Automatic extraction of anthropometric features and body composition parameters from computer tomography images enables improved BMI prediction at scale

1st Sebastian Salzmann *Center of Artificial Intelligence*
ZHAW School of Engineering Zuerich, Switzerland
salzmseb@students.zhaw.ch

***Abstract*—The Body Mass Index (BMI) is a known indicator of socio-economic status and health and is usually correlated with obesity. With the prevalence of overweight rising among many populations it inherently becomes an important indicator of population health. Automatic assessment of the BMI at scale can therefore serve as a epidemological health indicator in population studies as well as enable automatic patient health tracking in a clinical context. This study explores the automatic prediction of the Body Mass Index (BMI) out of Computer Tomography(CT) scan images of the lumbar vertebra 3 (L3) region using techniques of Deep Learning, Machine Learning and classical Computer Vision. By automating the measurement of the body diameter and tissue-specific body composition parameters we improve compared to the most accurate method to our knowledge, a linear model using the effective diameter D_{eff}. Combining the body composition parameters with the body diameter we predict the BMI with simpler regression methods resulting in our best model: a Random Forest predicting the BMI with a mean average error of 2.0 and root mean squared error of 2.8 on the test set. We conclude that only through the combination of the various feature extraction techniques outlined we were able to obtain a dataset on which training led to our best result. We furthermore discuss the performance of our methods in context of the clinical environment and find that while the approach works well there are still many possibilities for future research to follow up on.**

***Index Terms*—computer vision, computer tomography, biomedical imaging, image segmentation, body composition parameters, body mass index**

I. Introduction

The worldwide prevalence of obesity has nearly tripled since 1975 and become a major public health concern [1]. Overweight as well as obesity is linked to a variety of diseases such as coronary heart disease [2], strokes [3]. certain types of cancer [4] and Alzheimer [5]. A commonly used indicator correlated with obesity is the Body Mass Index (BMI) [6] which serves as epidemiological health indicator of populations.

The BMI is usually assessed as patient metadata and requires manual labor such as providing infrastructure and personal to guide a patient and capture the parameters. Aside from the labor investment more importantly the manual steps necessary for measurement stand in the way of automation. Automatic assessment of the BMI at scale from existing or routinely acquired data enables more efficient downstream analytics applicable in for example monitoring of population health or assessing a person's level of obesity in hindsight.

Computer Tomography (CT) has become such a routinely used technique as it is used to create scans of areas inside the human body and therefore used as a screening tool for lesions and abnormalities [7]. The derived images contain important information about the anatomy and body composition of a patient [8] and are commonly used to find kidney stones or tumors without having to perform an invasive biopsy [9]. CT body scans consist of stacked cross-sectional 2D images (=slices) building up a 3D representation of the body offering information in abundance. In CT analysis features of interest are usually extracted with the aid of software which can for example detect edges and similarities in areas automatically, enhance contrast and lightning and generally enable interaction with the volumetric 3D or sliced 2D data. With the rise of deep learning in computer vision many of the modern feature extraction tools in radiology rely on neural networks which aid in pattern recognition and label assignment [10], [11]. As input either the entire 3D volume or extracted, representative 2D slices are used [12]. The slice regions are named after their relative position to the spine‘s vertebrae as for example the lumbar vertebra number 3 (L3 vertebra).

Since recognizing and classifying pathological patterns from the image data is performed by trained radiologists increasing the degree of automation with an acceptable margin of error can reduce costs and speed up the process with the hope to improve prognosis, diagnostics and clinical decision support [13], [14]. Predicting important clinical parameters such as the BMI automatically and as reliable as possible is therefore a goal of modern analytical software systems supporting todays and future hospitals.

In this work different techniques of computer vision, machine and deep learning in order to measure the BMI from computer tomography images automatically and as accurate as possible are explored, compared using the same metrics and discussed in terms of performance, resource investment and the possibility of clinical application.

A. Motivation

Measurement of the BMI is primarily of interest in the health sciences as it is a measure correlated with obesity and represents a risk factor for a variety of diseases. With the availability of public image data growing, reflected in the efforts of platforms such as the Cancer Imaging Archive [15] or Stanford AIMI Shared Datasets [16] which de-identify and host a large array of medical image studies . Both initiatives aim for providing high quality image data enabling more large scale analysis in biostatistical, machine learning and deep learning applications. Many datasets however [17], [18] enrich single modalities focusing on the particular research question they address restricting the usability in follow up research that might want to explore a different direction. Additionally ignoring stratification in deep learning model training might encode unwanted biases when records like age, gender, race or health status are not available. Datasets like the RadFusion dataset [19] try to overcome this by publishing image data along with electronic health records and thereby making it possible to detect and remove potential biases which might be unknowingly encoded into resulting models.

BMI prediction from existing CT image data can therefore enrich a dataset which lacks this information and give clues about the health status of a patient at the time of image acquisition. Furthermore it can be used as an automatic routine assessment of a patients state of health over time and thereby support clinical decision making by recommending for example an adjustement to radiation and medication dosages [20].

B. Related work

BMI prediction from image data in general has been performed in various ways mainly without the usage of CT data. Velardo et al. [21] and Jiang et al. [22] for example tried to map anthropometric features found in human body images to BMI values. Another commonly used approach are the images of faces which were also used in different studies [23], [24] by mapping facial features to the BMI.

Closer related is the work of Vakli et al. [25] who used magnetic resonance imaging on brains to determine the BMI using a deep convolutional neural network for feature extraction and regression and therefore bears methodological similarities to our approach. They also used an inductive bias (age, gender) which improved their training performance but not their prediction. Since they used a different imaging technology and selected the brain instead of the abdomen as region of interest their work can be regarded as complimentary.

Our main contributions are therefore addressing the lack of research as regards CT data which contains a rich representation of anatomy and composition of the body and is a commonly used medical imaging technology.

Studywise partially related is the work of O‘Neill et al. who tried to tackle the inverse problem by trying to predict the effective body diameter from the BMI in order to improve size-specific patient radiation dosage recommendation [20]. They were able to build a linear model that can predict the BMI based on the body‘s lateral and anterior-posterior diameter on which we were able to build and improve upon.

BMI Range	BMI class	KITS21 occ.	KSA occ.
0-18.5	Underweight	2	16
18.5-25	Normal	29	118
25-30	Overweight	49	57
30-35	Obese I	32	33
35-∞	Obese II	37	10

TABLE I: BMI class stratification into 5 classes and their occurrence in the datasets

The impact of our approach is making use of the wealth of information CT data offers as we introduce a combinatorial approach: For building our models we do not only use anthropometric measures but also body tissue parameters determined by a semantic segmentation model and are able to improve BMI predictions without much optimization.

