Master's Thesis in Artificial Intelligence

Radboud University Nijmegen

Automatic muscle and fat segmentation in 3D abdominal CT images for body composition assessment

---

Radboud University Medical Center

Author:
Thijs van den Hout
S4597400
T.vandenHout@student.ru.nl

Co-authors:
Nikolas Lessmann
Postdoc researcher
Diagnostic Image Analysis Group
Radboudumc, Nijmegen

Alina Vrieling
Assistant professor
Health Evidence
Radboudumc, Nijmegen

Internal supervisor:
Nikolas Lessmann
Postdoc researcher
Radboudumc

Assessor:
Luca Ambrogioni
Assistant professor
Dept. of Artificial Intelligence
Radboud University

August 28, 2020

***Abstract*—Body composition is an informative biomarker in the treatment of cancer. In particular, low muscle mass has been associated with higher chemotherapy toxicity, shorter time to tumor progression, poorer surgical outcomes, impaired functional status, and shorter survival. However, because CT-based body composition assessment requires outlining the different tissues in the image, which is time-consuming, its practical value is currently limited. To form an estimate of body composition, different tissues are often segmented manually in a single 2D slice from the abdomen. For use in both routine care and in research studies, automatic segmentation of the different tissue types in the abdomen is desirable. This study focuses on the development and testing of an automatic approach to segment muscle and fat tissue in the entire abdomen. The four classes of interest are skeletal muscle (SM), inter-muscular adipose tissue (IMAT), visceral adipose tissue (VAT), and subcutaneous adipose tissue (SAT). A deep neural network is trained on two-dimensional CT slices at the level of the third lumbar vertebra. Three experiments were carried out with the goal of improving the network with information from other, unannotated data sources. Active learning methods were applied to sample additional data to annotate and include in the training of the model. The proposed algorithm combines two models to segment muscle and fat in the entire abdomen and achieves state-of-the-art results. Dice scores of 0.91, 0.84, 0.97, and 0.97 were attained for SM, IMAT, VAT, and SAT, respectively, averaged over five locations throughout the abdomen.**

1. Introduction

Body composition is defined as the relative amount of muscle and fat in the body and it is an important biomarker in the treatment of various diseases. In particular, low muscle mass is associated with higher chemotherapy toxicity in cancer treatment, which in turn is linked to decreased survival rate, and a higher risk of post-surgery complications [1], [2], [3]. A high fat percentage is also associated with higher risk of cardiovascular disease [4], and a lower skeletal muscle volume is linked to complications in major surgery [5], [6].

An assessment of body composition can reveal important information about the increased risks in numerous treatments. However, because of the difficulty in obtaining accurate body composition assessments, these efforts remain mostly unexplored. Body composition is assessed using many different techniques, including Body Mass Index (BMI), Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Dual-energy X-ray absorptiometry [7]. These methods differ greatly in accuracy, speed, and availability. In practice, simple but crude measures such as BMI are often used instead of more precise ones. CT imagery provides accurate images of the composition of the full body, but analysing these by hand is extremely time consuming. In practice, one axial slice of a CT scan at the height of the third lumbar vertebra (L3) is examined and manually segmented to obtain an approximate measure of body composition. Research has shown this single slice at the L3 vertebra provides a good estimate of the full body composition [8]. In Figure 1, an example of a CT scan is shown.

Although the segmentation of an axial slice at L3 gives an approximation of full body composition, being able to segment the complete abdomen in a 3D CT image will result in more accurate assessments and volumetric quantification instead of surface quantification. The manual segmentation of a single axial slice takes 15 minutes on average. Segmenting the entire abdomen manually is therefore infeasible. The current research focuses on the automation of body composition assessment by automatically segmenting 2D CT slices at L3, as well as the entire abdomen.

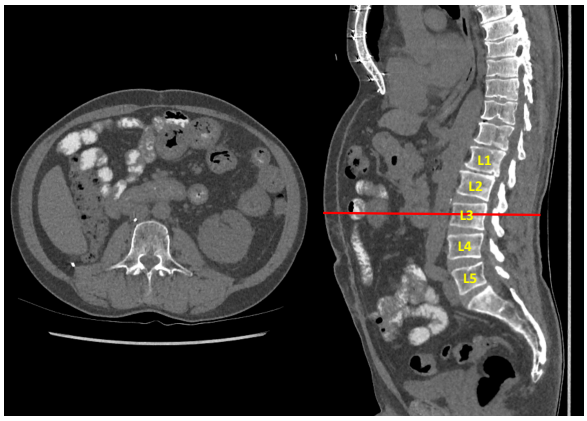

Figure 1: Example of a CT scan. On the right a sagittal slice through the center of the body. The five lumbar vertebrae are marked as L1 - L5. On the left an axial slice at the level of L3.

2. Related work

Previous research has focused mainly on the automatic segmentation of muscle and fat in a single axial slice at the location of L3. Popuri et al. introduce a FEM (Finite element method)-based approach to segment skeletal muscle from adipose tissue using template-based registration of the CT image and applying a statistical deformation model. They achieved an average Jaccard score of 91% [9]. Kullberg et al. applied thresholding of the images in combination with a lean tissue (muscle) filter, used to classify tissue as inside or outside the abdominal wall, to segment visceral and subcutaneous adipose tissue (VAT and SAT) with a Dice coefficient of 0.97 [10]. Dabiri et al. proposed a two-branch neural network in to segment skeletal muscle at L3 and achieved a Dice coefficient of 98% [11]. Although these studies report high accuracies, they are based on 2D CT images at L3 or the fourth thoracic vertebra (T4).

Little research has been done on the segmentation of 3D abdominal CT images for body composition assessment, largely because no annotated data is available for full abdominal scans. A few researchers have recently

ventured in this task, with promising first results. Hu et al. proposed a framework to segment skeletal muscle (SM), psoas muscle, VAT and SAT. In their study a PCA-based representation was used to estimate the atlas class and segment abdominal wall using Multi-Atlas Label Fusion (MALF). The abdominal wall is used in combination with pre-defined Hounsfield Unit (HU) ranges to segment SM, VAT, and SAT. They reported a Dice coefficient of 0.854, 0.740, 0.887 and 0.933 for skeletal muscle, psoas muscle, VAT, and SAT respectively [12]. In a 2019 study by Weston et al., a deep neural approach was employed to segment SAT, SM, VAT, organs, and bone in CT images. They report a Dice score of 0.93, 0.88, and 0.97 for SAT, SM and VAT, respectively, in unseen CT slices at L3 [13]. They applied their network to all abdominal slices between L1 and L5 and showed it generalized quite well to areas in the mid abdomen. No quantitative results for areas other than L3 were reported. In 2020, Grainer et al. described a method for segmenting VAT and SAT in 3D abdominal CT scans. They used a U-Net architecture to segment the two classes and reported a mean Dice score of 0.94 [14]. The ground truth labels used in the training of their network were created using thresholding of the CT image. SAT was selected as the area between the skin and the abdominal wall filtered with a threshold, while the area inside the abdominal wall was used to select VAT. Furthermore, they showed that BMI did not correlate well with VAT, reiterating the need for more advanced methods for body composition assessment.

3. Methods

The goal of this study is to obtain a novel method for assessing and quantifying the body composition of patients. Currently, automatic body composition measures are often too crude, and more accurate measures consume too much time or resources to obtain. This research will focus on the automatic segmentation of four classes in the abdomen: skeletal muscle (SM), Inter-muscular adipose tissue (IMAT), visceral adipose tissue (VAT) and sub-cutaneous adipose tissue (SAT). A system will be developed to segment these four classes in 2D axial slices at the level of L3, and it will be explored how this segmentation model can be generalized to segment the entire abdomen. The result of the algorithm is a segmentation mask of the abdomen and the quantification of each class in cm^3. The quantification of each class in a single slice at L3 is also given in cm^2. In contrast with previous research, this study will also segment IMAT, a challenging task since it only covers a small percentage of the scan. A larger quantity of IMAT is indicative for various pathophysiological conditions, such as type 2 diabetes, growth hormone-induced insulin resistance, and conditions characterized mostly by a decrease of muscle mass [15]. Furthermore, this research is one of the first to both quantitatively and qualitatively evaluate results throughout the abdomen.

This section will first describe the data that was available to develop the network. Then, the experiments that were carried out to realize the application will be explained. In short, these experiments comprise a base model trained on all the 2D annotated slices that were available, a model that included additional annotations from another data set in training, an experiment attempting a 3D approach, and experiments that apply active learning methods to increase the amount of annotated data.

3.1. Data

All data used in this project are images obtained with Computed Tomography (CT). These images are 3D volumes consisting of multiple 2D slices of 512 by 512 pixels. The value of pixels, or voxels in a volume, is the relative radiodensity on the Hounsfield scale. Values on this scale are often referred to as Hounsfield Units (HU). Water has an attenuation of 0 HU and air -1000 HU. The attenuation for skeletal muscle lies between -30 and 150 HU, for SAT and IMAT between -190 and -29 HU, and for VAT between -150 and -50 HU.

