# A Deep Learning Model to Automate Skeletal Muscle Area Measurement on Computed Tomography Images

Kaushalya C. Amarasinghe [1,2], Jamie Lopes [3], Julian Beraldo [4], Nicole Kiss [5,6], Nicholas Bucknell [2,7], Sarah Everitt [2,4], Price Jackson [2,8], Cassandra Litchfield [1], Linda Denehy [6,9], Benjamin J. Blyth [3], Shankar Siva [2,7], Michael MacManus [2,7], David Ball [2,7], Jason Li [1,2†] and Nicholas Hardcastle [8,10*†]

[1] Bioinformatics Core Facility, Cancer Research Division, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia, [2] Sir Peter MacCallum Department of Oncology, The University of Melbourne, Parkville, VIC, Australia, [3] Cancer Research Division, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia, [4] Radiation Therapy, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia, [5] Institute for Physical Activity and Nutrition (IPAN), Deakin University, Geelong, VIC, Australia, [6] Allied Health Department, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia, [7] Radiation Oncology, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia, [8] Physical Sciences, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia, [9] Melbourne School of Health Sciences, The University of Melbourne, Melbourne, VIC, Australia, [10] Centre for Medical Radiation Physics, University of Wollongong, Wollongong, NSW, Australia

**Background:** Muscle wasting (Sarcopenia) is associated with poor outcomes in cancer patients. Early identification of sarcopenia can facilitate nutritional and exercise intervention. Cross-sectional skeletal muscle (SM) area at the third lumbar vertebra (L3) slice of a computed tomography (CT) image is increasingly used to assess body composition and calculate SM index (SMI), a validated surrogate marker for sarcopenia in cancer. Manual segmentation of SM requires multiple steps, which limits use in routine clinical practice. This project aims to develop an automatic method to segment L3 muscle in CT scans.

**Methods:** Attenuation correction CTs from full body PET-CT scans from patients enrolled in two prospective trials were used. The training set consisted of 66 non-small cell lung cancer (NSCLC) patients who underwent curative intent radiotherapy. An additional 42 NSCLC patients prescribed curative intent chemo-radiotherapy from a second trial were used for testing. Each patient had multiple CT scans taken at different time points prior to and post- treatment (147 CTs in the training and validation set and 116 CTs in the independent testing set). Skeletal muscle at L3 vertebra was manually segmented by two observers, according to the Alberta protocol to serve as ground truth labels. This included 40 images segmented by both observers to measure inter-observer variation. An ensemble of 2.5D fully convolutional neural networks (U-Nets) was used to perform the segmentation. The final layer of U-Net produced the binary classification of the pixels into

muscle and non-muscle area. The model performance was calculated using Dice score and absolute percentage error (APE) in skeletal muscle area between manual and automated contours.

**Results:** We trained five 2.5D U-Nets using 5-fold cross validation and used them to predict the contours in the testing set. The model achieved a mean Dice score of 0.92 and an APE of 3.1% on the independent testing set. This was similar to inter-observer variation of 0.96 and 2.9% for mean Dice and APE respectively. We further quantified the performance of sarcopenia classification using computer generated skeletal muscle area. To meet a clinical diagnosis of sarcopenia based on Alberta protocol the model achieved a sensitivity of 84% and a specificity of 95%.

**Conclusions:** This work demonstrates an automated method for accurate and reproducible segmentation of skeletal muscle area at L3. This is an efficient tool for large scale or routine computation of skeletal muscle area in cancer patients which may have applications on low quality CTs acquired as part of PET/CT studies for staging and surveillance of patients with cancer.

Keywords: deep learning, convolutional neural networks, skeletal muscle, image segmentation, sarcopenia, lung cancer

## INTRODUCTION

Loss of skeletal muscle (SM) mass is an important consideration in oncologic patients as a key component of cancer-related malnutrition, sarcopenia and cachexia (1, 2). The loss of skeletal muscle occurring in these conditions has been linked with diminishing physical function (3, 4), increased risk of chemotherapy-related toxicities (5) and unfavorable survival outcomes (6–8). Early diagnosis and intervention with nutrition and exercise however, may improve outcomes in patients with loss of skeletal muscle (4). Importantly, weight and body mass index (BMI) alone are not good predictors of sarcopenia or cancer-related malnutrition (4, 9). Therefore, specifically in the oncology setting, there is a clear need to identify the presence of low skeletal muscle mass and intervene as necessary to reduce adverse effects.