II. Experimental Setup

A. Datasets - Preprocessing and description

1) KITS Dataset: The KITS21 dataset [26] consists of $N = 300$ images in NIfTI format together with metadata containing the BMI of the patients. Since the dataset is present as a 3D volume the lumbar vertebra 3 slices were manually selected using the ITKSnap software [27], converted to (1,1,1) pixel spacings (where each pixel to pixel distance corresponds to 1mm), normalized to zero mean and unit variance and saved as numpy arrays at the L3 level in order to perform further analysis. All data was processed using the SimpleITK library [28] in Python3.

For this only a subset of $N = 148$ images was used due to time constraints which is also the subset used in all further analysis except if otherwise indicated.

2) KSA Dataset: The KSA dataset [26] was collected at the Kantons Spital Aargau and consists of a total of $N = 235$ patients. The data is present in Dicom format as 2D slices at lumbar vertebra 3 level together with additional metadata containing the BMI of the patients. Slices were converted to (1,1,1) pixel spacings, normalized to zero mean and unit variance and saved as numpy arrays.

3) BMI value and class distribution: The BMI is usually categorized into different classes [29] as indicated in Table I together with their occurrence in the particular dataset. Figure 1 shows the BMI value distribution of both datasets.

Notably both datasets have a different data distribution with the KITS21 dataset containing more people with a higher BMI while in the KSA dataset most are in the normal range.

B. Evaluation Metrics

1) Regression: As a metric for regression the mean absolute error (MAE) and root mean squared error (RMSE) were used. Both are scale dependant measurements and therefore serve to compare different model performances on the same feature/dataset. N denotes the sample size, y_i the BMI of patient i and $\hat{y}_i$ the predicted BMI of patient i.

$$MAE = \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{N} \tag{1}$$

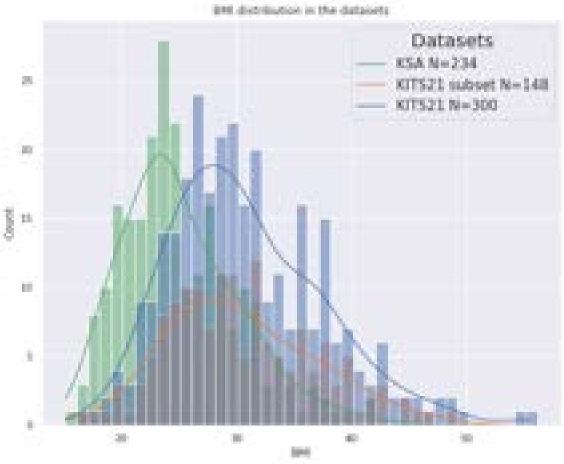


Fig. 1: Shows the BMI value distribution of the KITS21, KITS21 subset (selected as described in Section II-A1) and the KSA dataset.

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(y_i - \hat{y_i})^2}{N}} \tag{2}$$

2) Segmentation: For evaluating the segmentation performance as a first metric the Intersection Over Union (IoU) was used where A are all labelled pixel of the ground truth segmented image and B all labelled pixels of the prediction image. The labels are described in section Figure 4. The equation is denoted below.

$$IoU = \frac{A \cap B}{A \cup B} \tag{3}$$

Additionally the Dice coefficient *Dice* was calculated which is denoted in 4 and represents the ratio of the intersection of A and B over the sum of all pixels in both segmentation images.

$$Dice = \frac{2 * A \cup B}{A + B} \tag{4}$$

These metrics, on which neural net loss functions are based on, are used due to the unbalanced statistical distribution of labels in the image (eg. around half is background) which cannot be appropriately evaluated on image level by for example cross entropy loss calculations [30].

III. Results

Since predicting the BMI is the prediction of a continuous value regression seems to be the more appropriate method compared to classification since stratification into the BMI classes can be performed after obtaining the predicted continuous value.

A. BMI prediction directly from image data using deep learning

The BMI was predicted using raw image data using several difference architectures. The experiment served to get a first impression on how good a deep learning model using standard architectures performs in this case and to be able to select the best encoder given the results. The model architectures were taken from the pytorch torchvision library [31] and enhanced with a regressional output head. Models were trained with a batchsize of 20, learning rate of 5e-3 and an early stopping counter set to 8. When during 8 epochs the validation error would not decrease the training would stop. As loss function a L1 Loss was chosen which minimizes the mean absolute error. Results are visible in Table II.

Model	Epochs	Dataset part	MAE	RMSE
ResNet18	81	train	**3.1**	**3.9**
ResNet18	81	test	**3.6**	**4.8**
ResNet34	72	train	4.0	5.1
ResNet34	72	test	4.8	5.8
VGGNet	51	train	4.6	5.8
VGGNet	51	test	5.1	6.4

TABLE II: Deep learning architectures with regressional output head trained on CT image data alone for BMI prediction. The best scores as indicated in bold font were obtained using ResNet18 encoder.

Out of the baselines the ResNet18 encoder performed the best which is why it was chosen for all subsequent experiments as a baseline architecture.

B. BMI prediction by axis computation

Since O'Neill et al. [20] tried to tackle the inverse problem by trying to predict body internal measurements given the BMI the backwards approach was tested. They found that there is a linear relationship (equation visible in Figure 3) between the effective diameter D_{eff} and the BMI. The effective diameter D_{eff} is composed of the anterior-posterior D_{ap} and the anterior-lateral diameter D_{lat} as visible in Equation 5.

$$D_{eff} = \sqrt{D_{ap} * D_{lat}} \tag{5}$$

To find the region of the body in the picture a connected component analysis [32] was performed by binarizing the image with a Houndsfelt Unit (HU) value of -800 and keeping the largest region as the cut out body part. Then the minor and major axis of the binary mask of the remaining elliptic body shape were computed using scikit-image [33] region property function. Examples of images overlayed with the computed diameters D_{ap} and D_{lat} are visible in Figure 2.

The combination of reducing image artifacts by cutting out the body and axis computation was visually checked for all images of the KITS21 and KSA dataset and considered good yet also imperfect as visible in the example Figure 2. D_{lat} and D_{ap} can be slightly overestimated though the final BMI approximation was found to still correspond well with the true BMI as visible in the results Table III.