Three different data sets were used in the training of the segmentation network. Firstly, a data set containing 1287 2D axial CT slices at the location of the third lumbar vertebra (L3) with annotations for SM, IMAT, VAT, and SAT. The scans are acquired from patients with kidney cancer and originate from six different hospitals in the Netherlands. In Table 1, the quantification of each class in this dataset is shown. The skeletal muscle class comprises the muscles around the vertebrae, the muscles lining the abdominal wall, the psoas muscles, and muscles of the pelvis. SM occupies 9.9% of the CT scan at L3 on average. IMAT comprises the fatty tissue between muscles and on average covers 0.9% of the area of a CT slice at L3. VAT is the fat tissue in the abdominal cavity, between the organs and covers 11.3% of the axial slice on average. SAT is the fat tissue directly under the skin and occupies 12.3% of the slice at L3 on average. The remaining 65.5% is background, containing any tissue not covered by the four classes of interest, as well as air, fluids, bones, and contents of the digestive system. An example of an annotated CT scan at L3 can be seen in Figure 2.

Second, a data set containing 102 3D CT images of the abdomen and thorax with no annotated labels. These scans originate from the same studies as the previously mentioned data set. Patients in this data set are also

Data set	N	Background	SM	IMAT	VAT	SAT
L3 training set	1159	$907.8 \pm 270.1 cm^2$	$138.1 \pm 35.2 cm^2$	$12.3 \pm 9.8 cm^2$	$157.6 \pm 106.0 cm^2$	$171.1 \pm 91.0 cm^2$
L3 validation set	128	$931.7 \pm 250.1 cm^2$	$137.6 \pm 32.7 cm^2$	$12.0 \pm 9.3 cm^2$	$153.1 \pm 116.4 cm^2$	$169.0 \pm 84.5 cm^2$

TABLE 1: Mean quantification of each class in cm^2 in the dataset containing annotated axial CT slices at L3.

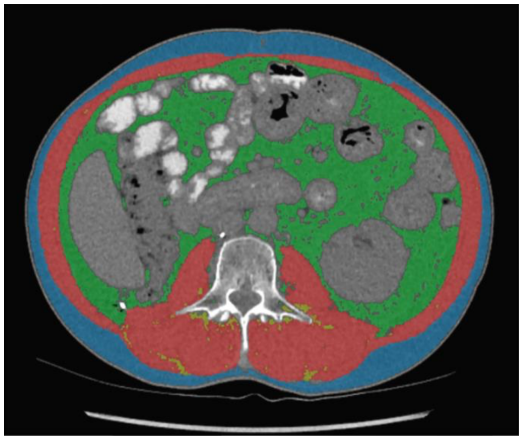

Figure 2: Example of an axial slice of a CT scan at the third lumbar vertebra (L3) with annotations. skeletal muscle (SM), inter-musclar adipose tissue (IMAT), visceral adipose tissue (VAT), subcutaneous adipose tissue (SAT)

present in the previous data set with an annotated L3 slice.

Lastly, the VISCERAL data set, a benchmarking data set for multi-organ segmentation containing 80 3D images, 40 of which are CT scans containing at least the thorax and abdomen. These images are segmented and annotated with 20 anatomical structures, most of which are abdominal organs [16]. The annotations do not include any of the classes in the focus of the current research. All images originate from the same hospital and were taken during clinical practice. The CT scans were acquired from patients with either confirmed bone marrow neoplasms or malignant lymphoma. Scans acquired from the latter were enhanced with an iodine-containing contrast agent. An example of a scan from the VISCERAL data set is shown in Figure 4.

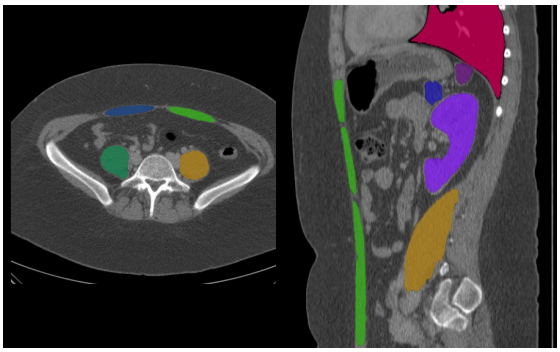

Figure 4: Example of a CT scan from the VISCERAL data set. On the right a sagittal slice on the left of the body, showing various organs, the left psoas muscle and the abdominal muscles. On the left an axial slice at L5, with annotations for the psoas muscles and abdominal muscles.

Test data. In order to quantitatively assess the performance of the proposed methods throughout the abdomen, some test data was prepared. Six CT images from the second dataset were selected with the criterion that their corresponding L3 slice was not present in the training set. From those six scans, the middle slice at the location of L1, L2, L4, and L5 was extracted and annotated. The annotations for the corresponding L3 slices existed in the validation set of the L3 data set. The annotations for the slices at L2 and L4 were created by a PhD student, and the annotations for the slices at L1 and L5 were created by the first author of this paper. The first author annotated the slices using 3D Slicer, an open source platform for medical image processing [17]. The annotation process was guided by an approved protocol for body composition analysis. Using thresholding of the image and a brush, annotations were drawn on top of the scan. The first author achieved an average Dice score of 0.933 on the other annotations in a test to validate aptitude, which was deemed adequate. This resulted in a test set of 6 annotated slices at L1, L2, L4, and L5, and 128 annotated test slices at L3. To maintain an equal number of test slices for each location in the abdomen, only the L3 slices from the six test scans will be used for testing.

3.2. Segmentation

The annotated data available for segmenting the aforementioned classes in the abdomen are axial slices at location of the third lumbar vertebra (L3). This restricts the methods that are viable to tackle this problem to two dimensional models. This research will focus on the application of deep neural networks in this domain; in particular, the U-Net model architecture will be used. U-Net is a fully convolutional network designed for biomedical image segmentation [18].

The U-Net adapted for this research comprises five down-sampling and up-sampling steps, where each step consists of two 3x3 padded convolution layers, each followed by the hyperbolic tangent activation function (tanh) and a batch normalization layer. Then, in the first half of the network, 2x2 max-pooling is applied to condense the image representation, while in the second half of the network, 2x2 up-sampling using nearest-neighbor interpolation is used to expand the representation to its original dimensions. Finally, a 1x1x5 convolutional kernel followed by a Softmax activation function is applied to obtain probabilities for each class per pixel. The final segmentation mask is obtained by selecting the most probable class for each pixel. In Figure 3, a schematic view of the network is shown. The models were trained for 1500 epochs with a batch size of 32 and a static learning rate of 0.0001, using the Adam optimizer. The model is optimized using a soft Dice loss function. The specific model architecture and hyper-parameters were experimentally discovered.

Before segmentation, the data is preprocessed by clipping the HU values outside the range [-400, 600] and subsequently normalizing the data between 0 and 1.

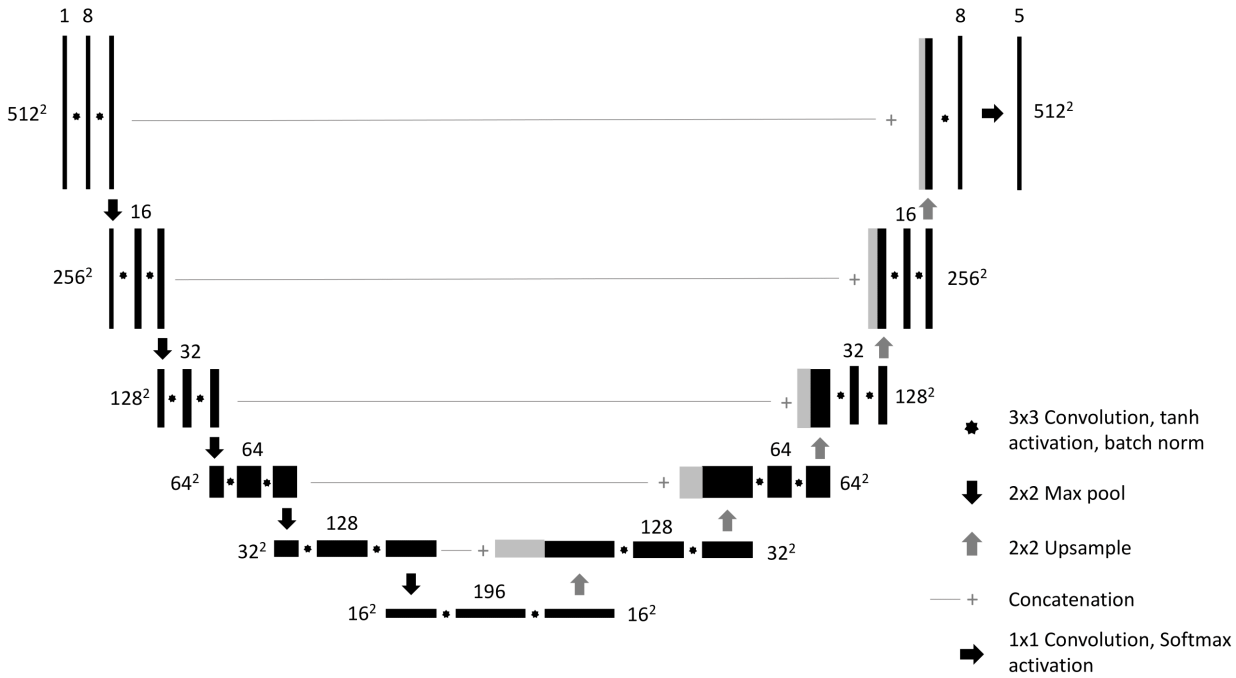

Figure 3: Schematic of the U-Net model architecture used in this study.

3.3. Experiments and intermediate results

To maximally utilize the data that was available, a number of experiments were carried out, encompassing various approaches to extract information from each data source. Here, these experiments will be described. Results of the experiments are briefly noted in this section if they are important for the following experiments. Often, the results of one experiment motivated the setup of the following, thus it is clarifying to present some summarized results here. The results of the experiments and their comparison are depicted in more detail in the results section.

