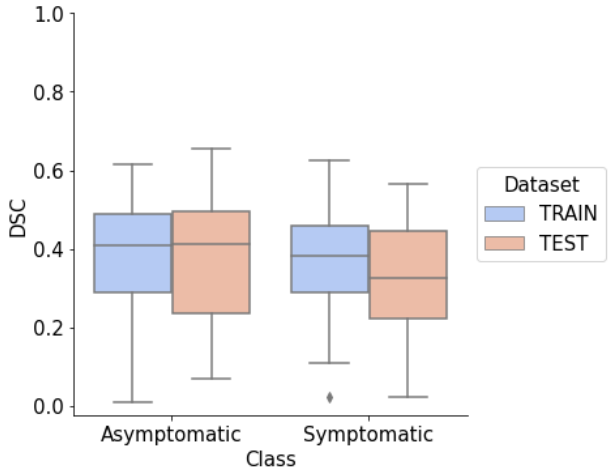

Figure 24: Train & Test AAA Wall DSC by Severity for BB-AAA-UNet

Displays the dice score coefficient (DSC) performance of BB-AAA-UNet based upon AAA severity. For the asymptomatic subgroup, the training set with 38 patients achieved mean 0.37, std 0.16, Q1 0.29, median 0.41, and Q3 0.49 while the test set with 25 patients achieved mean 0.37, std 0.17, Q1 0.24, median 0.41, and Q3 0.50. For the symptomatic subgroup, the training set with 35 patients achieved mean 0.36, std 0.14, Q1 0.29, median 0.38, and Q3 0.46 while the test set with 25 patients achieved mean 0.32, std 0.16, Q1 0.22, median 0.33, Q3 0.45.

For the asymptomatic subgroup, the model achieved DSCs of mean 0.37 and median 0.41 for the training set while the same for the test set were 0.37 and 0.41, respectively. For the symptomatic subgroup, the model achieved DSCs of mean 0.36 and median 0.38 while the same for the test set were 0.32 and 0.33, respectively (Figure 24).

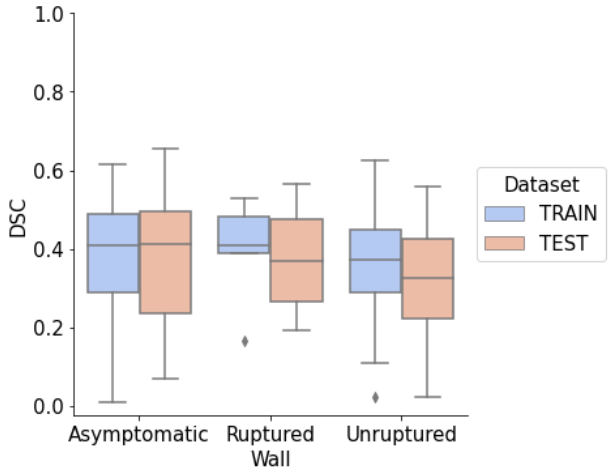

Figure 25: Train & Test AAA Wall DSC by Rupture for BB-AAA-UNet

Displays the dice score coefficient (DSC) performance of BB-AAA-UNet based upon the wall rupture status for the symptomatic subgroup with asymptomatic provided as reference. For the asymptomatic subgroup, the training set with 38 patients achieved mean 0.37, std 0.16, Q1 0.29, median 0.41, and Q3 0.49 while the test set with 25 patients achieved mean 0.37, std 0.17, Q1 0.24, median 0.41, and Q3 0.50. For the ruptured wall subgroup, the training set with 5 patients achieved mean 0.39, std 0.14, Q1 0.39, median 0.41, and Q3 0.48 while the test set with 4 patients achieved mean 0.37, std 0.17, Q1 0.26, median 0.37, and Q3 0.48. For the unruptured wall subgroup, the training set with 30 patients mean 0.35, std 0.14, Q1 0.29, median 0.37, and Q3 0.45 while the test set with 21 patients achieved mean 0.31, std 0.16, Q1 0.22, median 0.33, and Q3 0.43.

For the ruptured wall subgroup, the model achieved DSCs of mean 0.39 and median 0.41 for the training set while the same for the test set were 0.37 and 0.37, respectively. For the unruptured wall subgroup, the model achieved DSCs of mean 0.35 and median 0.37 while the same for the test set were 0.31 and 0.33, respectively (Figure 25).

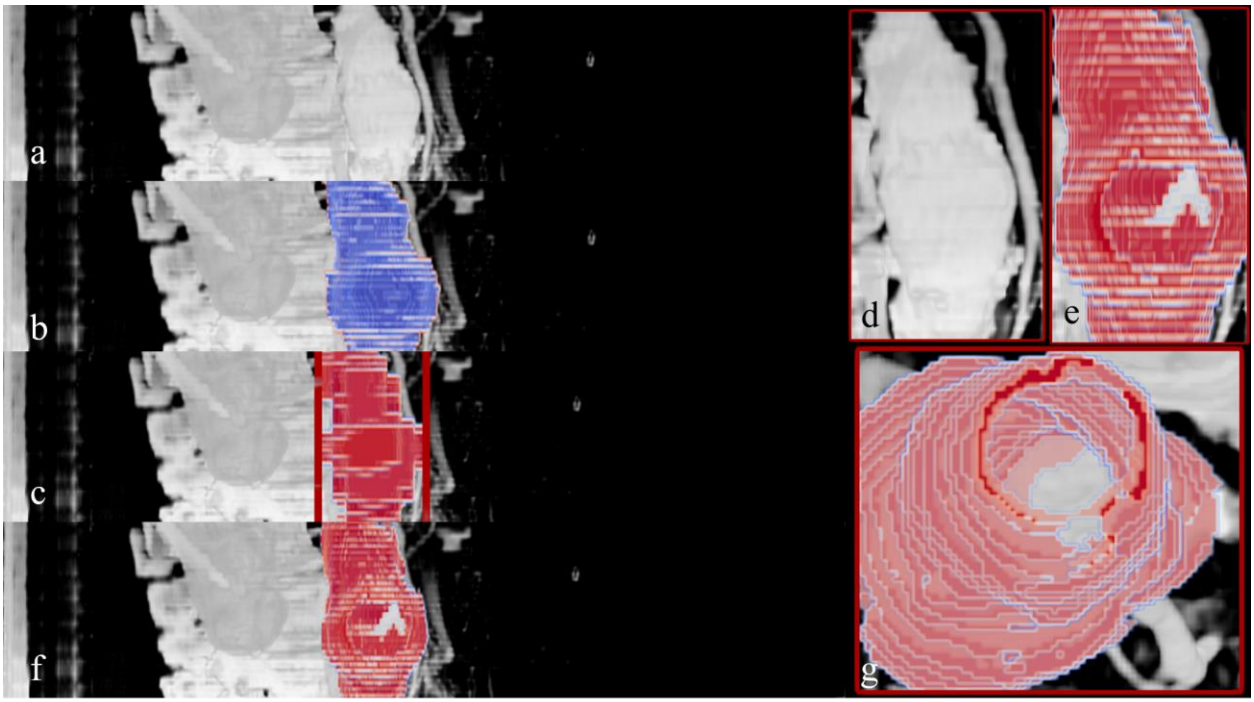

Figure 26: AAA Wall Segmentation Result of BB-AAA-UNet (Asymptomatic)

Illustrates the BB-AAA-UNet test set performance where a) depicts the computed tomography angiography (CTA) image labelled as asymptomatic; b) depicts the ground truth abdominal aortic aneurysm (AAA) wall; c) depicts BB-AAA-UNet's localized predicted segmentation of the AAA lumen, intraluminal thrombus (ILT), and wall; d) depicts the bounding box of the previous predicted segmentation applied to the CTA; e) depicts BB-AAA-UNet's high resolution predicted segmentation of the AAA wall; f) depicts the re-insertion of the high-resolution segmentation into the original CTA frame of reference, achieving a dice score coefficient (DSC) of 0.65; and g) highlights the previous from the top.