From the parameters D_{ap} and D_{lat} the linear model was computed on each of the datasets (KITS and KSA). To

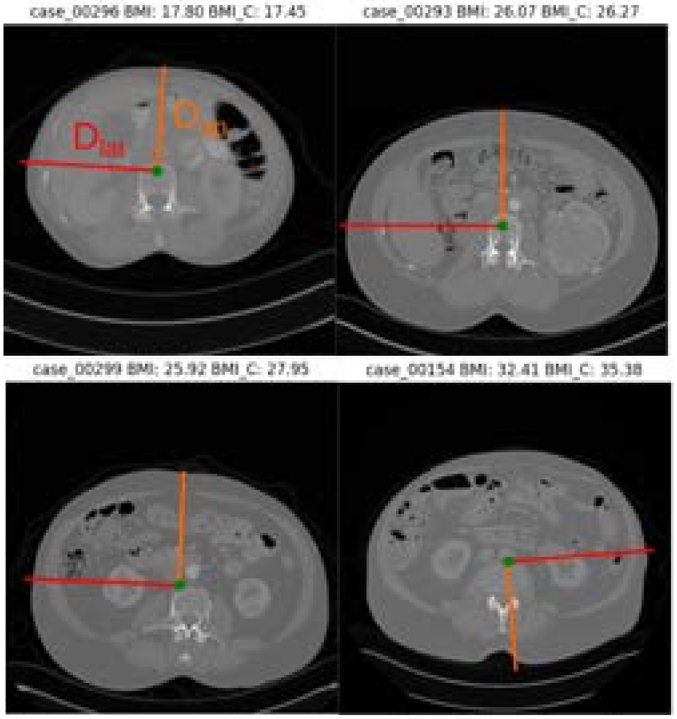

case_00296 BMI: 17.80 BMI_C: 17.45
case_00293 BMI: 26.07 BMI_C: 26.27
D_{lat}
D_{ap}
case_00299 BMI: 25.92 BMI_C: 27.95
case_00154 BMI: 32.41 BMI_C: 35.38

Fig. 2: Example cases where the lateral diameter D_{lat} (red) and anterior-posterior diameter D_{ap} (orange) of CT scans are computed and plotted as overlay on the images. The parameters are used for the linear model calculation resulting in the BMI. True BMI and computed BMI (BMI_C) according to the original model are indicated in the title.

investigate possible variations and also counteract the fact that the original dataset used relatively few samples (N=50) when compared to our datasets (KSA: N=248, Kits: N=148) linear models for each of the datasets was build. Since parameter extraction is automated this is now possible in a reasonable amount of time and without human intervention. The linear models correspond fairly well with the BMI as visible in Figure 3 and Table III.

Model	Dataset	MAE	RMSE
Original	KSA	3.1	3.8
Original	Kits	3.0	3.7
Kits	KSA	2.7	3.3
Kits	Kits	2.5	3.3
KSA	KSA	2.0	2.5
KSA	Kits	3.0	4.0

TABLE III: The table shows the MAE and RMSE metrics of three linear models which are based on the effective diameter D_{eff} and the equations visible in the legend of Figure 3. The models differ in the data they were fit on, namely the original model from the paper (=Original) [20], the the KITS21 dataset (=Kits) and the KSA dataset (=KSA) respectively.

Overall the model fit on the KITS dataset performed the best on our data and yields the same coefficients as when datasets are combined and then used to fit a linear model (data not shown).

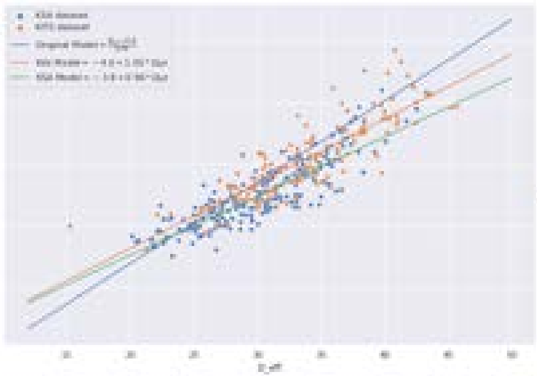


Fig. 3: Displaying the linear models based on different datasets. The orginal model was derived from the referenced paper [20] while the KitsModel was fitted on the KitsData II-A1 and the KSAModel was fitted on the KSAData II-A2 using the linear model function on the effective diameter D_{eff} in R.

C. BMI prediction with body composition parameters obtained by image segmentation

1) Image segmentation model training and performance: 38 Images from the KITS dataset [26] were selected with the following criteria:

1) Body needs to be completely inside the picture
2) Acceptable image quality
3) Stemming evenly drawn from the whole BMI range

The images were labelled according to the labels visible in Figure 4 and processed according to the same preprocessing procedure described in Section II-A1.

The images were split into a test (N=8) and training set (N=29) and fed into the network for either a fixed amount of epochs while taking checkpoint based on the best metric performance of the IoU. A batchsize of 16 and a learning rate of $2e^{-}2$ were chosen. In order to make sure that the training

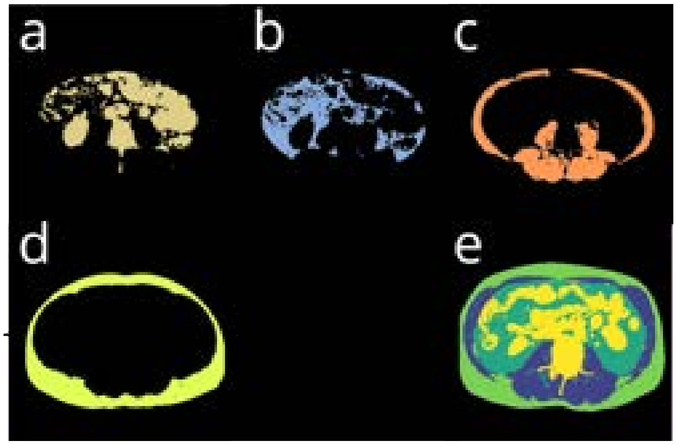


Fig. 4: Showing from top left to bottom right the segmented labels of (a) body background, (b) visceral adipose tissue (VAT), (c) muscles, (d) subcutaneous adipose tissue (SAT) and (e) the whole segmentation.

sees at least once all BMI classes (as defined in Table I) the training and test set were constructed to be evenly distributed. Only the underweight class was decided to be only put into the training so that training would see the only two examples available.

Visual examples of the resulting test set segmentations are shown in Figure 5 comparing the prediction and ground truth side by side.