Computed Tomography (CT) is proven to be an effective method to evaluate total body SM mass. In particular, cross-sectional area of SM at the third lumbar (L3) vertebra on abdominal CTs has been found to be highly correlated with the total body SM mass (10, 11). SM area at L3 normalized by patient height is commonly used as a surrogate marker of sarcopenia in cancer (9, 12) and as a component of recent diagnostic criteria for malnutrition and sarcopenia (2, 13). This marker is known as the L3 skeletal muscle index (SMI). Accurate segmentation of SM on CT is a time-consuming task and requires specific skill, training and experience, which limits the measurement in routine clinical practice as well as for large cohort studies. The advances in deep learning and computing resources in recent years provide novel opportunities to revisit these types of manual, time-consuming and routine tasks. In particular, deep learning has been shown to be particularly well suited to segmentation tasks (14, 15).

Previous work has demonstrated high accuracy deep learning segmentation in skeletal muscle of diagnostic quality CT scans acquired for a range of cancer and non-cancer indications (16–18). Positron Emission Tomography PET/CT studies are standard of care in staging and surveillance for a range of cancers (19–21), and are typically whole body acquisitions therefore are well suited for measurement of L3 skeletal muscle area. CT scans acquired during these studies are typically low quality, often only acquired for attenuation correction (22). These CT scans are acquired with reduced current to reduce patient dose, and no intravenous contrast, resulting in increased noise and reduced soft-tissue contrast (23). AI segmentation as trained on high quality diagnostic CT images may thus not be applicable to low quality CT scans such as those obtained in PET/CT studies. The current study aims to use a 2.5D convolutional neural network (CNN) based model to automatically segment SM area at L3 on low quality CT scans acquired as part of PET/CT studies.

## METHODS

### Study Design

A CNN based deep learning model was trained to automatically segment the skeletal muscle in an axial L3 slice of a full body CT scan. Manual segmentation was performed by a single observer according to the Alberta protocol (10, 11). A data set consisting of 147 scans obtained from 66 patients were used as the training and validation set (**Table 1**). A separate data set of 116 CTs from 42 patients were used to independently test the model (**Table 1**). The accuracy of the CNN model was assessed by comparing the overlap between manual and automatic skeletal muscle contours (Dice score) and percentage error between manual and CNN

**TABLE 1 |** Patient and scanner information.

|  | Cohort 1 (Training and validation set) | Cohort 2 (Independent testing set) |
|---|---|---|
| No. of Patients | 66 | 42 |
| No. of scans | 147 | 116 |
| Age at first scan (mean ± SD) | 66.94 ± 9.81 | 67.03 ± 8.72 (based on 35 patients) |
| Gender (female/male) | 24/42 | 9/26 (based on 35 patients) |
| Slice thickness | Average 3.0 mm (range 0.6-5.0) | Average 3.3 mm (range 3.0-3.3) |
| PET/CT Scanner | **Nine** different scanners from **eight providers** | **Two** different scanners from **one provider** |
| Scanner energy | 80 – 140 kV | 140 kV |
| Manual segmentation | Observer A | Observer B |

based contours. Approval to conduct this retrospective study was granted by the Institutional Research and Ethics Committee.

## Training and Validation Dataset

The training and validation dataset consisted of 66 patients who underwent repeat FDG PET/CT scans as part of a prospective observational trial investigating lung function PET/CT (ACTRN12613000061730). The trial protocol has been previously described (24); inclusion criteria was any patient receiving curative intent radiotherapy for NSCLC, with or without chemotherapy. Each patient had one baseline scan and up to three follow-up scans. Overall, the training dataset consists of 147 CTs (**Table 1** and **Supplementary Table 1**). Manual segmentation was performed by a single observer (observer A). This dataset will be referred to as Cohort 1.

### Independent Testing dataset

Independent Testing Dataset The dataset was derived from a prospective observational trial of 42 patients with NSCLC, which investigated the associations between interim tumor responses on 18F-flurodeoxyglucose (18F-FDG) PET/CT and 18F-fluoro-thymidine (18F-FLT) PET/CT and patient outcomes including progression-free survival and overall survival (ACTRN 12611001283965). The methodology for the original study has previously been described (25); inclusion criteria was any patient receiving curative intent radiotherapy for NSCLC, with or without chemotherapy. The survival outcomes and the relationship between skeletal muscle loss was described in (26). There were 42 patients in the testing set with NSCLC, with a base-line scan and up to 4 follow-up scans. In total, the testing dataset consists of 116 CTs (**Table 1** and **Supplementary Table 1**). Manual segmentation was performed by a single observer (observer B). This dataset will be referred to as Cohort 2.