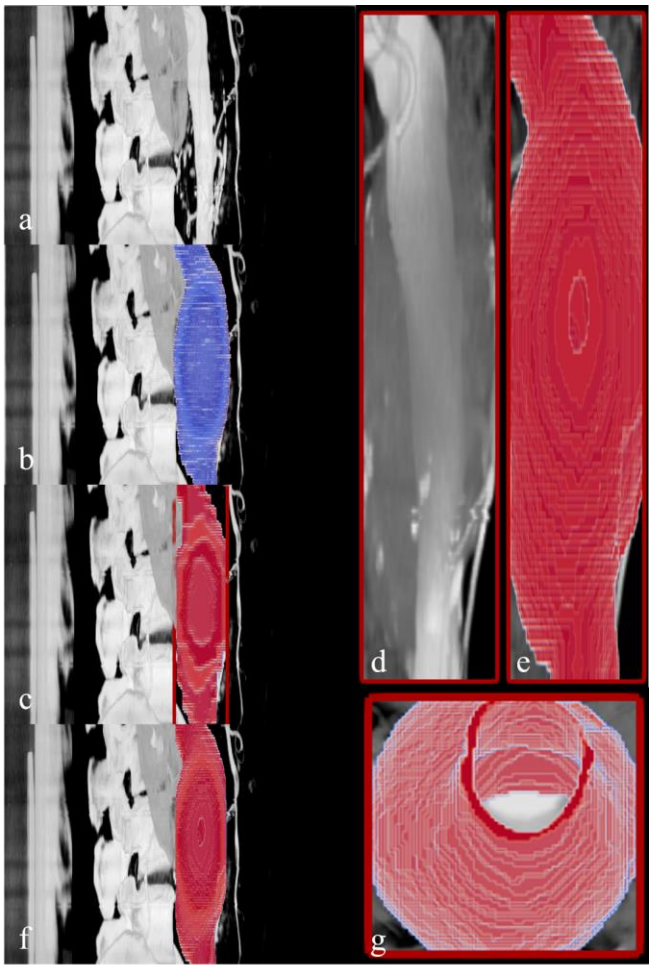

Figure 27: AAA Wall Segmentation Result of BB-AAA-UNet (Symptomatic, Ruptured)

Illustrates the BB-AAA-UNet test set performance where a) depicts the computed tomography angiography (CTA) image labelled as symptomatic, ruptured wall; b) depicts the ground truth abdominal aortic aneurysm (AAA) wall; c) depicts BB-AAA-UNet's localized predicted segmentation of the AAA lumen, intraluminal thrombus (ILT), and wall; d) depicts the bounding box of the previous predicted segmentation applied to the CTA; e) depicts BB-AAA-UNet's high resolution predicted segmentation of the AAA wall; f) depicts the re-insertion of the high-resolution segmentation into the original CTA frame of reference, achieving a dice score coefficient (DSC) of 0.56; and g) highlights the previous from the top.

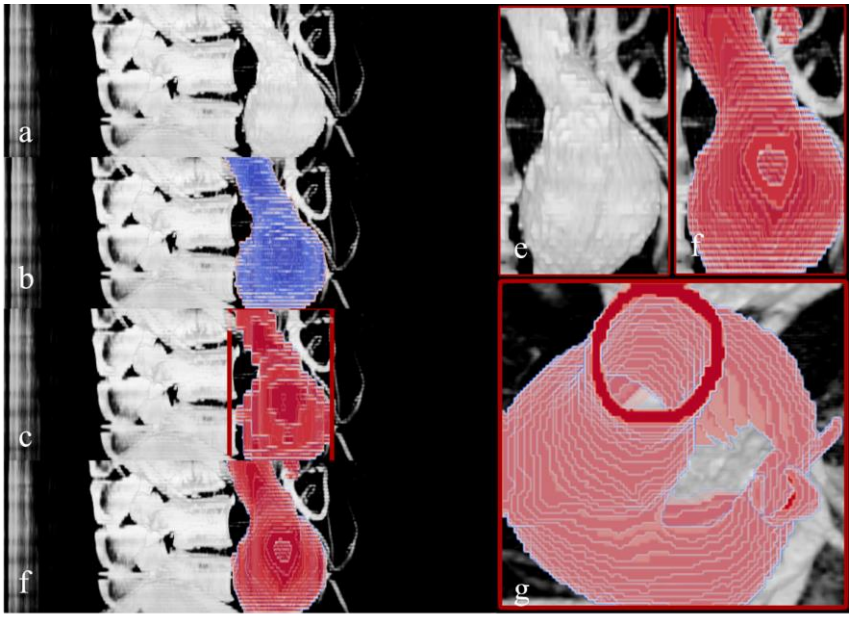

Figure 28: AAA Wall Segmentation Result of BB-AAA-UNet (Symptomatic, Unruptured)

Illustrates the BB-AAA-UNet test set performance where a) depicts the computed tomography angiography (CTA) image labelled as symptomatic, unruptured wall; b) depicts the ground truth abdominal aortic aneurysm (AAA) wall; c) depicts BB-AAA-UNet's localized predicted segmentation of the AAA lumen, intraluminal thrombus (ILT), and wall; d) depicts the bounding box of the previous predicted segmentation applied to the CTA; e) depicts BB-AAA-UNet's high resolution predicted segmentation of the AAA wall; f) depicts the re-insertion of the high-resolution segmentation into the original CTA frame of reference, achieving a dice score coefficient (DSC) of 0.56; and g) highlights the previous from the top.

The predicted segmentations are consistent with the outer wall of the AAA. However, the bounding box which is responsible for reducing the effects of interpolation appears to cut the volume of interest (Figures 26, 27, and 28).

4.4 Patch Segmentation UNet: Prediction of Aneurysm Wall by Medical Image Sub-volumes

Note, the DSC ranges from 0 to 1 where 1 is considered perfectly accurate. The whiskers of the boxplot indicate the distribution of the data, except for the outliers.

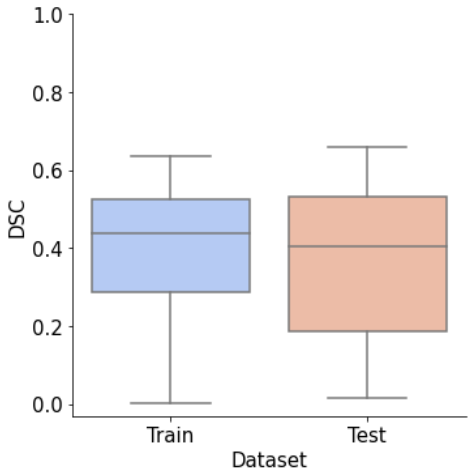

Figure 29: AAA Wall Train & Test DSC for Patch Segmentation UNet

Displays the dice score coefficient (DSC) performance of the Patch Segmentation UNet . The training set with 73 patients achieved mean 0.40, std 0.17, Q1 0.29, median 0.44, and Q3 0.53. The test set with 50 patients achieved mean 0.37, std 0.19, Q1 0.19, median 0.40, and Q3 0.53.

The model achieved DSCs of mean 0.40 and median 0.44 for the training set and the same for the test set were 0.37 and 0.40, respectively (Figure 29).

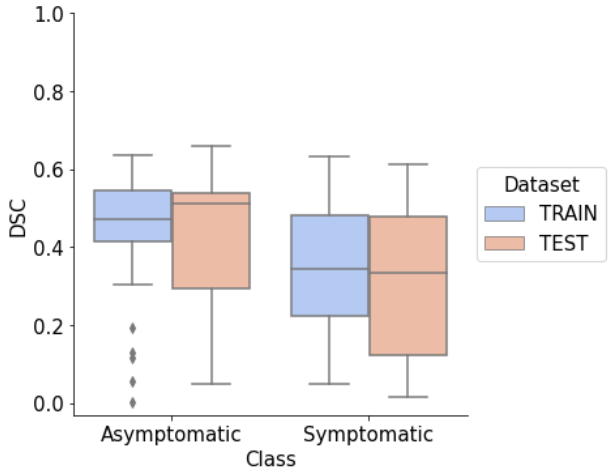

Figure 30: AAA Wall Train & Test DSC by Severity for Patch Segmentation UNet

Displays the dice score coefficient (DSC) performance of the Patch Segmentation UNet based upon AAA severity. For the asymptomatic subgroup, the training set with 38 patients achieved mean 0.44, std 0.16, Q1 0.41, median 0.47, and Q3 0.55 while the test set with 25 patients achieved mean 0.42, std 0.19, Q1 0.29, median 0.51, and Q3 0.54. For the symptomatic subgroup, the training set with 35 patients achieved mean 0.35, std 0.17, Q1 0.22, median 0.34, and Q3 0.48 while the test set with 25 patients achieved mean 0.32, std 0.19, Q1 0.12, median 0.33, Q3 0.48.

For the asymptomatic subgroup, the model achieved DSCs of mean 0.44 and median 0.47 for the training set while the same for the test set were 0.42 and 0.51, respectively. For the symptomatic subgroup, the model achieved DSCs of mean 0.35 and median 0.34 while the same for the test set were 0.32 and 0.33, respectively (Figure 30).

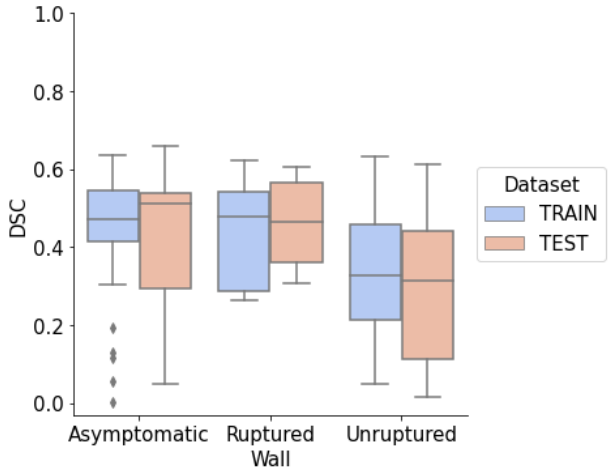

Figure 31: AAA Wall Train & Test DSC by Rupture for Patch Segmentation UNet

Displays the dice score coefficient (DSC) performance of the Patch Segmentation UNet based upon the wall rupture status for the symptomatic subgroup with asymptomatic provided as reference. For the asymptomatic subgroup, the training set with 38 patients achieved mean 0.44, std 0.16, Q1 0.41, median 0.47, and Q3 0.55 while the test set with 25 patients achieved mean 0.42, std 0.19, Q1 0.29, median 0.51, and Q3 0.54. For the ruptured wall subgroup, the training set with 5 patients achieved mean 0.44, std 0.16, Q1 0.29, median 0.48, and Q3 0.54 while the test set with 4 patients achieved mean 0.46, std 0.14, Q1 0.36, median 0.46, and Q3 0.56. For the unruptured wall subgroup, the training set with 30 patients mean 0.34, std 0.17, Q1 0.21, median 0.33, and Q3 0.46 while the test set with 21 patients achieved mean 0.30, std 0.19, Q1 0.11, median 0.31, and Q3 0.44.