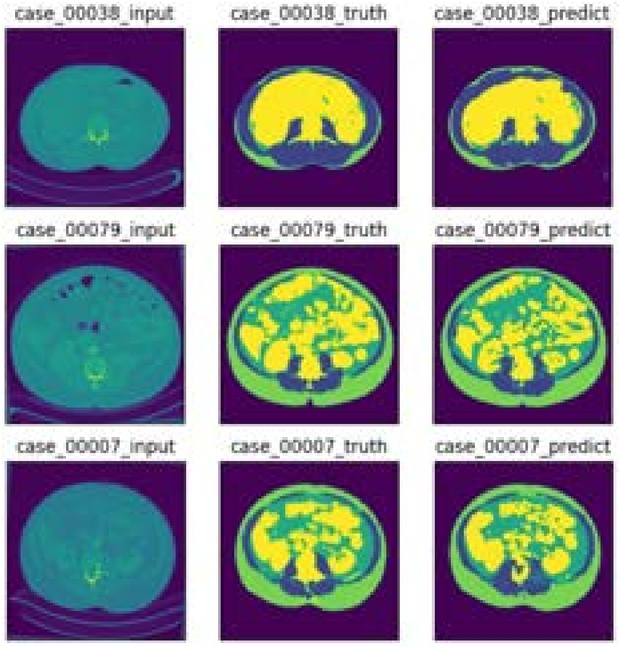


Fig. 5: Segmentation performed on test set using the Unet architecture [34]

As evaluation metrics we used the Intersection over Union and Dice coefficient as described in II-B2. The results are shown in Table V. To obtain the scores each score was calculated for each label separately and averaged across all labels. Additionally the procedure was done using two different Unet model encoders namely the ResNet18 and ResNet34. The ResNet18 encoder here performed better and the results are in bold for the best performance on the test and train set.

The tissue specific IoU and Dice scores were also looked at and are displayed in Table VI. They show a good performance on the test set with an IoU and Dice score above 70% on the test set.

BMI Class	Training set occurrence	Test set occurrence
Underweight	2	0
Normal	7	2
Overweight	6	2
Obese I	5	2
Obese II	2	2

TABLE IV: BMI classes of the constructed segmentation training and test set.

Model	Epochs	Dataset part	IoU	Dice
ResNet18	60	train	0.77 $\mp$ 0.04	0.84 $\mp$ 0.03
ResNet18	60	test	0.67 $\mp$ 0.07	0.76 $\mp$ 0.06
ResNet34	60	train	0.77 $\mp$ 0.04	0.85 $\mp$ 0.03
ResNet34	60	test	0.65 $\mp$ 0.10	0.74 $\mp$ 0.10
ResNet18	100	train	**0.86 $\mp$ 0.03**	0.87 $\mp$ 0.03
ResNet18	100	test	0.67 $\mp$ 0.07	0.78 $\mp$ 0.07
ResNet18	checkpoint	train	0.81 $\mp$ 0.03	**0.88 $\mp$ 0.03**
ResNet18	checkpoint	test	**0.70 $\mp$ 0.08**	**0.79 $\mp$ 0.07**

TABLE V: Image Segmentation metrics IoU and Dice score of the KITS dataset. Two model architectures were used and results are displayed for training and test set as well as for running a fixed epoch amount or using a model checkpoint based on a running IoU evaluation during the training.

Data	VAT	SAT	muscle	body background
train	0.82 $\mp$ 0.11	0.93 $\mp$ 0.05	0.86 $\mp$ 0.06	0.87 $\mp$ 0.06
test	0.67 $\mp$ 0.23	0.82 $\mp$ 0.17	0.73 $\mp$ 0.11	0.78 $\mp$ 0.08

TABLE VI: Tissue specific IoU scores for the KITS dataset when evaluated with the best performing ResNet18 (trained with checkpoint) from Table V.

2) Using Sklearn Regressors on body composition parameters: Body composition parameters were obtained by using image segmentation as described in Section III-C1. The tissue specific pixel counts were used as feature. Using the scikit-learn [35] library the data entries per case were split into train and test set in a 0.8 to 0.2 ratio. Parameters were scaled by subtracting the mean and scaling to unit variance. Scaling on train and test set was performed with the scaling obtained from the training set in order to prevent data leakage.

Regression analysis was performed using Support Vector Regressor (SVR) and Random Forest (RF). The features for training were obtained from the axis computation III-B, body composition prediction III-C. The body perimeter was obtained by using the perimeter function of the scikit image on the cut out body, as described in section III-B. In order to test whether the ground truth segmentations produce better results the original training and test setup used in the image segmentation was used which is a subset of the KITS dataset (N=38) as described in Section III-C1. Random state was fixed to 42 and all combination of the features, D_{ap} , D_{eff}, D_{lat}, body background, VAT, SAT, muscle and body perimeter were fed into the models. Results are visible in Table VII. The feature importance obtained from the Random Forest model was 0.55 for D_{eff}, 0.29 for SAT and 0.16 for D_{ap}. The Support Vector regression solely used D_{eff} and since D_{ap} is essentially contained in D_{eff} this analysis concludes that D_{eff} is a strong predictor of the BMI which contributes to the majority of the predictive power of the models.

Regressor	MAE	RMSE	Feature combination
SVR	2.3	2.7	D_{ap}
RF	1.8	2.2	SAT, D_{ap}, D_{eff}

TABLE VII: Regressions run with the parameters obtained from body axis computation and body composition analysis. The dataset used in this analysis was the subset of data used in the image segmentation using the ground truth body tissue segmentation parameters as performed by the radiologist.

The experiment was repeated using the predicted segmentation parameters for the Random Forest Regressor in order to assess possible performance differences. The Support Vector

Regressor does not include a predicted body composition parameter and was therefore omitted. Results are visible in Table VIII.

Regressor	MAE	RMSE	Feature combination
RF	1.6	2.0	SAT, D_{ap}, D_{eff}

TABLE VIII: Regressions run with the parameters obtained from body axis computation and body composition analysis by image segmentation. The dataset used in this analysis was the subset of data used in the image segmentation with predicted body tissue segmentation parameters.

The results between Table VII and Table VIII for the Random Forest regression stay comparable points to similarity of predicted and ground truth pixel counts as regards SAT tissue.

Since the Random Forest model performed better the model was selected for the follow up experiments.

The experiment was repeated on the KITS dataset (N=148) using the best feature combination obtained from results in Table VII. Additionally once the SAT values from the ground truth (for the train and test subset) segmentation were used combined with the predicted ones and once only predicted ones in order to also get a hint about segmentation prediction performance and possible impact on the regression results. The results are visible in Table IX.