3.3.1. L3 base model. The first experiment made use of only the first data set containing annotated axial slices at the location of L3. The data set was split in 90% training data and 10% validation data. The previously described deep neural network was trained for 2000 epochs to establish the L3 model, which forms the basis for the other experiments and is used as a baseline.

Table 2 shows Dice scores obtained by the L3 base model on the test slices. It can be seen that the two vertebrae farthest away from L3 have the worst performance, which is to be expected, as the model is only trained on slices at L3. Especially in L5, regular and obvious misclassifications occur because of the deviating structure of the scan. As can be seen in Figure 5, the psoas muscles are disconnected from the spine, and pelvic bones show. The algorithm fails to classify the psoas muscles as muscle, and wrongly classifies the marrow in the pelvis bones as muscle.

	SM	IMAT	VAT	SAT	Average
Dice score	0.8939	0.8325	0.9669	0.9686	0.9155

	L1	L2	L3	L4	L5
Dice score	0.8929	0.9263	0.9304	0.9341	0.8936

TABLE 2: Dice scores achieved by the L3 base model. The first row shows Dice scores per class, averaged over all locations in the abdomen. The second row shows Dice scores per location in the abdomen, averaged over all classes.

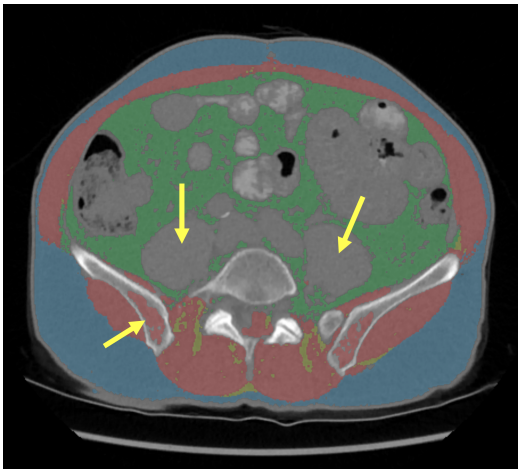

Figure 5: L5 slice segmented by the L3 base model. The arrows indicate areas of clear misclassification.

3.3.2. VISCERAL data set. Creating more annotated data by manually annotating scans in 3D is time-consuming. However, existing data sets unrelated to body composition assessment may provide annotations that are in some way useful to the task. To include more information about the structure of the abdomen farther away from L3, the VISCERAL dataset could be used to incorporate information about areas that certainly belong to the background class in the current research. The relevant classes that are segmented in the VISCERAL data set are the abdominal organs, psoas muscles, and abdominal muscles. The muscle groups are considered skeletal muscle in the context of this research, while the rest is considered as background.

From each of the 40 CT scans in the data set, one slice near L5 was chosen, since these slices were the most difficult to segment using the L3 base model. For these slices, pseudo-annotations were generated by the L3 model. On top of this, corrections were made by using the organ and muscle annotations in the data set, and the knowledge of the range of HU values each class could be. Each pixel that was segmented as organ is sure to be background in the current research, while the psoas muscles and abdominal muscles belong to the skeletal muscle class. Hounsfield values of the scan that were below -190 or above 150 also correspond to background. Finally, the pelvic bones were filled in manually as background, since the inside of these bones did not have a density high enough to be recognized as bone. This resulted in annotations which consisted partly of pixels automatically segmented by the L3 base model, and partly of corrections made on top of that using the VISCERAL annotations and HU values.

These scans were accompanied with a weight mask to lower the impact of the pseudo annotations on the loss function during training, since these annotations are not the ground truth but rather a decision made by a flawed model. The weight of the pixels segmented by the L3 base model (the pseudo annotations) was set to 0.6, whereas the pixels in the annotation which were corrected maintained a weight of 1. These weights were applied during training in the Dice loss function by multiplying the gradient of each pixel with its corresponding weight. The network was trained on the L3 training slices in combination with the enhanced slices from the VISCERAL data set.

	SM	IMAT	VAT	SAT	Average
Dice score	0.9054	0.8261	0.9654	0.9678	0.9162
	L1	L2	L3	L4	L5
Dice score	0.8919	0.9250	0.9268	0.9290	0.9080

TABLE 3: Dice scores achieved by the Visceral model. The first row shows Dice scores per class, averaged over all locations in the abdomen. The second row shows Dice scores per location in the abdomen, averaged over all classes.

When looking at the results in Table 3, slight improvements are made over the L3 base model at L5, while the performance in other locations drops slightly.

3.3.3. 3D approach. 102 scans in the L3 data set have a corresponding 3D CT scan. To make use of the information of the slices surrounding the annotated L3 slice this experiment was carried out. For these 102 scans, the three slices above and below the annotated L3 slice were extracted from their 3D counterpart. For the L3 slices that did not have a corresponding 3D scan, the same L3 slice was duplicated to obtain the same shape as the others. This resulted in *slabs* of CT images of 7 x 512 x 512.

The neural network was adapted such that the first layer of the U-Net consisted of 3D convolutions. Three consecutive blocks of 3D convolution, hyperbolic tangent activation, and 3D batch normalization reduce the image to 1 x 512 x 512, after which the rest of the network can be applied as normal.

	SM	IMAT	VAT	SAT	Average
Dice score	0.8530	0.6449	0.8788	0.9369	0.8284
	L1	L2	L3	L4	L5
Dice score	0.7499	0.8068	0.9394	0.8267	0.8192

TABLE 4: Dice scores achieved by the 3D approach. The first row shows Dice scores per class, averaged over all locations in the abdomen. The second row shows Dice scores per location in the abdomen, averaged over all classes.

The results in Table 4 show a decrease in performance in almost all respects. Slight improvements over the base model are made at L3, but at a great cost in effectiveness at other locations.

3.4. Active learning

The problem with many machine learning applications in the medical domain is the lack of annotated data. In this project too, the amount of annotated data was limited. The intention of this study is to create a model that is capable of segmenting the entire abdomen, while only annotated slices at L3 are available. To increase the amount of annotated data while avoiding unnecessary work, a method of training called active learning can be applied. In an active learning approach, data points are selected from a large pool of unannotated data that would be the most informative to the task when annotated. The selected data is then annotated by a so-called ”oracle”, the person or other information source that can be queried for additional annotations, after which the model is retrained with the increased training set. Using this method, the network can be improved more efficiently because less annotated data is required. Usually, active learning is an iterative process: determining the most informative samples, annotating these, and re-training the network.

3.4.1. Sample selection. The pool of unannotated data from which samples were selected for annotation was created by dissecting each of the 102 3D CT scans into

separate axial slices and selecting the abdominal slices that did not include the L3 vertebra. This resulted in a set of 2750 unannotated abdominal CT slices from L1 to L5, excluding L3.

Many sampling methods have been proposed in previous research, with a clear distinction between two approaches: uncertainty sampling and representation sampling. Uncertainty sampling encapsulates the sampling methods that make use of a measure of uncertainty of the model about a particular sample. This uncertainty measure is often the variance between the outputs of multiple runs of the same model after introducing a factor of randomness to the model. An example of uncertainty sampling is the Monte-Carlo Dropout method (MC dropout), where a dropout layer is activated during inference to introduce randomness in outputs. A dropout layer randomly removes a fixed percentage of nodes in the network at its location, essentially resulting in a part of the network being "turned off". The higher the variance between the outputs of the same sample, the more uncertain the model is about that sample, and the more informative it is deemed to be.

Representation sampling methods compare a representation of the candidate sample with the representation of the annotated data and compare the two to determine the difference (or distance). A sample whose representation is much different from the representation of the annotated data is assumed to be more informative to the model than a similar one. Such a representation can be a parameterized function, for example the latent space of a variational auto-encoder. If a candidate sample is far away in the latent space from the annotated data points, it is more useful to the model than a sample that is close to the annotated data in the same latent space.

The two methods, uncertainty and representation sampling, are combined in this study to form a sample selection strategy similar to one defined as MedAL in previous research by Smailagic et al. [19]. Firstly, Monte-Carlo Dropout is applied to select M candidate slices based on model uncertainty. All unannotated slices were segmented 25 times using the baseline model trained on all annotated slices with an 80% dropout layer active at the most abstract level of the network. The variance between the 25 segmentation maps was computed for each sample and the 100 most uncertain samples were selected. Afterwards, the latent representations from the U-Net model were extracted for these M candidates and compared to the latent variables of the annotated data.

$$d(c) = \frac{1}{N}\sum_{i=0}^{N} cosine\_distance(f(c), f(a_i))$$

Equation 1: Distance measure between the latent representation of a candidate slice c and the annotated data ($a \in$ annotated data). $f(x)$ extracts the latent variables representing sample x from the U-Net model.

The slices that are the furthest away in latent space to the pool of annotated data are considered the most different and thus the most informative. Equation 1 shows the distance measure for a candidate slice and the annotated pool. It computes the average cosine distance between the feature vector of the candidate slice and all feature vectors of the annotated slices. The cosine distance measure was used because it obtained good predictive entropy scores between the samples. Instead of computing $d(c)$ for all slices, first the 100 slices about which the model was most uncertain as determined by the Monte-Carlo Dropout were selected, and for this subset $d(c)$ was computed. The 50 slices that maximize $d(c)$ were selected for annotation and were annotated by the first author of this study.

The additionally annotated data was included in training with a ratio of 2:30; two additional annotated slices and 30 L3 slices in each batch. This constraint was put in place to evenly distribute the underrepresented data set of additional annotated slices in the training process.