For the ruptured wall subgroup, the model achieved DSCs of mean 0.44 and median 0.48 for the training set while the same for the test set were 0.46 and 0.46, respectively. For the unruptured wall subgroup, the model achieved DSCs of mean 0.34 and median 0.33 while the same for the test set were 0.30 and 0.31, respectively (Figure 31).

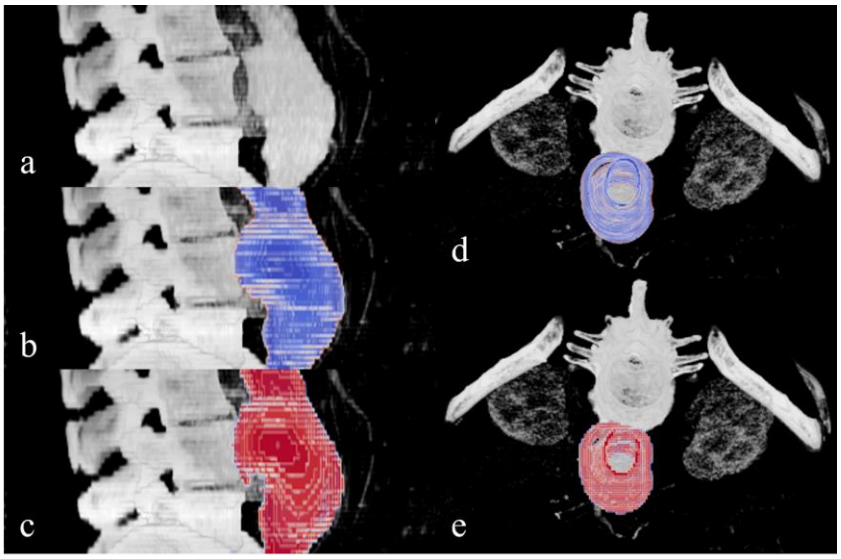

Figure 32: AAA Wall Segmentation Result of Patch Segmentation UNet (Asymptomatic)

Illustrates the Patch Segmentation UNet test set performance where a) depicts the computed tomography angiography (CTA) image that was labelled as asymptomatic; b) depicts the ground truth abdominal aortic aneurysm (AAA) wall; c) depicts the Patch Segmentation UNet's prediction; d) displays the top view of the ground truth AAA wall; e) displays the top view of the Patch Segmentation UNet's prediction. The model's predicted segmentation achieved a dice score coefficient (DSC) of 0.66.

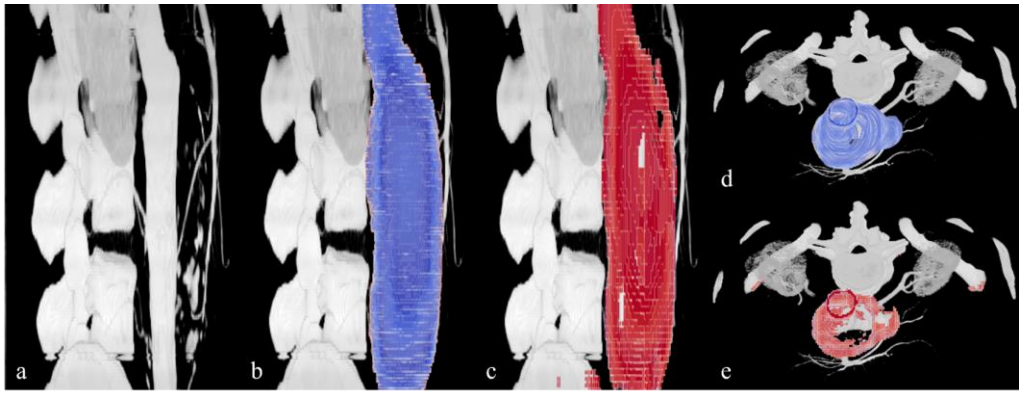

Figure 33: AAA Wall Segmentation Result of Patch Segmentation UNet (Symptomatic, Unruptured)

Illustrates the Patch Segmentation UNet test set performance where a) depicts the computed tomography angiography (CTA) image that was labelled as symptomatic, unruptured wall; b) depicts the ground truth abdominal aortic aneurysm (AAA) wall; c) depicts the Patch Segmentation UNet's prediction; d) displays the

top view of the ground truth AAA wall; e) displays the top view of the Patch Segmentation UNet's prediction. The model's predicted segmentation achieved a dice score coefficient (DSC) of 0.61.

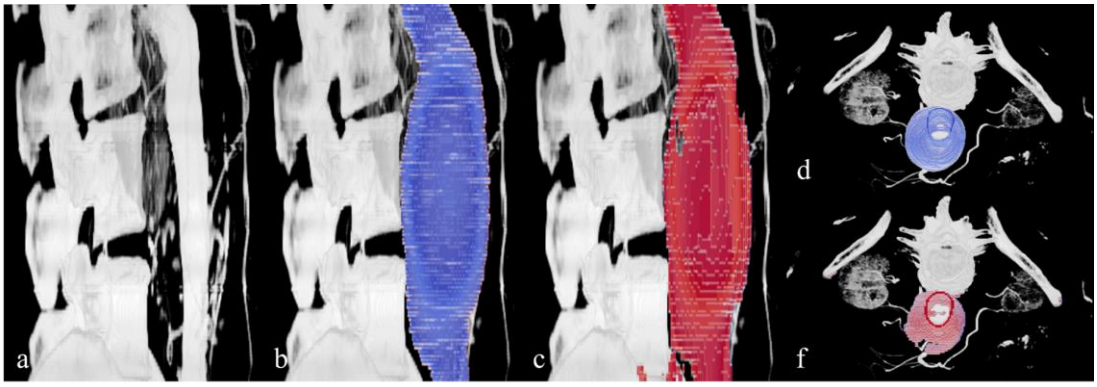

Figure 34: AAA Wall Segmentation Result of Patch Segmentation UNet (Symptomatic, Ruptured)

Illustrates the Patch Segmentation UNet test set performance where a) depicts the computed tomography angiography (CTA) image that was labelled as symptomatic, ruptured wall; b) depicts the ground truth abdominal aortic aneurysm (AAA) wall; c) depicts the Patch Segmentation UNet's prediction; d) displays the top view of the ground truth AAA wall; e) displays the top view of the Patch Segmentation UNet's prediction. The model's predicted segmentation achieved a dice score coefficient (DSC) of 0.61.

As before, the predicted segmentations are consistent with the outer wall of the AAA. However, despite the removal of the bounding box, the effects of interpolation are still present (Figures 32, 33, and 34). For the asymptomatic test set AAA wall DSCs, BB-AAA-UNet's Anderson Darling statistic was below the respective critical value while the same for Patch Segmentation UNet was above the respective critical value. The Mann-Whitney U and Wilcoxon rank-sum tests determined *p-values* = 0.21 and 0.20, respectively. For the symptomatic test set AAA wall DSCs, the Anderson Darling statistics of both BB-AAA-UNet Patch Segmentation UNet were below their respective critical values. The t-test determined *p-value* = 0.98.

4.5 AAA Image Transformers: Classifying Medical Images by Aneurysm Severity with Latent Representations

Asymptomatic cases were labelled as the negative class whereas the symptomatic cases were labelled as the positive class.

Table 1: Baseline Image Classification Performance

Details the classification performance of the image classifier on the AAA Wall dataset, specifically the 64x64x64 z-normalized computed tomography angiography (CTA) images. The optimal probability threshold (OPT) is noted.

Dataset	Probability Threshold	Precision	Negative Predictive Value	Recall	Specificty	F1	Accuracy
Train	0.59 (OPT)	0.71	0.77	0.77	0.71	0.74	0.74
	0.50	0.48	NA	1	0	0.65	0.48
Test	0.59 (OPT)	0.71	0.69	0.68	0.72	0.69	0.70
	0.50	0.50	NA	1	0	0.67	0.50

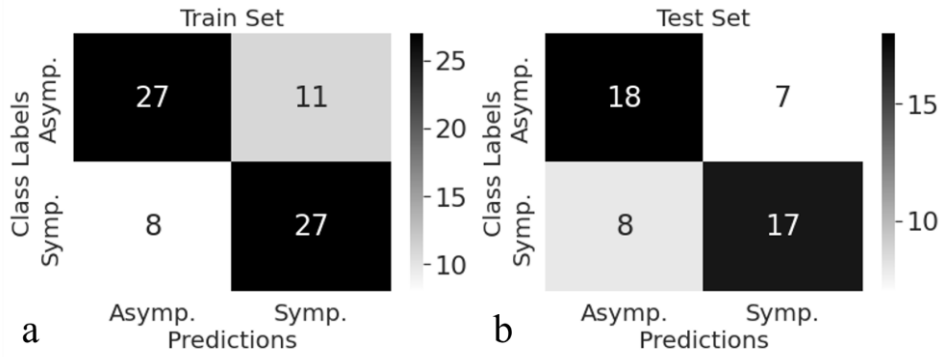

Figure 35: Baseline Image Classification by OPT

Details the classification performance by patient sample of the image classifier based upon the optimal probability threshold (OPT) of 0.59 where a) represents the training set; and b) represents the test set. The class labels are specified on the y-axis and the model predictions are specified on the x-axis.