Experiment	MAE	RMSE
Ground truth parameters (train + test)	3.0	3.9
Predicted parameters	3.0	4.4

TABLE IX: Random Forest regression run with the parameters obtained from body axis computation and body composition analysis by image segmentation. The dataset used in this analysis was all KITS data where the L3 region was selected from. In order to also include a measure of segmentation performance once ground truth parameters were mixed with prediction ones and once only predicted ones were used in order to assess the impact on BMI prediction of the model.

At last the correlation coefficients were plotted for body composition and body axis parameters and are visible in the heatmap displayed in Figure 6. The results together with the regression results in Table VII further confirm that D_{eff} (and hence D_{lat} and D_{ap}) is indeed a good predictor of the BMI, moreso than the body composition parameters. The perimeter, a parameter that was assumed to work well for BMI prediction but omitted due to not resulting in the best regression models (visible in Table VII), exhibited also a stronger correlation than the body composition parameters which nevertheless also show interaction with the BMI. Interesting is also the interplay where SAT and VAT positively correlate with each other while muscle mass seems to correlate negatively with SAT and VAT.

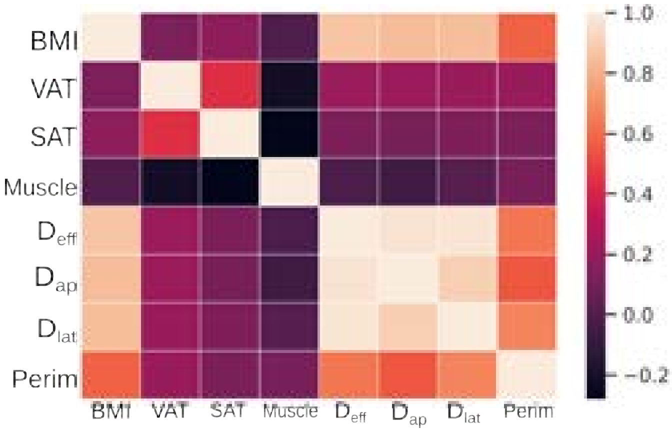


Fig. 6: Correlation coefficient heatmap of body composition parameters and the body axis computations

D. Experiments with combined datasets

In order to gain a larger dataset the subset of the KITS dataset (N=146) was combined with the KSA dataset (N=248) and split into test(0.2), training(0.72) and validation set(0.08). For the combined dataset body composition parameters were predicted using the best performing segmentation model, trained as described in Section III-C. Body axis parameters were determined as described in Section III-B.

1) Deep learning BMI prediction with D_{eff} as inductive bias: Since deep learning model can be trained with inductive biases we trained models using the best predictor D_{eff} found in the previous regression analysis III-C2.

Models were trained with a standard ResNet18 architecture with L1Loss , 1e-2 learning rate, batchsize of 16 and early stopping enabled (8 non consecutive validation epoch decreases of the loss trigger early stopping). ResNetAxis represents a model class which inherits from ResNet18 and concatenates the effective diameter D_{eff} with the model output obtained as described in section III-B. The two values are then passed onto a final fully connected interpretation layer.

Model	Type	MAE	RMSE
ResNetAxis18	test	**2.4**	**3.1**
ResNetAxis18	train	**2.4**	**3.0**
ResNet18	train	3.1	3.9
ResNet18	test	3.6	4.8

TABLE X: The table shows a different model architecture results when trying to predict the BMI value by regression using an L1 Loss on the combined dataset. ResNet18 is a standard Resnet18 as described in section III-A. ResNetAxis18 is the standard ResNet18 which concatenates the with the D_{eff} value and forwards it to a final interpretation output head.

The ResNetAxis18 shows improved performance over the initial model using only raw image data.

2) Sklearn regression on all parameters: In order to compare the above results with the performance of the same regressional analysis as performed in III-C2 the dataset constructed as described in III-D was performed as already done in Section III-C2.

The results are displayed in Table XI for using the best combinations as found by the minimal RMSE and MAE as well as using only the D_{eff} feature in order to make a fair comparison with the ResNetAxis18 and body axis linear model possible.

Regressor	Dataset	MAE	RMSE	Feature combination
SVR	test	2.1	2.7	SAT, D_{ap}, perimeter, D_{eff}
SVR	train	2.1	2.8	SAT, D_{ap}, perimeter, D_{eff}
RF	test	**2.0**	**2.8**	**Muscle, SAT, $\boldsymbol{D_{eff}}$**
RF	train	**0.9**	**1.1**	**Muscle, SAT, $\boldsymbol{D_{eff}}$**
RF	test	2.4	3.2	D_{eff}
RF	train	1.0	1.4	D_{eff}

TABLE XI: Regressions run with the parameters obtained from body axis computation and body composition analysis. The dataset used in this analysis was the combined KITS and KSA dataset with the same datasplit used as in the inductive bias deep neural network experiment in Section III-D1.

Feature importance for the best Random Forest was 0.13 for muscle, 0.15 for SAT and 0.72 for D_{eff}.

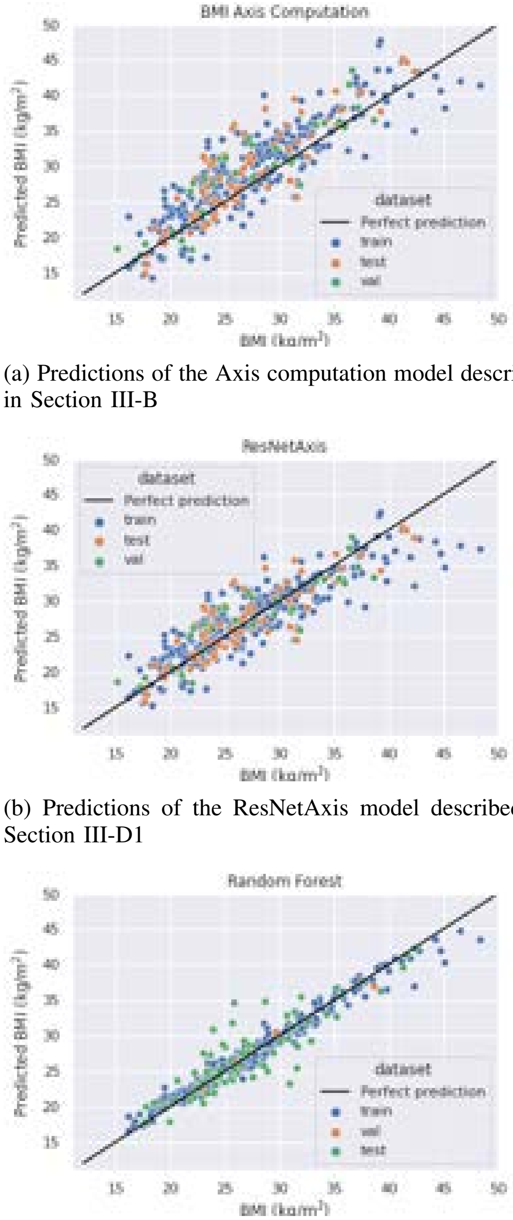
3) Visual comparison of Axis, ResnetAxis and Random Forest predictions: In order to have a look at the performance of the three main methodologies used the predictions for the combined datasets were plotted against the true BMI. The results are visible in Figure 7. Random Forest also visually outperforms the other models in terms of performance which is visible in Figure 7c.