3.4.2. Validation. To validate the MedAL sample selection strategy in the current domain, a different, more naive strategy was also applied to establish a baseline.

Slices with the highest mean squared error to the annotated slices were selected for annotation in the more naive control strategy. This measure was computed by taking the mean of the mean squared errors of a candidate slice to every annotated slice. Using both sampling strategies, 50 annotated slices were selected, annotated, and added to the annotated data set, on which the models were retrained.

For both the MedAL and control sampling methods, a maximum of two slices were selected from the same patient and those slices were at least 5 slices apart. This constraint was applied to prevent selecting slices from only a few patients that might be outliers in the data set.

The results of these two approaches are compared in the results section. Naturally, deep neural networks generalize better with more annotated data, so another model was trained with the union of both additional annotated data sets. Four slices were selected by both sampling strategies, resulting in a set of 96 unique additional annotations. The model was trained using a ratio of 4:30 additionally annotated slices to L3 slices per batch. In comparison with the models resulting from the other experiments, the active learning model will refer to the model trained with all additional annotated data. The results of the active learning model trained with all additional annotated data are concisely shown in Table 5, because of their relevance for the following section.

	SM	IMAT	VAT	SAT	Average
Dice score	0.9038	0.7864	0.9589	0.9621	0.9028

	L1	L2	L3	L4	L5
Dice score	0.8728	0.9001	0.9158	0.9099	0.9154

TABLE 5: Dice scores achieved by the Active Learning model, trained with the additional annotations from both sampling strategies.

The active learning model improved the performance at the location of L5, while performance at the other locations declines.

Separately, another experiment to validate the MedAL sampling strategy was carried out using the annotated slices at L3. Since this large data set was already annotated, both previously mentioned sampling strategies, as well as random sampling, could be applied in a more traditional, iterative setting. A random 10% split of the training data was used to train a baseline model. Afterwards, all three sampling strategies were applied to select an additional 10% of data from the remaining 90% of annotated data. With this increased training set, the models were retrained. This was repeated until all 100% of the training data was used. The results of this experiment are shown in section 5.1.3 since they are of no immediate importance to the following.

3.5. Twin networks

The results from the active learning experiment show that the active learning network outperforms the L3 base model at L5, but does not improve the performance elsewhere in the abdomen. To combine the competence of both the active learning model and the L3 base model, a twin network setup was employed for segmenting 3D abdominal scans.

First, a vertebra segmentation algorithm is applied to the scan. This algorithm was developed by the Diagnostic Image Analysis Group (DIAG) at Radboud University Medical Center, and returns a segmentation map of the vertebrae in the 3D CT scan [20]. Using this segmentation, the abdomen is extracted by selecting all axial slices that contain any of the vertebrae L1 through L5. Second, the axial slices containing the L5 vertebra are segmented using the active learning model, while the other abdominal slices are segmented using the L3 base model. Finally, the two segmentations are concatenated to form the final segmentation map of muscle and fat in the abdomen.

Table 6 shows the Dice scores for the twin networks approach.

	SM	IMAT	VAT	SAT	Average
Dice score	0.909	0.835	0.966	0.969	0.920
	L1	L2	L3	L4	L5
Dice score	0.893	0.926	0.930	0.934	0.915

TABLE 6: Dice scores achieved by the Twin networks approach. The first row shows Dice scores per class, averaged over all locations in the abdomen. The second row shows Dice scores per location in the abdomen, averaged over all classes.

4. Application

To enable an easy interface with the algorithm, the complete algorithm pipeline is hosted on grand-challenge.org, a platform for end-to-end development of machine learning solutions in biomedical imaging[1]. The algorithm is wrapped in a docker container, which is ran on the grand-challenge computing cluster. Users can upload CT scans through a web-interface and will be returned the quantification of each of the classes for the entire abdomen in cm^3 and for a slice at the location of L3 in cm^2, and a segmentation map of the abdomen. The latter can be viewed in the browser through CIRRUS, a workstation platform that allows scrolling through 3D scans, applying mask overlays, among other functionality. An example of the output of the segmentation model viewed in the CIRRUS workstation is shown in appendix 1. A flow chart of the algorithm pipeline is shown in Figure 6.

5. Results

In the experiments section, results of the different experiments were briefly presented. In this section, the results will be depicted and compared in more detail.

1. https://grand-challenge.org/

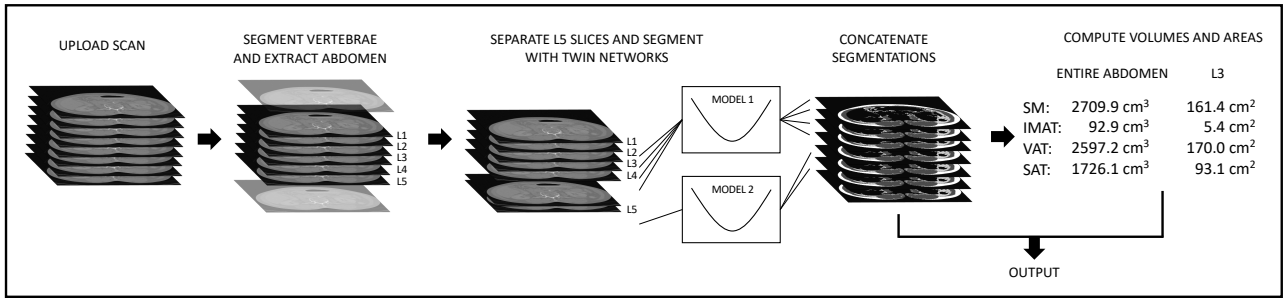

Figure 6: A flow chart of the application pipeline. First, the user uploads a scan to the website. A vertebra segmentation algorithm is applied to extract the abdominal axial slices. The slices from the lower abdomen (L5) are separated and the scan is segmented using the twin networks. The segmentations are concatenated and the volumes of each class in the entire abdomen, as well as the area of each class at the location of L3 are computed. The segmentation map is overlaid on the scan in the output.

5.1. Quantitative results

The accuracy of the algorithm can be assessed quantitatively in two ways: the accuracy of the segmentation maps and the accuracy of the resulting quantification of areas.

5.1.1. Segmentation maps. The accuracy of the segmentation maps will be expressed using the Dice coefficient of similarity between two areas, which is defined as $2|A \cap B|/|A+B|$. In Figure 7 and 8, box plots depicting the Dice scores for each of the experiment models are shown. Note that the box plot for the twin networks is not shown in Figure 8 since it is equivalent to the performance of the L3 base model for L1 - L4, and the active learning model at L5. From Figure 7 it can be observed that the IMAT class is significantly more difficult to segment than the other classes, due to its small area in the scan. Small errors quickly lead to a large decrease in Dice score. On the contrary, VAT and SAT are segmented with relative ease and confidence, obtaining high Dice scores.

When comparing the models, it is apparent that the 3D approach is inferior to the other models. At the location of L3, the network performs well, which is predictable since it is trained on L3 slices with surrounding information. Most notable is the large variance in performance by this model, which could be indicative of overall uncertainty.

The VISCERAL model is rather indiscernible from the L3 base model in performance. It is slightly better at segmenting skeletal muscle, however, at the location where improvement was expected, namely L5, it does not advance the baseline.

The active learning model improves the baseline model in a few aspects. Firstly, the segmentation of skeletal muscle is more accurate. This is largely because

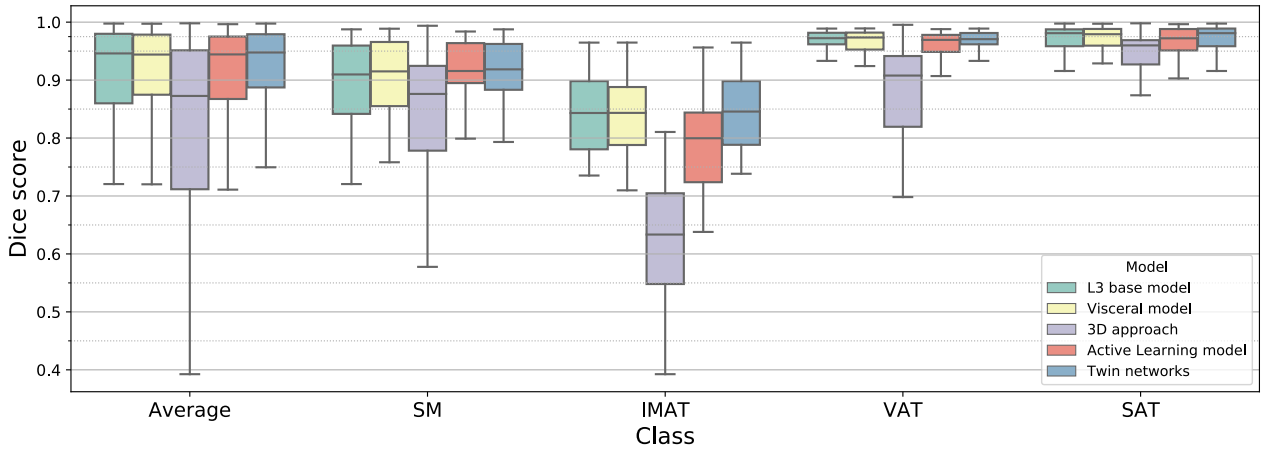

Figure 7: Box plots showing the Dice score (averaged over all locations in the abdomen) for each class and each model.