For the baseline CTA image set, the classifier achieves AUCs of 0.80 and 0.74 for the train and test sets, respectively (Table 1 and Figure 35). Retrieving the OPT from the baseline image study significantly boosted the accuracy to ~70% for the test set.

Table 2: Class Agnostic Embedding Classification Performance

Details the classification performance of the class agnostic embedding classifier on the AAA Wall dataset, specifically the 64x64x32 z-normalized embeddings from the class agnostic image transformer. The optimal probability threshold (OPT) is noted.

Dataset	Probability Threshold	Precision	Negative Predictive Value	Recall	Specificty	F1	Accuracy
Train	0.47 (OPT)	0.71	0.81	0.83	0.68	0.76	0.75
	0.50	0.89	0.67	0.49	0.95	0.63	0.73
Test	0.47 (OPT)	0.63	0.70	0.76	0.56	0.69	0.66
	0.50	0.67	0.54	0.24	0.88	0.35	0.56

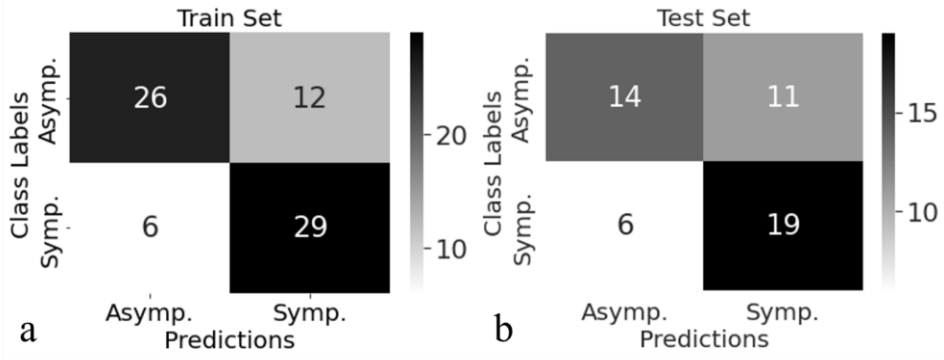

Figure 36: Class Agnostic Embedding Classification by OPT

Details the classification performance by patient sample of the class agnostic embedding classifier based upon the optimal probability threshold (OPT) of 0.47 where a) represents the training set; and b) represents the test set. The class labels are specified on the y-axis and the model predictions are specified on the x-axis.

For the embeddings derived from the class agnostic image transformer, the classifier achieves AUCs of 0.83 and 0.67 for the train and test sets, respectively (Table 2 and Figure 36).

Table 3: Asymptomatic Embedding Classification Performance

Details the classification performance of the asymptomatic embedding classifier on the AAA Wall dataset, specifically the 64x64x32 z-normalized embeddings from the asymptomatic image transformer. The optimal probability threshold (OPT) is noted.

Dataset	Probability Threshold	Precision	Negative Predictive Value	Recall	Specificty	F1	Accuracy
Train	0.43 (OPT)	0.88	0.83	0.80	0.89	0.84	0.85
	0.50	0.93	0.64	0.40	0.97	0.56	0.70
Test	0.43 (OPT)	0.76	0.76	0.76	0.76	0.76	0.76
	0.50	0.73	0.56	0.32	0.88	0.44	0.60

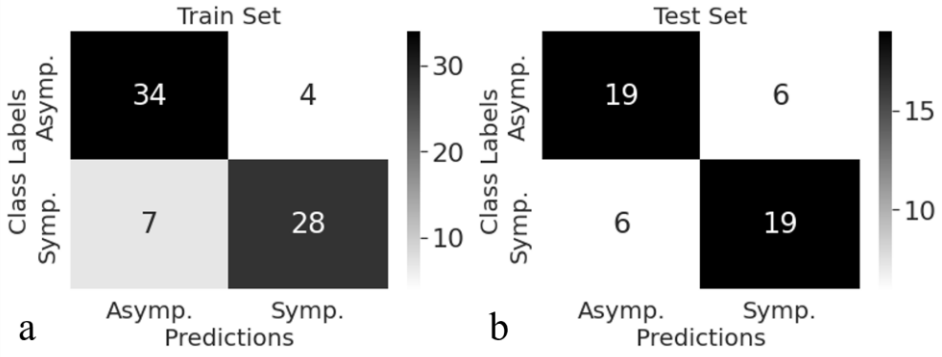

Figure 37: Asymptomatic Embedding Classification by OPT

Details the classification performance by patient sample of the asymptomatic embedding classifier based upon the optimal probability threshold (OPT) of 0.43 where a) represents the training set; and b) represents the test set. The class labels are specified on the y-axis and the model predictions are specified on the x-axis.

For the embeddings derived from the asymptomatic image transformer, the classifier achieves AUCs of 0.88 and 0.75 for the train and test sets, respectively (Table 3 and Figure 37).

The asymptomatic embedding classifier outperforms the baseline after application of the respective OPT, test set accuracy of 76% vs 70%. Though, there are signs of overfitting.

Table 4: Symptomatic Embedding Classification Performance

Details the classification performance of the symptomatic embedding classifier on the AAA Wall dataset, specifically the 64x64x32 z-normalized embeddings from the symptomatic image transformer. The optimal probability threshold (OPT) is noted.

Dataset	Probability Threshold	Precision	Negative Predictive Value	Recall	Specificty	F1	Accuracy
Train	0.60 (OPT)	0.84	0.79	0.74	0.87	0.79	0.81
	0.50	0.64	1	1	0.47	0.78	0.73
Test	0.60 (OPT)	0.68	0.68	0.68	0.68	0.68	0.68
	0.50	0.57	0.88	0.96	0.28	0.72	0.62

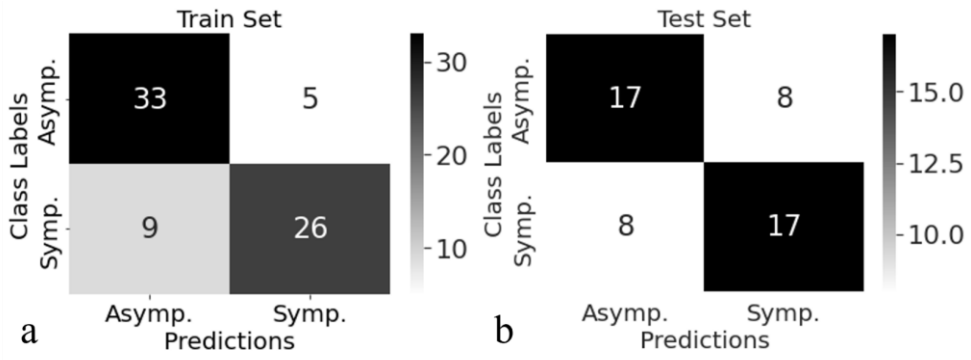

Figure 38: Symptomatic Embedding Classification by OPT

Details the classification performance by patient sample of the symptomatic embedding classifier based upon the optimal probability threshold (OPT) of 0.60 where a) represents the training set; and b) represents the test set. The class labels are specified on the y-axis and the model predictions are specified on the x-axis.

For the embeddings derived from the symptomatic image transformer exclusively, the classification model achieved AUCs of 0.89 and 0.72 for the train and test sets, respectively (Table

4 and Figure 38). As such, the asymptomatic image transformer outperformed its symptomatic counterpart.

4.6 AAA-ViT: Moving Towards Detection with Classification of Aneurysm Severity with Anatomical Explanation

Asymptomatic cases were labelled as the negative class whereas the symptomatic cases were labelled as the positive class.

Table 5: Classification Performance of the Vision Transformer

Details the classification performance of AAA-ViT on the AAA Wall dataset, specifically the 128x128x64 z-normalized CTAs. The optimal probability threshold (OPT) is noted. Asymptomatic cases were labelled as the negative class whereas the symptomatic cases were labelled as the positive class.