For clinical importance the actual BMI (mis-)classification rates matter and therefore here we plotted the results of the mean average error distribution of the test set for each BMI class together with the total misclassification rate of the particular model for all classes in Figure 8.

IV. Discussion

Measuring the BMI from image data has been performed in various ways though to our knowledge not with the usage of predicted body composition parameters and body axis diameters from computer tomography images.

Our approach can be outlined in the following manner: In Section III-A we applied deep learning directly on image data determining a baseline of what we can achieve without much optimization. In Section III-B we started with extracting anthropometric body axis parameters from the sliced images using classical computer vision techniques. Confirming the predictive capabilities of the effective diameter D_{eff} we incorporated it as an inductive bias into the previous baseline deep learning model in Section III-D1 improving upon the linear model given by inverting the equation of O'Neill et al. [20] (equation is given in the legend of Figure 3). The improved model yielded better MAE, RMSE values than the linear model though when used on a BMI classification task (see Figure8) the missclassification rate does not show improvement and also visual inspection when plotted against a perfect prediction we cannot make out much difference between the two approaches (see Figure 7b and 7a). From the prediction plots it appears as if the linear model got slightly corrected by the deep learning approach though this would have to be confirmed with further experiments as for example comparing the least square fits of both predictions. Taking advantage of the information content of CT images we determined body composition parameters using a Unet [34] trained on labelled image segmentations in Section III-C and ran some initial test and quality controls on the obtained segmentation results by looking at the IoU and Dice values. The segmentation results obtained are not perfect as visible in Table V, VI and Figure 5 though considered good with a score above 0.5 [36]. We continued our analysis using Support Vector and Random Forest with usage of predicted, tissue specific pixel counts which again yielded a slight improvement VII when compared to our previous approaches. We investigated the performance of using predicted pixel counts only and found it to be working similarly when compared to usage of ground truth parameters (see Table VII and VIII). In order

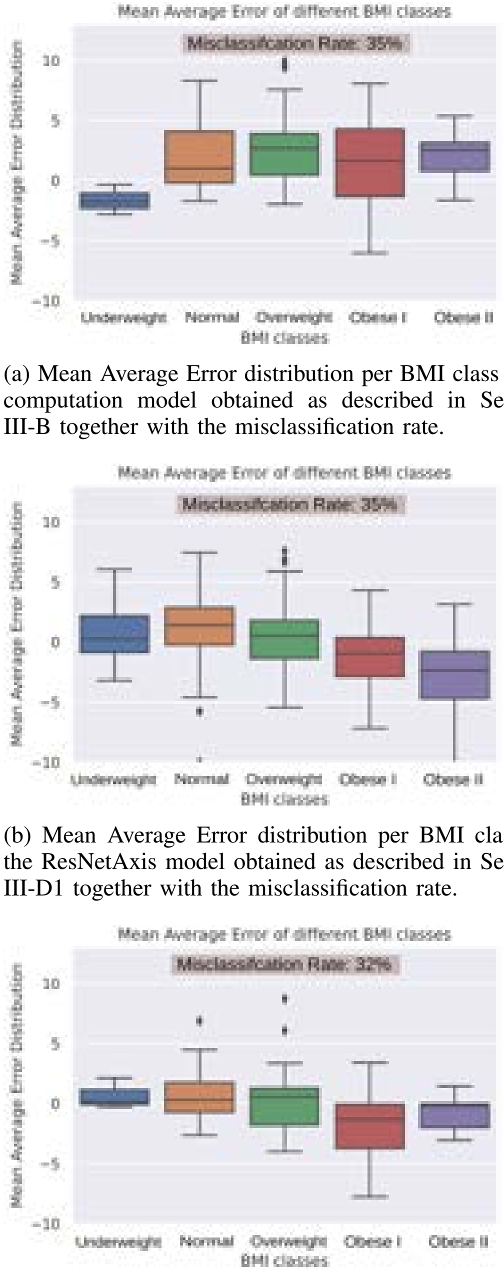


(a) Predictions of the Axis computation model described in Section III-B

(b) Predictions of the ResNetAxis model described in Section III-D1

(c) Predictions of the Random Forest model described in Section III-D2

Fig. 7: Side by side comparison of prediction obtained from three different modelling approaches.



(a) Mean Average Error distribution per BMI class Axis computation model obtained as described in Section III-B together with the misclassification rate.

(b) Mean Average Error distribution per BMI class of the ResNetAxis model obtained as described in Section III-D1 together with the misclassification rate.

(c) Mean Average Error distribution per BMI class of the Random Forest model obtained as described in Section III-D2 together with the misclassification rate.

Fig. 8: Side by side comparison of Mean Average Error distributions per BMI class (as denoted in Section II-A3) obtained from three different modelling approaches on the test set.

to acquire more data to train on we combined the KITS and KSA dataset measuring body axis parameter and predicting the body composition parameters (assuming that the segmentation model would work similarly well on the KSA dataset) with the model obtained in Section III-C1 and ran some simpler regression models which in the end achieved the best results as visible in Table XI and led to our best and final model: a Random Forest trained on the effective diameter D_{eff} as well as muscle and SAT pixel counts. We compare the model side by side as regards missclassication rate and MAE per class(see Figure 8a) which visually shows the increased per class performance and also indicates that the class with the biggest error seems to be Obese I. True vs. predicted BMI are also plotted for all models in Figure 7 which also visually supports Random Forest as the most appropriate model since the dataspread of test and validation data points seem less compared to the axis computation and ResNetAxis model. The resulting best model therefore was built on a combination of features obtained from body axis computation and body composition parameter prediction yielding an improved BMI prediction model.