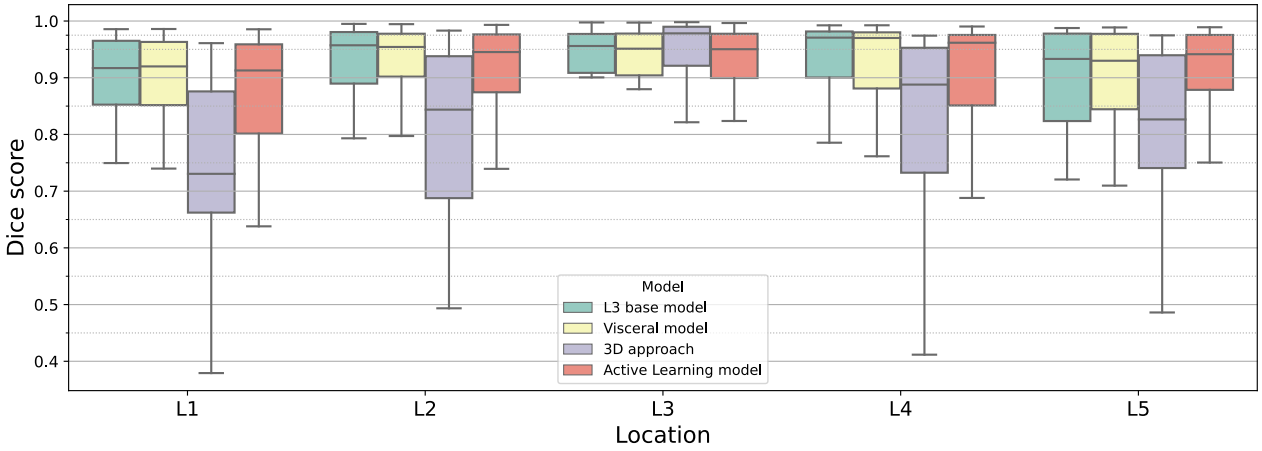

Figure 8: Box plots showing the Dice score (averaged over all classes) in each location in the abdomen, for each model.

the segmentation of skeletal muscle by the L3 base model at L5 is rather poor. At L5, the psoas muscles are detached from the spine and pelvic bones surrounded by muscle are visible. This differs greatly from the contents of a slice at L3. Since the selected slices in the active learning sampling strategy were mostly those around L5, the active learning model has adapted to the features of such slices. This effect is shown visually in the qualitative results section. The adaption to L5 slices lead to a slightly worse performance in the other areas of the abdomen when compared to the baseline model, hence the twin networks setup was devised.

The twin networks approach performs the best overall. Therefore, the following results regarding the quantification of each class will refer to the segmentations created by this model. More detailed box plots showing the Dice score for each class at each location for the L3 base model, active learning model, and twin networks approach can be seen in appendix 2.

5.1.2. Volume quantification. The quantification of the volume of each class is of great importance to the application of this research. The volume of each class is computed by counting the number of pixels in that class and multiplying it with the size of each voxel and is reported as cm^3.

The evaluation of the quantification of each class in this research is limited to the surface areas from the test set. No three-dimensional annotations exist to compare our quantifications with. The surface area of each class is computed by counting the number of pixels in that class and multiplying it with the size of the pixel, resulting in a reported surface area for each class in cm^2.

In Figure 9, the predicted area for each class is plotted against the ground truth area. The computed R^2 values of each correlation plot reveal similar results as were observed from the Dice score box plots in figures 7 and 8. SM and IMAT are more difficult to segment than VAT and SAT.

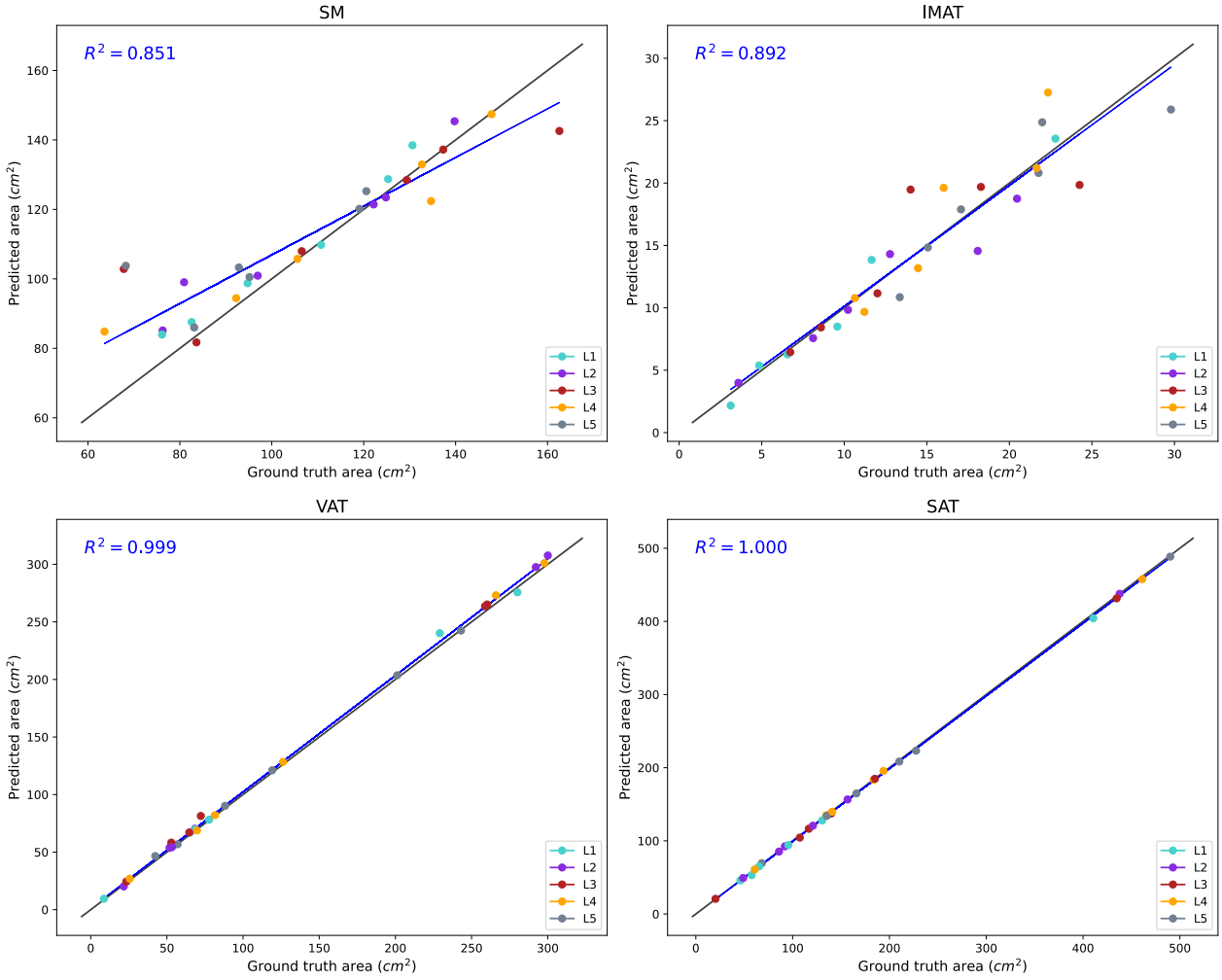

Figure 9: Correlation plots for the predicted area in cm^2 and the ground truth area in cm^2 for each class. The predictions are those of the twin networks approach. The 30 data points are the test slices, 6 for each location in the abdomen (L1 - L5). The black line is x=y, the blue line is a line fitted to the data.