Dataset	Probability Threshold	Precision	Negative Predictive Value	Recall	Specificity	F1	Accuracy
Train	0.52 (OPT)	0.74	0.67	0.57	0.82	0.65	0.70
	0.50	0.69	0.68	0.63	0.74	0.66	0.68
Test	0.52 (OPT)	0.69	0.71	0.72	0.68	0.71	0.70
	0.50	0.69	0.71	0.72	0.68	0.71	0.70

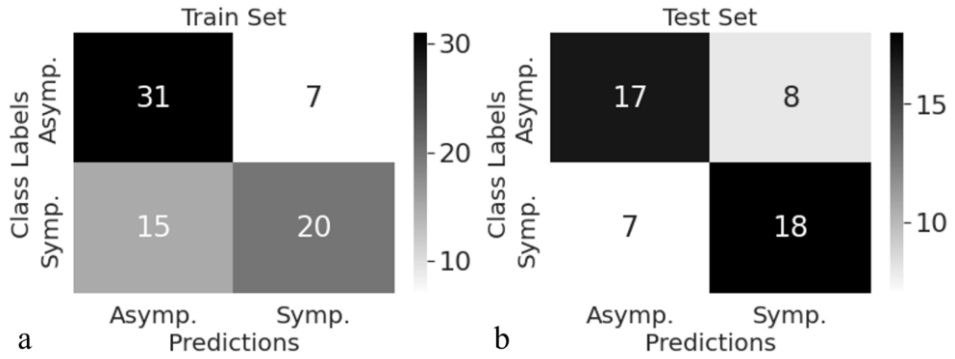

Figure 39: Vision Transformer Image Classification by OPT

Details the classification performance by patient sample of AAA-ViT based upon the optimal probability threshold (OPT) of 0.52 where a) represents the training set; and b) represents the test set. The class labels are specified on the y-axis and the model predictions are specified on the x-axis.

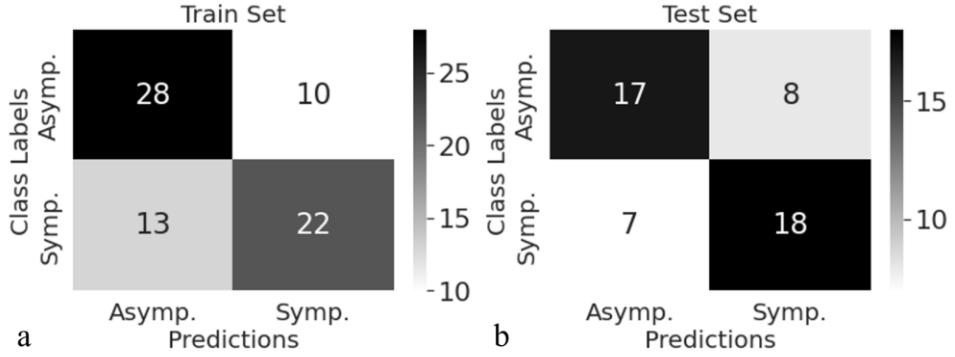

Figure 40: Vision Transformer Image Classification by Normal Threshold

Details the classification performance by patient sample of AAA-ViT based upon the probability threshold of 0.50 where a) represents the training set; and b) represents the test set. The class labels are specified on the y-axis and the model predictions are specified on the x-axis.

The 3D ViT achieved AUCs of 0.73 for both the train and test sets (Table 5 and Figures 39 and 40). The classification performance between the positive and negative classes of the train and tests sets remained balanced, though the overall accuracy was limited to ~70%. AAA-ViT indicated no signs of overfitting.

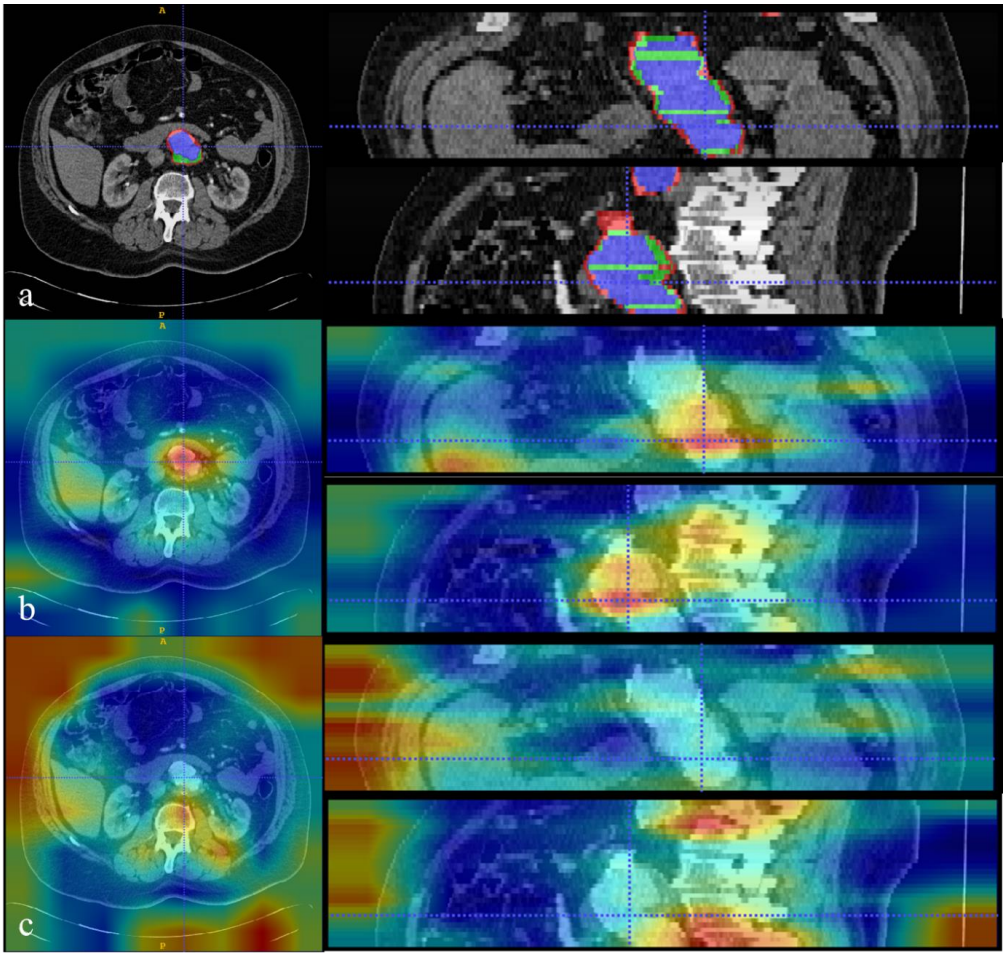

Figure 41: Heatmap Results of the AAA-ViT (Asymptomatic)

The computed tomography angiography (CTA) of a test set patient labelled as asymptomatic and predicted as asymptomatic at both the OPT and 0.50 threshold where a) represents the annotation of the abdominal aortic aneurysm (AAA) lumen in blue, intraluminal thrombus (ILT) in green, and AAA wall in red; b) represents the heatmap of the model predicted negative class, average precision (AP) was 0.14; and c) represents the heatmap of the positive class, AP was 0.01.

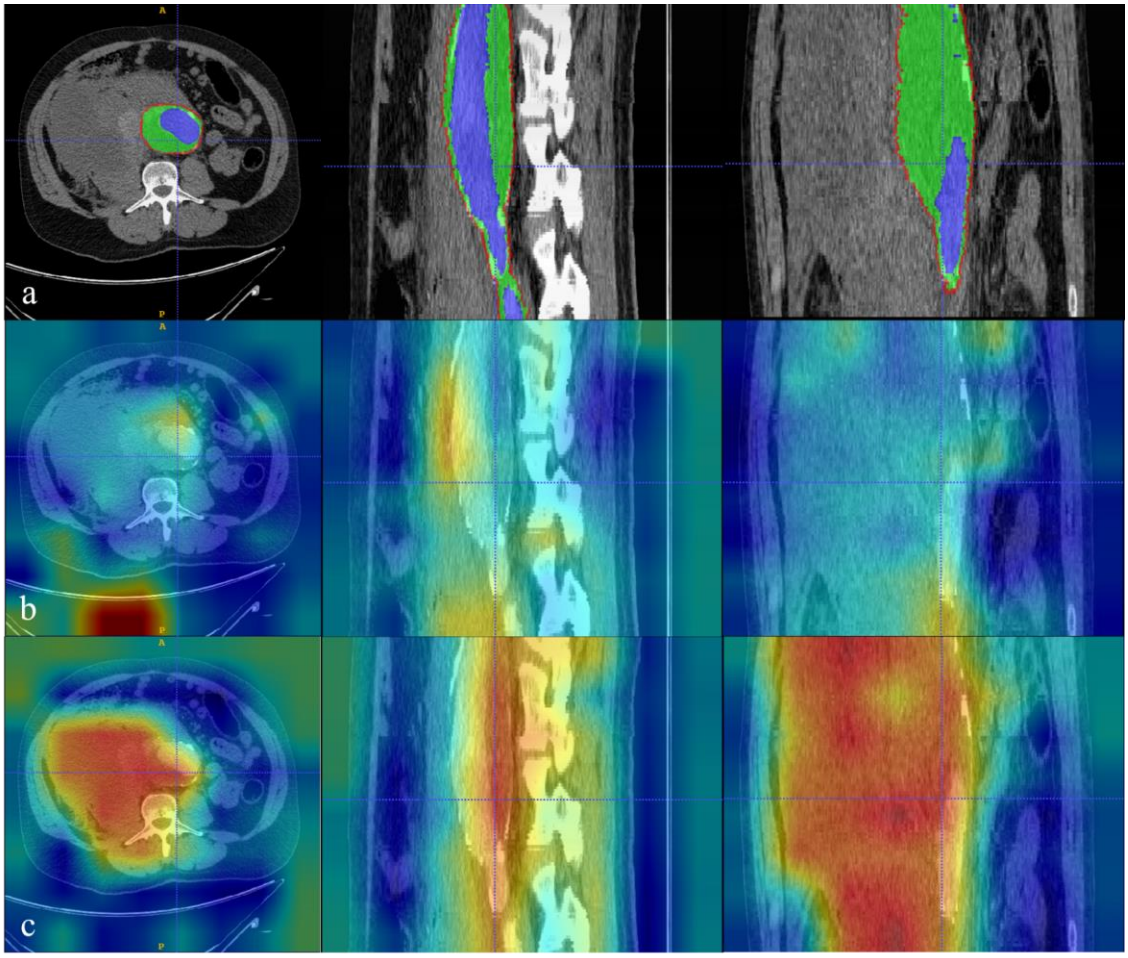

Figure 42: Heatmap Results of the AAA-ViT (Symptomatic)

The computed tomography angiography (CTA) of a test set patient labelled as symptomatic and predicted as symptomatic at both the OPT and 0.50 threshold where a) represents the annotation of the abdominal aortic aneurysm (AAA) lumen in blue, intraluminal thrombus (ILT) in green, and AAA wall in red; b) represents the heatmap of the negative class, average precision (AP) was 0.04; and c) represents the heatmap of the model predicted positive class, AP was 0.07.