A. *Limitations*

A limitation quite common for biological datasets was the difficulty to find public datasets which contain both BMI and the lumbar region L3. As mentioned in our motivational Section I-A this might be an informational gap that we can ameliorate with this work since predictions can now be obtained at scale on image data alone. As a result of the difficulty to find appropriate data our collection merely contains $N = 406$ images. We assume that models would to do better with additional data since especially marginal BMI groups as the underweight or obese II class (see Table I) are underrepresented and therefore lead to class imbalance. As a result the random datasplit performed in the last experiments using both datasets does not include any BMI‘s (see Figure 7a) above 42 and our tissue segmentation model does not include an underweight example in the test set (see Table IV) since it was deemed more important to use the two examples available for training. These are merely two examples from our current experimental setup yet would additional data also allow further clinically relevant stratification according to for example gender and age since the BMI is clearly dependant on the latter [37] and body composition on the first [38].

Additionally we have to ask what performance is needed and how much should we invest. The data for using the linear model in Section III-B is easily obtained and already performs quite well compared to our final model from Section III-D2. The final Random Forest trained on additional body composition parameters requires the training of a more sophisticated segmentation deep learning model and the additional time and resource investment which needs to put into labelling and training might not be feasible for the increase in performance. On the other hand could the obtained body composition parameter enable other types of analysis such as direct body fat or muscle mass measurements and therefore find further use. As regards application in a clinical environment our work would consequently need to be examined given a particular context as our best performance (MAE: 2.0, RMSE: 2.8 on the test set from Table XI) might be sufficient for some applications but lacking for others and body composition parameters more or less important to obtain.

B. *Considerations for future work and conclusion*

There are many research questions still left unexplored which might be interesting to follow up. As regards our modelling approaches the inclusion of the effective diameter as an inductive bias to deep learning model using image data directly as an input yielded an improvement over the original linear model (see Section III-D1). Having said that the design decision to concatenate D_{eff} at the end of the network and usage of a subsequent interpretation layer before the output head are purely intuitive decisions where a different approach might lead to better results. An idea for example could be to train the model first without the parameter then freeze the model parameters and only then introduce the effective diameter as inductive bias in order to not steer the inital model learning too much towards the already strong predictive bias of the parameter.

Another avenue unexplored is introducing the body composition parameters in a meaningful manner. We tried to incorporate them as additional binary masks as paralell image channels to the original image though the trained models exhibited worse performance and were therefore omitted in this work. Since body composition parameters do represent features useful in predicting the BMI (see Section III-C2) a different design decision might lead to increased performance.

Combining our results with predictions from other scientific findings might also lead to further interesting scientific insights. Itani et al. for example tried to predict the body fatness percentage given the BMI and several other parameters such as age and gender [39]. Their results could be used for validating and fine tuning our tissue segmentation model which yields SAT and VAT values enabling a body fat measurement which for example has been shown to be a better indicator of insulin resistance than the BMI [40].

Another open question that could be investigated is whether the the L3 actually the appropriate slice to study. While it does represent a normative region [41] which is often looked at by radiologists volumetric CT contains information about the entire acquired body shape and composition. Scanning of various body regions in 3D or in a combinatorial approach rather than singular slices might build up improved representations of body shape and composition and thereby increase predictive performance of BMI models.

As regards the BMI as population health metric volumetric CT data might yield yet another advantage. Since muscle mass actually mediates associations of the BMI with mortality and adiposity [42] body composition analysis in CT data could result in a body composition adjusted BMI which takes internal body measurements under consideration correcting the current shortcomings of the BMI.

Conclusively we were able to automatically extract features from CT images of the spinal L3 region and build accurate BMI models with different degrees of sophistication that can be used to enrich existing datasets, indicate a patients state of health in hindsight or be used as a population health indicator on epidemiological studies performed at larger scale.

References

[1] World, Health, and Organisation. Obesity and overweight. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[2] E. W. Gregg, Y. J. Cheng, B. L. Cadwell, G. Imperatore, D. E. Williams, K. M. Flegal, K. M. V. Narayan, and D. F. Williamson, "Secular Trends in Cardiovascular Disease Risk Factors According to Body Mass Index in US Adults," *JAMA*, vol. 293, no. 15, pp. 1868–1874, 04 2005. [Online]. Available: https://doi.org/10.1001/jama.293.15.1868

[3] C. Curioni, C. André, and R. Veras, "Weight reduction for primary prevention of stroke in adults with overweight or obesity," *Cochrane Database of Systematic Reviews*, no. 4, 2006. [Online]. Available: https://doi.org//10.1002/14651858.CD006062.pub2

[4] A. McTiernan, "Obesity and cancer: the risks, science, and potential management strategies," *Oncology (Williston Park, N.Y.)*, vol. 19, no. 7, pp. 871—81; discussion 881—2, 885—6, June 2005. [Online]. Available: http://europepmc.org/abstract/MED/16053036

[5] S. Alford, D. Patel, N. Perakakis, and C. S. Mantzoros, "Obesity as a risk factor for alzheimer's disease: weighing the evidence," *Obesity Reviews*, vol. 19, pp. 269–280, 2 2018.

[6] J. Pasco, G. Nicholson, S. Brennan, and M. Kotowicz. Prevalence of obesity and the relationship between the body mass index and body fat: cross-sectional, population-based data.

[7] Computed tomography (ct). [Online]. Available: https://www.nibib.nih.gov/science-education/science-topics/computed-tomography-ct

[8] M. Paris, "Body composition analysis of computed tomography scans in clinical populations: The role of deep learning," *Lifestyle Genomics*, vol. 13, pp. 1–4, 12 2019.

[9] W. Brisbane, M. R. Bailey, and M. D. Sorensen, "An overview of kidney stone imaging techniques," *Nature reviews. Urology*, vol. 13, p. 654, 11 2016. [Online]. Available: /pmc/articles/PMC5443345/ /pmc/articles/PMC5443345/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5443345/

[10] S. Koitka, L. Kroll, E. Malamutmann, A. Oezcelik, and F. Nensa, "Fully automated body composition analysis in routine CT imaging using 3D semantic segmentation convolutional neural networks," *European Radiology*, vol. 31, no. 4, pp. 1795–1804, Apr. 2021. [Online]. Available: https://doi.org/10.1007/s00330-020-07147-3

[11] Y. Fu, J. E. Ippolito, D. R. Ludwig, R. Nizamuddin, H. H. Li, and D. Yang, "Automatic segmentation of CT images for ventral body composition analysis," *Medical Physics*, vol. 47, no. 11, pp. 5723–5730, Nov. 2020, arXiv: 2009.01965. [Online]. Available: http://arxiv.org/abs/2009.01965