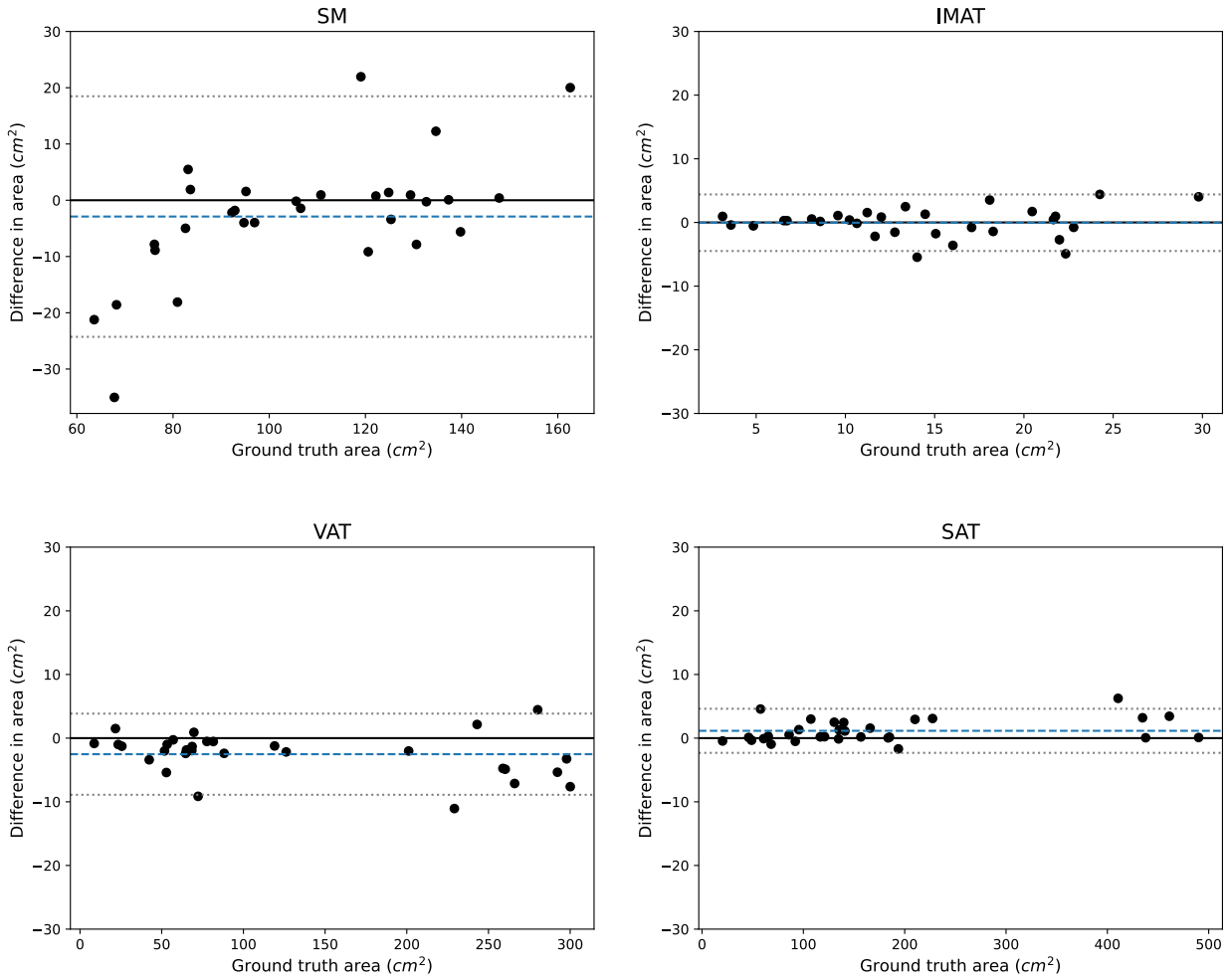

Figure 10: Bland-Altman plots showing the difference in area for each class between the predicted quantification and the ground truth in cm^2. The predictions are those of the twin networks approach. The 30 data points are the test slices, 6 for each location in the abdomen (L1 - L5). The mean difference is indicated by the dashed blue line. The 95% confidence intervals are indicated by the gray dotted lines.

The algorithm does not seem to regularly either overestimate or underestimate the area of any class. In Figure 10, Bland-Altman plots show more clearly the difference in area between the algorithm's prediction and the ground truth. The mean absolute difference between the predictions and ground truth is 8.83% for SM, 11.50% for IMAT, 3.47% for VAT, and 1.08% for SAT.

5.1.3. Active learning. In this section, the results of the active learning experiments will be described. In Table 7 and 8, Dice scores are shown for the two models trained with the additional annotated data. Both sampling strategies, MedAL and a naive mean squared error approach, seem to be equally effective. Both networks increase performance in L5 over the L3 base model, but lower performance elsewhere in the abdomen.

	SM	IMAT	VAT	SAT	Average
Dice score	0.8977	0.7881	0.9584	0.9651	0.9023

	L1	L2	L3	L4	L5
Dice score	0.8733	0.9032	0.9200	0.9128	0.9023

TABLE 7: Dice scores achieved by the network trained with the additional annotated data from the MedAL sampling strategy. The first row shows Dice scores per class, averaged over all locations in the abdomen. The second row shows Dice scores per location in the abdomen, averaged over all classes.

The second active learning experiment, using the annotated slices from the L3 dataset, compared the two sampling strategies with random sample selection in an iterative setting. The results are shown in Table 9. It seems the three sampling strategies are equally effective

	SM	IMAT	VAT	SAT	Average
Dice score	0.9004	0.7870	0.9566	0.9634	0.9019

	L1	L2	L3	L4	L5
Dice score	0.8736	0.9058	0.9144	0.9120	0.9035

TABLE 8: Dice scores achieved by the Active Learning model using the MSE sample selection strategy.

on this data set. One explanation for this effect is that the L3 slices are all very similar to each other and the difference between them is not significant enough to make a difference in the effectiveness as training data. Another explanation might be that any 10% of the training data (∼106 slices) encompasses the breadth of variation between images, making all 10% split effectively equivalent in information richness.

5.2. Qualitative results

In this section, the performance of the algorithm will be evaluated qualitatively, by visually inspecting the resulting segmentation maps and comparing them to the ground truth. This is useful for the evaluation of the proposed algorithm because too little reference data is available to thoroughly evaluate it solely quantitatively.

In Figure 11, an example segmentation of a 3D CT scan by the proposed algorithm is shown. Images A-C

% of training data	Mean Dice, MedAL	Mean Dice, MSE	Mean Dice, random
10% random selection, 500 epochs	0.838	0.838	0.838
20%, 500 epochs	0.874	**0.878**	0.873
30%, 500 epochs	0.883	**0.884**	0.881
40%, 1000 epochs	0.907	0.907	0.907
50%, 1000 epochs	0.922	0.922	0.922
60%, 1000 epochs	**0.921**	0.920	0.919
70%, 1500 epochs	**0.926**	0.924	0.924
80%, 1500 epochs	**0.928**	0.925	0.927
90%, 1500 epochs	0.928	0.928	**0.929**

TABLE 9: Active learning experiment on L3 data set. Comparing MedAL sample selection strategy with largest MSE to the annotated pool and random selection.

show segmented axial slices at L1, L3, and L5, respectively. Here, differences between the various locations in the abdomen are clearly visible. In L1, the liver, spleen, and often the bottom of the lungs are visible. The psoas muscles at this location are just slivers of muscle on both sides of the vertebra. Here, errors are made in the ribs, where the marrow is sometimes classified as muscle. The boundary between the muscles along the abdominal wall and directly adjacent organs is also difficult to discern, since they often have a similar density.

In image A, there are some uncertainties in the boundary between the liver and abdominal wall on the left of the image.

In image B (L3), the segmentation is naturally the

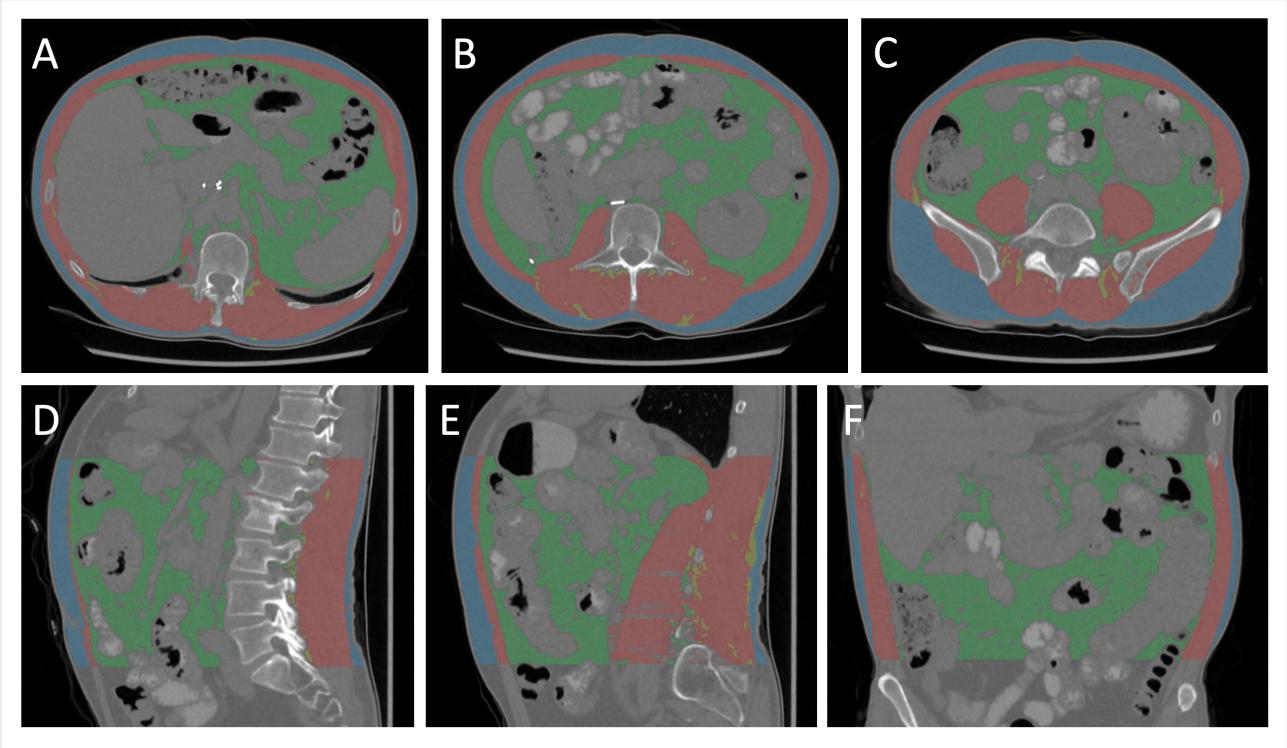

Figure 11: Representative example of abdominal segmentation of muscle and fat. **A-C:** axial slices at L1, L3, and L5. **D, E:** sagittal slices through the middle and to the left of the body. **F:** coronal slice through the center of the abdomen.

most accurate. The bowels are more prominently visible, as well as the kidneys. The psoas muscles around the vertebra are larger than they are higher up in the abdomen.

Image C shows a slice at L5. We can see the performance using the twin networks has improved over the L3 base model when comparing this segmentation to one created by the L3 base model in Figure 5. The psoas muscles are now correctly classified as muscle, and the pelvic bones are largely correctly classified as background. Still, the model incorrectly classifies some sporadic remnants of muscle inside of the pelvic bones.