The model successfully identified a region of the AAA itself and the accompanying ILT as seen in the heatmaps generated for the AAA-ViT predicted class (Figures 41 and 42). However, the model also considers the unrelated abdominal tissue and occasionally the image background.

Table 6: mean AP of AAA-ViT Heatmaps

Displays the mean average precision (AP) based upon the original class label as well as the heatmap class indication specified for the modified layer-wise relevance propagation (LRP). Note, the mean is taken over all the patient APs for each dataset.

Dataset	Class Label	Heatmap Class Indication	mean AP
Train	Asymptomatic	Positive	0.02
	Symptomatic	Positive	0.06
	Asymptomatic	Negative	0.03
	Symptomatic	Negative	0.07
Test	Asymptomatic	Positive	0.02
	Symptomatic	Positive	0.06
	Asymptomatic	Negative	0.03
	Symptomatic	Negative	0.06

Upon calculating the AP with respect to both a positive and negative class indication for the LRP based heatmap generation, it was observed that the performance remained quite low. Considering these metrics as well as the visual inspection, it can be inferred that the 3D ViT includes the AAA, but the regions of high intensities are not limited to the AAA (Table 6).

5.0 Discussion

5.1 AAA-UNet: Baseline Aneurysm Segmentation

Despite the data augmentation and regularization applied to AAA-UNet, we note that the model shows signs of overfitting as demonstrated by the training set performance being significantly higher than that of the test set. The model achieves mean 0.74 and median 0.79 DSCs for the training set while the same for the test set were 0.63 and 0.72, respectively. Furthermore, the predicted segmentations suffer from the effects of interpolation to the original frames of reference due to the lack of smooth boundaries along the lumen edges. While reducing model input dimensions alleviates compute constraints it comes at the expense of accuracy. Further, segmentation performance is highly variable though this can be attributed to the limited amount of data the model was trained upon. We observe that the surveillance cases have slightly higher variation in DSC performance as compared to the same for elective repair cases. Likewise, the AGH hospital subgroup has significantly higher variation in test set DSCs as compared to the same for NMH.

5.2 BB-AAA-UNet: Memory Efficient High-Resolution Segmentation with Prior Aneurysm Localization

BB-AAA-UNet still predicts with the high variance but is also less overfit. The model achieves DSCs of mean 0.80 and median 0.85 for the training set and the same for the test set were

0.70 and 0.85, respectively. As such our memory efficient computational pipeline outperforms the current approach to medical image segmentation described by AAA-UNet. This conclusion is further validated by the results of the Anderson Darling, Mann-Whitney U, and Wilcoxon rank-sum tests. As determined by Anderson Darling, the null hypothesis is rejected, indicating a non-normal distribution. Likewise, Mann-Whitney U and Wilcoxon rank-sum nonparametric tests reject the null hypothesis, indicating that test set DSCs for AAA-UNet and BB-AAA-UNet are significantly different.

A limitation to 3D CNNs as compared to their 2D counterparts is the increase in GPU memory owing to the substantially greater number of model training parameters [41]. Our novel computational approach enables 3D models to reap the rewards of high-resolution segmentation despite facing memory constrained development environments. Note, BB-AAA-UNet was trained in the same computational environment as AAA-UNet.

However, a shortcoming of the methods underpinning BB-AAA-UNet is that the bounding box itself can crop out AAA lumen regions. In consequence, the AAA lumen can have cuts, or the appearance of flattened regions incongruous with the anatomical AAA geometry. Comparing the hospital and treatment indication subgroups, we do not observe any significant differences from the results reported for AAA-UNet.

The focus of our study was accurate segmentation the AAA lumen while preserving memory efficiency; however, we can improve the applicability of our work by serving in lieu of conventional image processing pipelines. Semiautomatic segmentation of 2D CTs followed by reconstruction allowed researchers to compute geometric features of the ILT surface and volume ratios associated, enabling the investigators to observe that the ILT reduces peak wall stress [42]. To

accelerate discovery of AAA biomechanical properties, we could modify our model to perform multiclass and automatic 3D segmentation of the lumen and ILT.

5.3 BB-AAA-UNet: As Applied to Aneurysm Wall Segmentation

We find that while BB-AAA-UNet was successful for localizing the AAA, however, the high-resolution segmentations suffered from poor segmentation accuracy. By visual inspection, the model appears to be missing specific regions of the AAA wall, generally along the boundaries. Despite the model showing no signs of overfitting, segmentation performance across both the train and tests sets never outperformed DSC of 0.50 in terms of both the mean and median DSC. Notably, the symptomatic subgroup had significantly lower DSC performance as compared to the same for the asymptomatic subgroup. Provided the results, we presume that interpolation for the sake of efficient training removed a significant number of relevant labels and voxels corresponding to the AAA wall. Nevertheless, the AAA wall that was confined to the bounding box is accurately segmented by BB-AAA-UNet. Automatic 3D segmentation of the AAA wall obviates the need for assuming uniform wall thickness when performing stress analysis of biomechanical models [43].

5.4 Patch Segmentation UNet: Prediction of Aneurysm Wall by Medical Image Sub-volumes

While patch-based segmentation preserves memory efficiency, it can fail to capture the full global context of the medical image, or the full field of view [41]. However, we specify the training

parameters and loss function such that each sub-volume of size 128x128x64 voxels represents ¼ of the total image of size 256x256x64 voxels and handle the class imbalance by incentivizing the model to learn the image characteristics of the AAA wall. Furthermore, larger sub-volumes are more likely to incorporate the AAA wall itself. The disease itself, AAAs, manifests in a single location, as opposed to manifesting at multiple locations across the vasculature, so preserving global context, or full field of view, as much as possible was necessary. Though, selecting a larger sub-volume size reduces the overall number of sub-volumes sampled per image. We chose a random sub-volume sampler as it reduces overfitting.

As compared to our own memory efficient approach, we find that while the DSC improves for the asymptomatic cases, the symptomatic case performance remains the same. Patch-based segmentation outperforms our own memory efficient method for asymptomatic cases, but not for the symptomatic cases. For the test set of the asymptomatic subgroup, BB-AAA-UNet achieves DSCs of mean 0.37 and median 0.41 while the same for Patch Segmentation UNet is 0.42 and 0.51, respectively. For the test set of the symptomatic subgroup, BB-AAA-UNet achieves mean 0.32 and median 0.33 and Patch Segmentation UNet achieves 0.32 and 0.33 for the same, respectively. For asymptomatic cases, the null hypothesis of the Anderson Darling test is rejected for Patch Segmentation UNet but not for BB-AAA-UNet AAA wall DSCs. As such, we apply statistical tests for non-normal distributions. The following Mann-Whitney U and Wilcoxon rank-sum tests do not reject the null hypothesis. For symptomatic cases, the null hypothesis of the Anderson Darling test is not rejected for both Patch Segmentation UNet and BB-AAA-UNet, indicating normal distributions. Thereby, the t-test determines that the AAA wall DSCs are not statistically different. Nevertheless, as per the reported mean and medians, there remains increase

in Patch Segmentation UNet's DSC performance as compared to BB-AAA-UNet for AAA wall segmentation.

The symptomatic AAAs may require a greater field of view for more robust feature extraction or are more variable in geometry as compared to their asymptomatic counterparts. In the future, we hope to train at an even higher CTA resolution of 512x512x64 voxels to further bolster accuracy.

Conventional image processing pipelines for AAA or arterial wall segmentation are time consuming, inefficient, and prone to annotator variation for manual methods. Investigators found that calcification decreases AAA stability as indicated by an increase in peak wall stress after computing maximum stresses associated with AAA components. On the other hand, the same researchers noted that ILT reduced maximum AAA stress [44]. As part of the study which made the aforementioned discoveries, a semiautomated segmentation algorithm processed 2D CTs of 20 patients to annotate the lumen, ILT, wall, and calcifications. The authors report that 3D reconstruction demands 3 hours per AAA geometry [44]. It was unclear whether this includes segmentation and/or any post-processing steps like meshing and exporting. Patch Segmentation UNet requires 12.5 minutes to automatically and consistently segment the AAA wall of 123 patients (i.e., both the train and test sets) starting from a prepared model input to saving the 3D prediction in the original frame of reference. Note, Patch Segmentation UNet can be modified to predict multiple classes as in the above experiment and we perform inference in a low-cost development environment. Interpolation, thresholding, GPU memory, and input medical image dimensions are also factors which influence processing time. Our technology could augment downstream mechanics related computation regarding AAA anatomy in research or clinical settings.