[12] L. Song, H. Wang, and Z. J. Wang, "Bridging the gap between 2d and 3d contexts in ct volume for liver and tumor segmentation," *IEEE journal of biomedical and health informatics*, vol. 25, pp. 3450–3459, 9 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33905339/

[13] P. Esmaeilzadeh, "Use of AI-based tools for healthcare purposes: a survey study from consumer's perspectives," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 170, Dec. 2020. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01191-1

[14] S. Romero-Brufau, K. D. Wyatt, P. Boyum, M. Mickelson, M. Moore, and C. Cognetta-Rieke, "A lesson in implementation: A pre-post study of providers' experience with artificial intelligence-based clinical decision support," *International Journal of Medical Informatics*, vol. 137, p. 104072, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505619310123

[15] The cancer imaging archive. [Online]. Available: https://www.cancerimagingarchive.net/about-the-cancer-imaging-archive-tcia

[16] Standford aimi shared datasets. [Online]. Available: https://stanfordaimi.azurewebsites.net/about

[17] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C. Fu, X. Han, P. Heng, J. Hesser, S. Kadoury, T. K. Konopczynski, M. Le, C. Li, X. Li, J. Lipková, J. S. Lowengrub, H. Meine, J. H. Moltz, C. Pal, M. Piraud, X. Qi, J. Qi, M. Rempfler, K. Roth, A. Schenk, A. Sekuboyina, P. Zhou, C. Hülsemeyer, M. Beetz, F. Ettlinger, F. Grün, G. Kaissis, F. Lohöfer, R. Braren, J. Holch, F. Hofmann, W. H. Sommer, V. Heinemann, C. Jacobs, G. E. H. Mamani, B. van Ginneken, G. Chartrand, A. Tang, M. Drozdzal, A. Ben-Cohen, E. Klang, M. M. Amitai, E. Konen, H. Greenspan, J. Moreau, A. Hostettler, L. Soler, R. Vivanti, A. Szeskin, N. Lev-Cohain, J. Sosna, L. Joskowicz, and B. H. Menze, "The liver tumor segmentation benchmark (lits)," *CoRR*, vol. abs/1901.04056, 2019. [Online]. Available: http://arxiv.org/abs/1901.04056

[18] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Medical Image Computing and*

Computer-Assisted Intervention – MICCAI 2015, N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 556–564.

[19] Y. Zhou, S.-C. Huang, J. A. Fries, A. Youssef, T. J. Amrhein, M. Chang, I. Banerjee, D. Rubin, L. Xing, N. Shah, and M. P. Lungren, "Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr," 2021. [Online]. Available: https://arxiv.org/abs/2111.11665

[20] S. O'Neill, R. Kavanagh, B. Carey, N. Moore, M. Maher, and O. O'Connor, "Using body mass index to estimate individualised patient radiation dose in abdominal computed tomography," *European Radiology Experimental*, vol. 2, 12 2018.

[21] C. Velardo and J.-L. Dugelay, "Weight estimation from visual body appearance," in *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2010, pp. 1–6.

[22] M. Jiang and G. Guo, "Body weight analysis from human body images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2676–2688, 2019.

[23] G. Bolukbaş, E. Başaran, and M. E. Kamaşak, "Bmi prediction from face images," in *2019 27th Signal Processing and Communications Applications Conference (SIU)*, 2019, pp. 1–4.

[24] L. Wen and G. Guo, "A computational approach to body mass index prediction from face images," *Image and Vision Computing*, vol. 31, no. 5, pp. 392–400, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885613000462

[25] P. Vakli, R. J. Deák-Meszlényi, T. Auer, and Z. Vidnyánszky, "Predicting body mass index from structural mri brain images using a deep convolutional neural network," *Frontiers in Neuroinformatics*, vol. 14, p. 10, 3 2020.

[26] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han *et al.*, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge," *Medical Image Analysis*, p. 101821, 2020.

[27] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[28] R. Beare, B. Lowekamp, and Z. Yaniv, "Image segmentation, registration and characterization in r with simpleitk," *Journal of Statistical Software*, vol. 86, no. 8, p. 1–35, 2018. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v086i08

[29] C. B. Weir and A. Jan. Bmi classification percentile and cut off points. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK541070/

[30] S. Du. Understanding dice loss for crisp boundary detection. [Online]. Available: https://medium.com/ai-salon/understanding-dice-loss-for-crisp-boundary-detection-bb30c2e5f62b

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[32] M. Jephraim. Image processing with python: Connected components and region labelling. [Online]. Available: https://medium.com/swlh/image-processing-with-python-connected-components-and-region-labeling-3eef1864b951

[33] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "scikit-image: image processing in python," *PeerJ*, vol. 2, p. e453, 2014.

[34] W. Weng and X. Zhu, "U-net: Convolutional networks for biomedical image segmentation," *IEEE Access*, vol. 9, pp. 16 591–16 603, 5 2015. [Online]. Available: https://arxiv.org/abs/1505.04597v1

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[36] A. Rosebrock. (2016, 11) Intersection over union (iou) for object detection - pyimagesearch. [Online]. Available: https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

[37] I. K. Karlsson, K. Lehto, M. Gatz, C. A. Reynolds, and A. K. D. Aslan, "Age-dependent effects of body mass index across the adult life span on the risk of dementia: a cohort study with a genetic approach," *BMC medicine*, vol. 18, 6 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32513281/

[38] M. A. Bredella, "Sex differences in body composition," *Advances in experimental medicine and biology*, vol. 1043, pp. 9–27, 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29224088/

[39] L. Itani, H. Tannir, D. El Masri, D. Kreidieh, and M. El Ghoch, "Development of an easy-to-use prediction equation for body fat percentage based on bmi in overweight and obese lebanese adults," *Diagnostics*, vol. 10, no. 9, 2020. [Online]. Available: https://www.mdpi.com/2075-4418/10/9/728

[40] G. H. Goossens, "The metabolic phenotype in obesity: Fat mass, body fat distribution, and adipose tissue function," *Obesity facts*, vol. 10, pp. 207–215, 7 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28564650/

[41] S. Belharbi, C. Chatelain, R. Hérault, S. Adam, S. Thureau, M. Chastan, and R. Modzelewski, "Spotting l3 slice in ct scans using deep convolutional network and transfer learning," *Computers in Biology and Medicine*, vol. 87, pp. 95–103, 8 2017.

[42] M. K. Abramowitz, C. B. Hall, A. Amodu, D. Sharma, L. Androga, and M. Hawkins, "Muscle mass, bmi, and mortality among adults in the united states: A population-based cohort study," *PloS one*, vol. 13, 4 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29641540/