When looking at the slices on the sagittal plane in images D and E, some artefacts of using a 2D approach can be observed. Distinct layers are visible, especially in the psoas muscles, where the model is uncertain of the class and correctly classifies it in one slice but not in the next. This is the result of each axial slice being treated as independent from each other, while looking at it from this perspective it is clear that they form connected groups of tissue.

6. Discussion

This study introduces a deep neural network which accurately segments skeletal muscle, inter-muscular adipose tissue, visceral adipose tissue, and subcutaneous adipose tissue in the abdomen. The goal of this research was to achieve a new state-of-the-art method for body composition assessment using routine CT scans of the abdomen. The proposed method gives a more accurate assessment of body composition than previous research with the same objective. The algorithm segments a slice at the location of L3, as well as the entire abdomen, resulting in a segmentation map of the abdomen and the quantification of the volume each compartment.

The algorithm achieves Dice scores of 0.91, 0.84, 0.97 and 0.97 for SM, IMAT, VAT, and SAT respectively, with an average of 0.92. The IMAT class was particularly difficult to segment due to is small overall volume. Related work has segmented other compartments; for example the 2019 study by Weston et al. in which they segment SAT, SM, viscera, and bone. They achieve Dice scores of 0.93 and 0.88 for SAT and SM respectively [13]. Hu et al. achieve Dice scores of 0.85, 0.89 and 0.93 for SM, VAT, and SAT respectively on L3 slices in their 2018 study [12].

The quantification of the volume of each class is an important statistic and can be derived from the segmentation of a scan. The mean percentage differences between the predicted quantification and ground truth are 8.83%, 11.5%, 3.47%, and 1.08% for SM, IMAT, VAT, and SAT respectively. Related work by Grainer et al. in 2020 segment VAT and SAT in 3D abdominal scans and report an average volumetric difference of 6.2% and 3.0% for VAT and SAT respectively [14].

This study is unique in the fact that it segments SM, IMAT, VAT, and SAT in the entire abdomen. These four compartments form important biomarkers in the treatment of cancer and for the prediction of cardiovascular disease and complications in major surgery. By segmenting the entire abdomen, a more accurate assessment of body composition can be made than by segmenting a single slice at L3, which is the current standard. Moreover, the automatic segmentation of a 3D CT scan is much faster than the manual segmentation of even a single slice.

The proposed algorithm is hosted as an application on grand-challenge.org and is easily accessible through a web-interface. This allows for easy integration with follow-up research, and eventually for use in the clinic.

A number of things can be learned from the results of the various experiments that were carried out. Firstly, a model trained on only annotated slices at the location of L3 can form a solid base for the segmentation of the entire abdomen.

The baseline model was difficult to improve upon with little or sparsely annotated data, however, experiments with the VISCERAL data set show that it is possible. Perhaps using more pseudo-annotated slices, corrected by other information sources, can form a viable alternative to increasing the pool of annotated data in the current domain, which is labor intensive and time consuming.

The experiment employing a 3D-to-2D network proved to worsen the performance. At L3, Dice scores were high, but overall the model performance declined. This could be explained by the fact that only a small percentage of training data ($\sim$6.5%) had information about the surrounding slices, which may not have been enough to warrant this approach. If more annotated data with surrounding slices is available, this might be a concept to revisit.

Results from the active learning experiments point out that the sampling selection strategies were not as successful as anticipated. Especially the comparison between the MedAL-like and naive sampling strategies was underwhelming. One explanation for this could be that the first step of the MedAL strategy, the uncertainty sampling, eliminated too many candidates, damaging the balance between uncertainty and representation sampling. Another explanation is that the method might not lend itself well to the current domain, and different strategies should be explored. For example using an image atlas, a type of template, with registration methods to determine how different a slice is from the annotated data set, or using more informative latent variables for representation sampling. Nonetheless, the additional annotated data improved performance at L5, leading to the proposed twin networks approach.

Limitations of this study include the fact that the algorithm is not tested on data from other hospitals than those in the training set. Different CT hardware and settings such as resolution and slice thickness can affect the performance of the algorithm. Furthermore, the fact that no 3D annotated data exists limits the network architecture to 2D approaches. Given the flaws discussed in the qualitative results section, the objective would clearly benefit from a 3D approach. Finally, due to the lack of annotated data from locations throughout the abdomen, the algorithm remains most effective in the mid-abdomen.

More annotated data in the lower and upper abdomen would increase the overall performance of the algorithm.

Future research could focus on developing more advanced post-processing of 3D segmentation maps to tackle the layer gaps artefact resulting from a slice-wise segmentation model. Additionally, a more effective way to select informative training samples in the current domain could be researched.

References

[1] W. P. Dijksterhuis, M. J. Pruijt, S. O. van der Woude, R. Klaassen, S. A. Kurk, M. G. van Oijen, and H. W. van Laarhoven, "Association between body composition, survival, and toxicity in advanced esophagogastric cancer patients receiving palliative chemotherapy," *Journal of cachexia, sarcopenia and muscle*, vol. 10, no. 1, pp. 199–206, 2019.

[2] M. C. Gonzalez, C. A. Pastore, S. P. Orlandi, and S. B. Heymsfield, "Obesity paradox in cancer: new insights provided by body composition," *The American journal of clinical nutrition*, vol. 99, no. 5, pp. 999–1005, 2014.

[3] G. Malietzis, A. Currie, T. Athanasiou, N. Johns, N. Anyamene, R. Glynne-Jones, R. Kennedy, K. Fearon, and J. Jenkins, "Influence of body composition profile on outcomes following colorectal cancer surgery," *British Journal of Surgery*, vol. 103, no. 5, pp. 572–580, 2016.

[4] L. F. Van Gaal, I. L. Mertens, and E. Christophe, "Mechanisms linking obesity with cardiovascular disease," *Nature*, vol. 444, no. 7121, pp. 875–880, 2006.

[5] M. Sugimoto, M. B. Farnell, D. M. Nagorney, M. L. Kendrick, M. J. Truty, R. L. Smoot, S. T. Chari, M. R. Moynagh, G. M. Petersen, R. E. Carter *et al.*, "Decreased skeletal muscle volume is a predictive factor for poorer survival in patients undergoing surgical resection for pancreatic ductal adenocarcinoma," *Journal of Gastrointestinal Surgery*, vol. 22, no. 5, pp. 831–839, 2018.

[6] K. H. Sheetz, S. A. Waits, M. N. Terjimanian, J. Sullivan, D. A. Campbell, S. C. Wang, and M. J. Englesbe, "Cost of major surgery in the sarcopenic patient," *Journal of the American College of Surgeons*, vol. 217, no. 5, pp. 813–818, 2013.

[7] A. Andreoli, F. Garaci, F. P. Cafarelli, and G. Guglielmi, "Body composition in clinical practice," *European journal of radiology*, vol. 85, no. 8, pp. 1461–1468, 2016.

[8] M. Mourtzakis, C. M. Prado, J. R. Lieffers, T. Reiman, L. J. McCargar, and V. E. Baracos, "A practical and precise approach to quantification of body composition in cancer patients using computed tomography images acquired during routine care," *Applied Physiology, Nutrition, and Metabolism*, vol. 33, no. 5, pp. 997–1006, 2008.

[9] K. Popuri, D. Cobzas, N. Esfandiari, V. Baracos, and M. Jägersand, "Body composition assessment in axial ct images using fem-based automatic segmentation of skeletal muscle," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 512–520, 2015.

[10] J. Kullberg, A. Hedström, J. Brandberg, R. Strand, L. Johansson, G. Bergström, and H. Ahlström, "Automated analysis of liver fat, muscle and adipose tissue distribution from ct suitable for large-scale studies," *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[11] S. Dabiri, K. Popuri, E. M. C. Feliciano, B. J. Caan, V. E. Baracos, and M. F. Beg, "Muscle segmentation in axial computed tomography (ct) images at the lumbar (l3) and thoracic (t4) levels for body composition analysis," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 47–55, 2019.

[12] P. Hu, Y. Huo, D. Kong, J. J. Carr, R. G. Abramson, K. G. Hartley, and B. A. Landman, "Automated characterization of body composition and frailty with clinically acquired ct," in *International Workshop and Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging*. Springer, 2017, pp. 25–35.

[13] A. D. Weston, P. Korfiatis, T. L. Kline, K. A. Philbrick, P. Kostandy, T. Sakinis, M. Sugimoto, N. Takahashi, and B. J. Erickson, "Automated abdominal segmentation of ct scans for body composition analysis using deep learning," *Radiology*, vol. 290, no. 3, pp. 669–679, 2019.

[14] A. T. Grainger, A. Krishnaraj, M. H. Quinones, N. J. Tustison, S. Epstein, D. Fuller, A. Jha, K. L. Allman, and W. Shi, "Deep learning-based quantification of abdominal subcutaneous and visceral fat volume on ct images," *Academic Radiology*, 2020.

[15] R. Vettor, G. Milan, C. Franzin, M. Sanna, P. De Coppi, R. Rizzuto, and G. Federspil, "The origin of intermuscular adipose tissue and its pathophysiological implications," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 297, no. 5, pp. E987–E998, 2009.

[16] O. Jimenez-del Toro, H. Müller, M. Krenn, K. Gruenberg, A. A. Taha, M. Winterstein, I. Eggel, A. Foncubierta-Rodríguez, O. Goksel, A. Jakab *et al.*, "Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks," *IEEE transactions on medical imaging*, vol. 35, no. 11, pp. 2459–2475, 2016.