5.5 AAA Image Transformers: Classifying Medical Images by Aneurysm Severity with Latent Representations

We find that class specific image transformers can approach and even beat the classification performance of conventional CNNs and class agnostic image transformers. Classifier performance is as follows: 1) the asymptomatic embedding classifier with test set AUC 0.75 and accuracy 76%, 2) the baseline image classifier with test set AUC 0.74 and accuracy 70%, 3) the symptomatic embedding classifier with test set AUC 0.72 and accuracy 68%, and 4) the class agnostic embedding classifier with test set AUC 0.66 and 66% accuracy.

Therefore, the latent space of less severe or asymptomatic AAAs is more meaningful than the same for more severe or symptomatic cases. As established by both the bounding box and patch segmentation methods struggling to identify AAA wall volumes of the symptomatic class, we can infer that the latent space of an asymptomatic image transformer is more finely tuned to recognize departures from the more normal anatomy as compared to the symptomatic image transformer's ability to distinguish from the more diseased anatomy. Given that both class specific image transformers were more successful than their class agnostic counterpart, it is implied that the model learns in a one v. all fashion (i.e., the model has greater success at establishing the latent space of a single class and recognizing departures from it).

We not only prove that segmentation models successfully encode, in the latent space, meaningful information beyond that found while analyzing medical images, but the same information is clinically relevant to informing disease severity. During hyperparameter tuning (i.e., selection of batch size, learning rate, epochs) of the classifier, the model was more likely to overfit when trained with embeddings, suggesting that better feature selection or regularization strategies should be adopted to further boost accuracy.

In a prior study, researchers used backwards stepwise linear regression to identify the most relevant local wall strength predictors (i.e., age, gender, AAA size, AAA normalized local diameter, ILT thickness, etc.). Following the identification of the significant features, the researchers built a linear mixed effects model for noninvasive prediction of AAA wall strength (i.e., spatial measures taken from CTs). While the influence of each input could be related to the model output, the authors cited that the statistical model could not predict wall strength outside of the original data range [45]. The purpose of the image transformers study is to relate the intuitive task of segmentation to the more complex understanding of classification. The inherent advantage of the deep learning models we develop is that they are more robust to variation and perform automatic feature extraction. Thus, the models are more finely tuned to recognizing image characteristics beyond pre-specified metrics, trading some interpretability in consequence. Nevertheless, our end goal has remained to bring explainability to deep learning.

5.6 AAA-ViT: Moving Towards Detection with Classification of Aneurysm Severity with Anatomical Explanation

Our novel 3D medical image ViT approaches ~70% accuracy in the test set, similar to the image classifier described in the previous section. However, unlike the previously mentioned image transformers, the ViT not only classifies CTAs by AAA severity, but also provides anatomical explanation for its prediction in the form of class specific heatmaps. We observe that when the model predicts the correct class, the predicted heatmap of the model predicted class carries a strong likelihood of incorporating AAA anatomy (i.e., AAA lumen, ILT, and wall). In the visualizations provided, we note that the correctly predicted asymptomatic case highlights a

region of the AAA lumen while correctly predicted symptomatic case highlights the ILT, an image volume associated with elevated rupture risk.

As such, the heatmaps of asymptomatic cases are more likely to include the lumen while the heatmaps for symptomatic cases are more likely to incorporate the same anatomy as well as surrounding features like the ILT, calcifications, and wall. However, the ViT is not AAA specific because it can focus on other aspects of the image volume including the abdominal tissue. Therefore, we hope to elevate the AP, a measure of AAA localization by heatmap, through further training and tuning of the model.

A significant limitation to ViTs is that they are described as often being data hungry as compared to their CNN counterparts, in part due to their lack of strong inductive bias (i.e., the locality – neighboring pixels are related and weight sharing – pattern searching provided by convolutions) [46, 47]. Though, more recent methods such as student-teacher strategies which employ distillation tokens have been shown to maintain accuracy while remaining data efficient during training as well as the introduction of convolutions to ViTs has shown promise in merging each architecture's advantages [48, 49]. Nevertheless, as de facto CNNs have used GRAD-CAM explain the anatomical basis for classification decisions in the past, we hope to accomplish the same with our novel 3D AAA-ViT leveraging LRP [50].

Moving forward, we plan on reapplying our novel 3D medical image ViT for classification of normal v. peripheral arterial disease (PAD) which presents a new challenge. PAD is an illness typified by stenoses of the arteries which manifest at infrequent locations across the vasculature, unlike AAAs which are generally confined to the descending aorta. Thus, providing prediction of patient disease status and anatomical justification of PAD through heatmap based localization remains a future endeavor. In short, we engineer a novel, interpretable, and state-of-the-art deep

learning technology which is applied for detection of AAAs through patient specific medical image classification and weakly supervised heatmap localization.

6.0 Conclusion

We present several key contributions to the scientific literature. The bounding box localization for high-resolution image segmentation inference pipeline we develop outperforms the conventional CNN for AAA lumen segmentation. Memory efficient methods underpinned on bounding boxes and patch-based segmentation show promise for AAA wall delineation. Image transformers with class specific segmentation encoders aid the identification of AAA severity, approaching and even overcoming the performance achieved by basic medical image classifiers. Finally, we present a novel 3D medical image ViT which provides interpretable explanations for its prediction of patient disease status. In sum, we engineer several deep learning methods for the detection, segmentation, and classification of the AAA while respecting the inherent frame of reference provided by the medical image.

Appendix A Study References

Appendix A includes links to the results that were included as part of the Master's Thesis.

1) AAA-UNet: Baseline Aneurysm Segmentation
    a. Single Segmentation Model
        i. https://drive.google.com/drive/folders/1AdwupMSqXUQCz4ffgJ4scGovioJiMwyW?usp=sharing
2) BB-AAA-UNet: Memory Efficient High-Resolution Segmentation with Prior Aneurysm Localization
    a. Bounding Box Localization Model
        i. https://drive.google.com/drive/folders/1Jn_FRGXgw3QcS5pxBnMMIrlseg2i8LA6?usp=sharing
    b. High Resolution Segmentation Model
        i. https://drive.google.com/drive/folders/19jRj_YDSJhhrOh9WG3dHx9sCeYeBSTJb?usp=sharing
3) BB-AAA-UNet: As Applied to Aneurysm Wall Segmentation
    a. Bounding Box Localization Model
        i. https://drive.google.com/drive/folders/1THI1pXtyvfelGQgTo7RLyuf7PIrRIRqT?usp=sharing
    b. High Resolution Segmentation Model
        i. https://drive.google.com/drive/folders/15FxzBCYQxxAvkuseS5R_ek5I5Y3XAASS?usp=sharing
4) Patch Segmentation UNet: Prediction of Aneurysm Wall by Medical Image Sub-volumes

a. Single Segmentation Model
    i. https://drive.google.com/drive/folders/1whV1lRzaOOJHJxJx79SmjqpIFNrfTlpK?usp=sharing

5) AAA Image Transformer: Classifying Medical Images by Aneurysm Severity with Latent Representations
    a. Class Agnostic Image Transformer & Image Classification
        i. https://drive.google.com/drive/folders/1yM9nIQi000KJZC4S8RQgkPuLkb_cPHMz?usp=sharing
    b. Asymptomatic Image Transformer & Classification
        i. https://drive.google.com/drive/folders/16cjLCj8bW39OC93RY8RZYqi1pZrr3goj?usp=sharing
    c. Symptomatic Image Transformer & Classification
        i. https://drive.google.com/drive/folders/1qInJeze8ro_1y9-MgF9rsnX5_Bkz9cbQ?usp=sharing
6) AAA-ViT: Moving Towards Detection with Classification of Aneurysm Severity
    a. Single Classification Model
        i. https://drive.google.com/drive/folders/1p6PF0LOhdWt2VHriv_tMR8jO-G8riJtU?usp=sharing

Appendix B Miscellaneous

Visualization of medical images was accomplished by ITK-SNAP for the Dataset and AAA-VIT sections found within Methods [27]. Visualization of medical images in all other figures was accomplished using ParaView [51].

Bibliography

[1] Centers for Disease Control and Prevention, National Center for Health Statistics. Multiple Cause of Death 1999-2019 on CDC WONDER Online Database, released in 2020. Data are from the Multiple Cause of Death Files, 1999-2019, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.

[2] Singh, M. Abdominal Aortic Aneurysm. Society for Vascular Surgery.

[3] Claridge, R., Arnold, S., Morrison, N., & van Rij, A. M. (2017). Measuring abdominal aortic diameters in routine abdominal computed tomography scans and implications for abdominal aortic aneurysm screening. Journal of vascular surgery, 65(6), 1637-1642.

[4] Castro-Ferreira, R., Vidoedo, J., Peixoto, J., Canedo, A., Teixeira, J., Leite-Moreira, A., & Sampaio, S. (2021). Incidental abdominal aortic aneurysms are largely undocumented and unmonitored. Annals of Vascular Surgery.

[5] Darling, R. C., Messina, C. R., Brewster, D. C., & Ottinger, L. W. (1977). Autopsy study of unoperated abdominal aortic aneurysms. The case for early resection. Circulation, 56(3 Suppl), II161-4.

[6] Hall, A. J., Busse, E. F., McCarville, D. J., & Burgess, J. J. (2000). Aortic wall tension as a predictive factor for abdominal aortic aneurysm rupture: improving the selection of patients for abdominal aortic aneurysm repair. Annals of vascular surgery, 14(2), 152-157.

[7] Haller, S. J., Crawford, J. D., Courchaine, K. M., Bohannan, C. J., Landry, G. J., Moneta, G. L., ... & Rugonyi, S. (2018). Intraluminal thrombus is associated with early rupture of abdominal aortic aneurysm. Journal of vascular surgery, 67(4), 1051-1058.

[8] Buijs, R. V., Willems, T. P., Tio, R. A., Boersma, H. H., Tielliu, I. F., Slart, R. H., & Zeebregts, C. J. (2013). Calcification as a risk factor for rupture of abdominal aortic aneurysm. European Journal of Vascular and Endovascular Surgery, 46(5), 542-548.

[9] Reeps, C., Maier, A., Pelisek, J., Härtl, F., Grabher-Meier, V., Wall, W. A., ... & Gee, M. W. (2013). Measuring and modeling patient-specific distributions of material properties in abdominal aortic aneurysm wall. Biomechanics and modeling in mechanobiology, 12(4), 717-733.

[10] Gasser, T. C. Biomechanical Rupture Risk Assessment: A Consistent and Objective Decision-Making Tool for Abdominal Aortic Aneurysm Patients. Aorta. 2016; 4 (2).

[11] Zhao, W. Endoleaks (Type I-V). Society for Vascular Surgery.

[12] Corriere, M. A., Feurer, I. D., Becker, S. Y., Dattilo, J. B., Passman, M. A., Guzman, R. J., & Naslund, T. C. (2004). Endoleak following endovascular abdominal aortic aneurysm repair: implications for duration of screening. Annals of surgery, 239(6), 800.

[13] Kim, H. O., Yim, N. Y., Kim, J. K., Kang, Y. J., & Lee, B. C. (2019). Endovascular aneurysm repair for abdominal aortic aneurysm: a comprehensive review. Korean journal of radiology, 20(8), 1247-1265.

[14] Di Martino, E. S., Bohra, A., Geest, J. P. V., Gupta, N., Makaroun, M. S., & Vorp, D. A. (2006). Biomechanical properties of ruptured versus electively repaired abdominal aortic aneurysm wall tissue. Journal of vascular surgery, 43(3), 570-576.

[15] Khosla, S., Morris, D. R., Moxon, J. V., Walker, P. J., Gasser, T. C., & Golledge, J. (2014). Meta-analysis of peak wall stress in ruptured, symptomatic and intact abdominal aortic aneurysms. Journal of British Surgery, 101(11), 1350-1357.

[16] Vorp, D. A., Lee, P. C., Wang, D. H., Makaroun, M. S., Nemoto, E. M., Ogawa, S., & Webster, M. W. (2001). Association of intraluminal thrombus in abdominal aortic aneurysm with local hypoxia and wall weakening. Journal of vascular surgery, 34(2), 291-299.

[17] Doyle, B. J., Callanan, A., Burke, P. E., Grace, P. A., Walsh, M. T., Vorp, D. A., & McGloughlin, T. M. (2009). Vessel asymmetry as an additional diagnostic tool in the assessment of abdominal aortic aneurysms. Journal of vascular surgery, 49(2), 443-454.

[18] Chung, T. K., Liang, N. L., & Vorp, D. A. (2022). Artificial intelligence framework to predict wall stress in abdominal aortic aneurysm. Applications in Engineering Science, 10, 100104.

[19] Fantazzini, A., Esposito, M., Finotello, A., Auricchio, F., Pane, B., Basso, C., ... & Conti, M. (2020). 3D Automatic Segmentation of Aortic Computed Tomography Angiography Combining Multi-View 2D Convolutional Neural Networks. Cardiovascular engineering and technology, 11(5), 576-586.

[20] Hesamian, M. H., Jia, W., He, X., & Kennedy, P. (2019). Deep learning techniques for medical image segmentation: achievements and challenges. Journal of digital imaging, 32(4), 582-596.

[21] Salvi, A., Finol, E., & Menon, P. G. (2021, November). Convolutional Neural Network based Segmentation of Abdominal Aortic Aneurysms. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 2629-2632). IEEE.

[22] Habijan, M., Galić, I., Leventić, H., Romić, K., & Babin, D. (2020, September). Abdominal Aortic Aneurysm Segmentation from CT Images using Modified 3D U-Net with Deep Supervision. In 2020 International Symposium ELMAR (pp. 123-128). IEEE.

[23] Siriapisith, T., Kusakunniran, W., & Haddawy, P. (2019). 3D segmentation of exterior wall surface of abdominal aortic aneurysm from CT images using variable neighborhood search. Computers in biology and medicine, 107, 73-85.

[24] Salvi, A., Finol, E., & Menon, P. G. (2022, April). Image transformers with regional attention for classification of aneurysm rupture risk without explicit segmentation. In Medical Imaging 2022: Biomedical Applications in Molecular, Structural, and Functional Imaging (Vol. 12036, pp. 278-286). SPIE.

[25] Reeps, C., Maier, A., Pelisek, J., Härtl, F., Grabher-Meier, V., Wall, W. A., ... & Gee, M. W. (2013). Measuring and modeling patient-specific distributions of material properties in abdominal aortic aneurysm wall. Biomechanics and modeling in mechanobiology, 12(4), 717-733.

[26] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[27] Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 2006 Jul 1;31(3):1116-28.

[28] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Springer, Cham.

[29] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

[30] Z. Yaniv, B. C. Lowekamp, H. J. Johnson and R. Beare, "SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research", J Digit Imaging., vol. 31, no. 3, pp. 290-303, 2018.

[31] Beare, R. (2011). Histogram-based thresholding—some missing methods. Insight J.

[32] Pérez-García, F., Sparks, R., & Ourselin, S. (2021). TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Computer Methods and Programs in Biomedicine, 208, 106236.

[33] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Jorge Cardoso, M. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In Deep learning in medical image analysis and multimodal learning for clinical decision support (pp. 240-248). Springer, Cham.

[34] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[35] Pawlowski, N., Ktena, S. I., Lee, M. C., Kainz, B., Rueckert, D., Glocker, B., & Rajchl, M. (2017). Dltk: State of the art reference implementations for deep learning on medical images. arXiv preprint arXiv:1711.06853.

[36] He, K., Zhang, X., Ren, S., & Sun, J. (2016, October). Identity mappings in deep residual networks. In European conference on computer vision (pp. 630-645). Springer, Cham.

[37] Chefer, H., Gur, S., & Wolf, L. (2021). Transformer interpretability beyond attention visualization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 782-791).

[38] Reedha, R., Dericquebourg, E., Canals, R., & Hafiane, A. (2021). Vision Transformers For Weeds and Crops Classification Of High Resolution UAV Images. arXiv preprint arXiv:2109.02716.

[39] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7), e0130140.

[40] Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. Pattern recognition, 65, 211-222.

[41] Wang, Y., Blackie, L., Miguel-Aliaga, I., & Bai, W. (2022). Memory-efficient Segmentation of High-resolution Volumetric MicroCT Images. arXiv preprint arXiv:2205.15941.

[42] Wang, D. H., Makaroun, M. S., Webster, M. W., & Vorp, D. A. (2002). Effect of intraluminal thrombus on wall stress in patient-specific models of abdominal aortic aneurysm. Journal of vascular surgery, 36(3), 598-604.

[43] Raghavan, M. L., Vorp, D. A., Federle, M. P., Makaroun, M. S., & Webster, M. W. (2000). Wall stress distribution on three-dimensionally reconstructed models of human abdominal aortic aneurysm. Journal of vascular surgery, 31(4), 760-769.

[44] Li, Z. Y., Jean, U., Tang, T. Y., Soh, E., See, T. C., & Gillard, J. H. (2008). Impact of calcification and intraluminal thrombus on the computed wall stresses of abdominal aortic aneurysm. Journal of vascular surgery, 47(5), 928-935.

[45] Vande Geest, J. P., Wang, D. H., Wisniewski, S. R., Makaroun, M. S., & Vorp, D. A. (2006). Towards a noninvasive method for determination of patient-specific wall strength distribution in abdominal aortic aneurysms. Annals of biomedical engineering, 34(7), 1098-1106

[46] Cao, Y. H., Yu, H., & Wu, J. (2022). Training vision transformers with only 2040 images. arXiv preprint arXiv:2201.10728.

[47] d'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., & Sagun, L. (2021, July). Convit: Improving vision transformers with soft convolutional inductive biases. In International Conference on Machine Learning (pp. 2286-2296). PMLR.

[48] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning (pp. 10347-10357). PMLR.

[49] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22-31).

[50] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[51] Ayachit, Utkarsh, The ParaView Guide: A Parallel Visualization Application, Kitware, 2015, ISBN 978-1930934306