[17] "3d slicer." [Online]. Available: https://www.slicer.org/

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[19] A. Smailagic, P. Costa, H. Y. Noh, D. Walawalkar, K. Khandelwal, A. Galdran, M. Mirshekari, J. Fagert, S. Xu, P. Zhang *et al.*, "Medal: Accurate and robust deep active learning for medical image analysis," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 481–488.

[20] N. Lessmann, B. van Ginneken, P. A. de Jong, and I. Išgum, "Iterative fully convolutional neural networks for automatic vertebra segmentation and identification," *Medical image analysis*, vol. 53, pp. 142–155, 2019.

Appendix

1. CIRRUS workstation

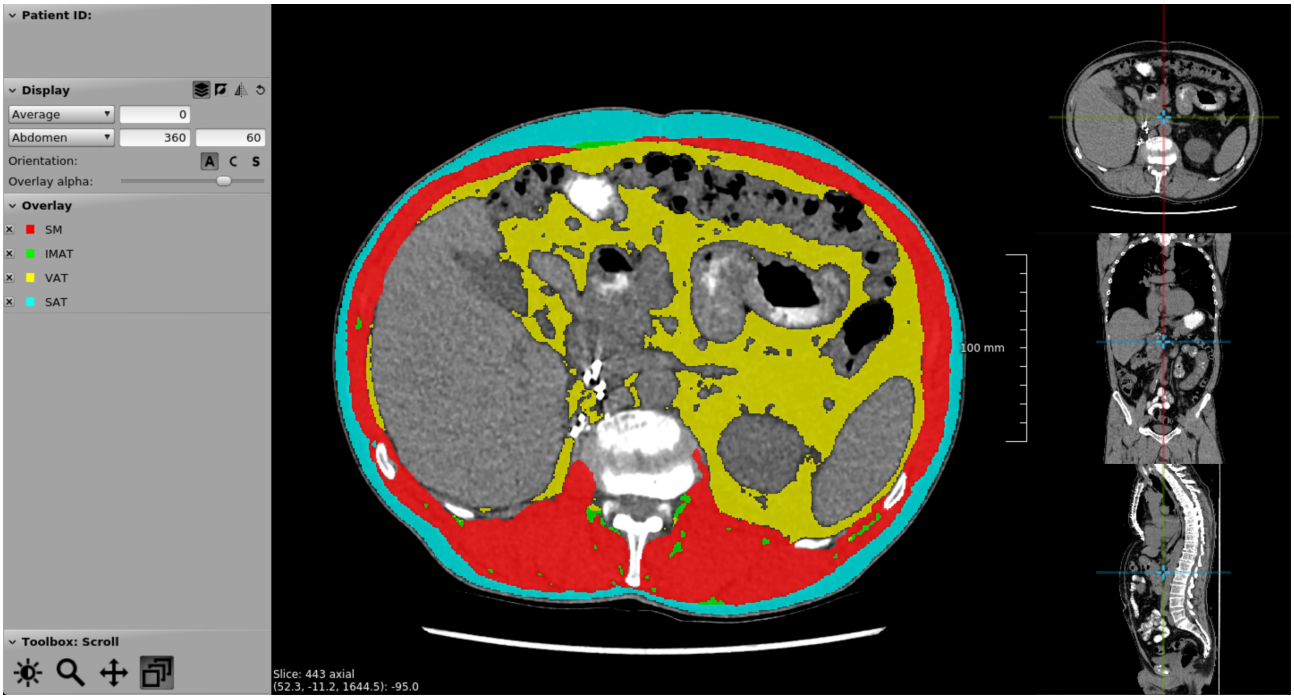

Figure 12: Example output of the model, viewed in the CIRRUS workstation. The workstation allows the user to scroll through the image and view it from different perspectives.

2. Detailed box plots

From the comparison of these two plots, it becomes clear that especially the SM compartment at L5 benefits from the active learning network.

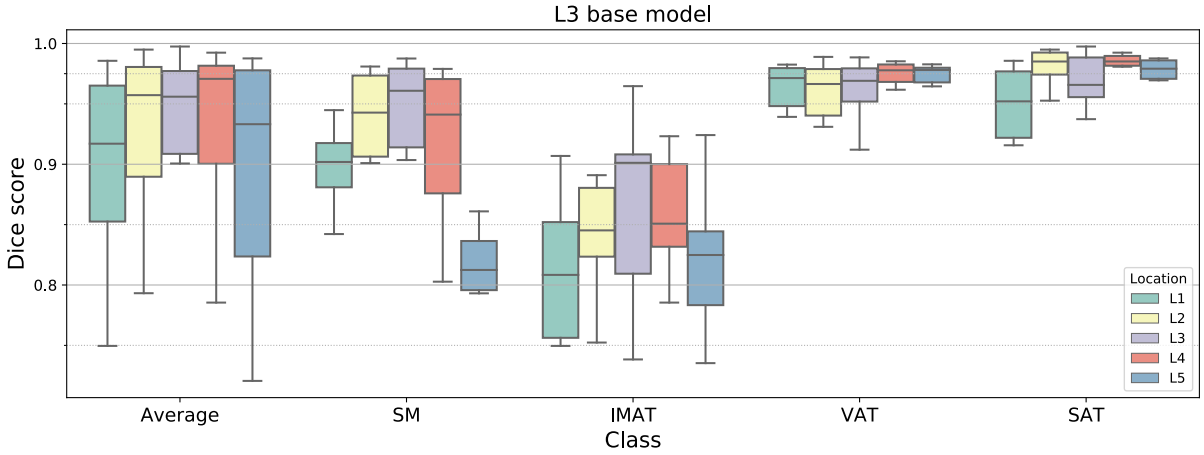

Figure 13: Box plots showing the Dice score for each class in each location for the L3 base model.

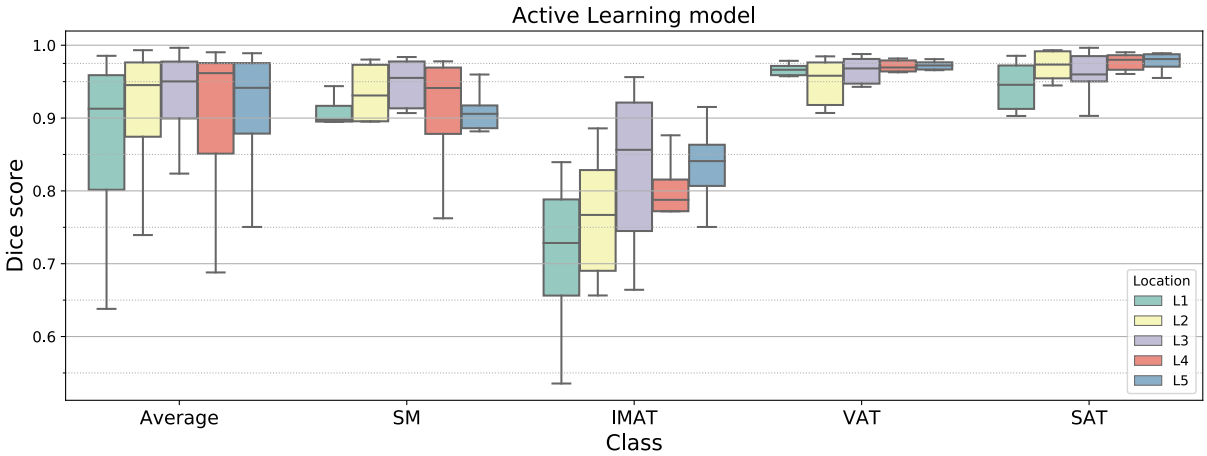

Figure 14: Box plots showing the Dice score for each class in each location for the Active Learning model.

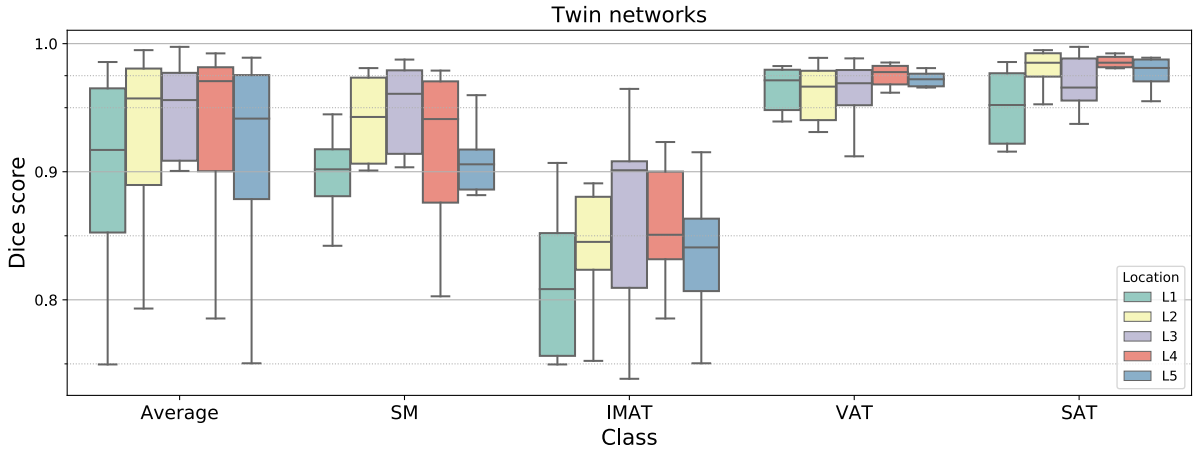

Figure 15: Box plots showing the Dice score for each class in each location for the Twin networks approach.