## Manual Segmentation

Manual segmentation of the skeletal muscle at an axial L3 slice of the full body CT scan was performed according to the Alberta protocol (10, 11). Briefly, the skeletal muscle including external and internal oblique, psoas, paraspinal, transverse abdominis and rectus abdominis was segmented using Hounsfield Unit (HU) thresholds of -29 – 150. This was manually adjusted to exclude ligaments and connective tissue around the vertebra. A single expert trained in the Alberta protocol (NK) supervised two observers to perform the segmentation according to the protocol. The training and test datasets were contoured by different observers, who performed cross comparison and review of segmentations with the expert as well as medical staff on the study (NB). Subsequent revisions of segmentations were performed if deemed necessary to adhere to the Alberta protocol. To measure inter-observer consistency and to provide inter-observer context to the automated segmentation results, the two observers each performed segmentation on 20 images of the other observer's data, resulting in 20 images from each data set with segmentation from both observers. The inter-observer difference was computed using Dice score and absolute percentage difference.

## Neural Network Development

We experimented with several model architectures, cross-validation designs, loss functions, augmentation techniques, and optimization methods. All the models were implemented using PyTorch. The initial model consisted of a variation of 2D U-Net (27) (**Figure S1**). The model was trained by dividing the patients in Cohort 1 into two disjoint groups: training and validation. This ensures that our model is never trained and validated on scans from the same patient. The validation set had 10% of randomly selected patients. The model was trained using several loss functions including binary cross entropy loss, Dice loss (28) and focal loss (29). The trained model was evaluated using the validation set at each epoch in terms of network loss and Dice score. The model with the best average validation Dice score was retained and tested using Cohort 2. We achieved a median absolute percentage error (APE) in skeletal muscle area of 3.82% under the 2D model. The 2D model had poor performance mainly in distinguishing SM from other organs such as liver.

A 3D U-Net (30) model can improve the predictions by analyzing 3D volumes simultaneously, which mimics the manual segmentation procedure. Therefore, we tested a 2.5D U-Net architecture for SM segmentation; we have defined this as 2.5D as we have constrained the model to three axial slices, as opposed to a full 3D volume. Two setups were used to train the 2.5D model. Firstly, we used the same training and validation sets from Cohort 1 and trained one 2.5D model. Secondly, we trained an ensemble of 2.5D models. We divided the training patients (Cohort 1) into five groups, namely CV1 to CV5, to train five 2.5D models (5-fold cross validation). At each training round, we held back one group of patients as validation set and used the other four groups for the training process. This model achieved a median APE of 1.46%. Therefore, we discarded the 2D model in favor of the 2.5D model. The ensemble technique is used to tackle overfitting potential in a small sample size problem. It is expected that the five individual models may have varying performance, but can collectively give a consensus decision that outperforms traditional training.

The following sections describe in detail, the 2.5D model architecture, loss function, optimization and neural network training.

## 2.5D U-Net Architecture

U-Net (27, 30) is a type of fully convolutional neural network (FCNN). Mainly, it consists of a contracting path and an expanding path (**Figure 1**). First, the input goes through the contracting path, which consists of convolutional blocks and focuses on finer details of the image at the expense of losing spatial information. Each convolutional block in the contracting path consists of two sets of 2.5D convolutional steps, batch normalization and rectified linear unit (ReLU) activations. All convolutions are 3x3x3 with stride 1 and padding 1. Finally, a max pooling step is performed for down sampling (halving the feature set) from one encoder block to the other encoder block down the line.

Secondly, during the expanding path, the spatial information is recovered by means of skip connections (**Figure 1**). Before going through the convolutional block, the input undergoes an up-sampling step to expand dimensions. The up-sampling is done by means of trilinear interpolation. The expanding path consists of three convolutional blocks. Here, each convolution block comprises of three sets of 2.5D convolutions followed by batch normalization and ReLU. All convolutions are 3x3x3 with stride 1 and padding 1.

## Input to the Model

The input to the model is the three axial slices from the CT scan, which consists of the L3 slice and adjacent slices on top and bottom of L3. The input is pre-processed by replacing the pixel values outside -29 to +150 HU (Hounsfield unit) range by 0. Further, input is resized to 256x256x3 from a 512x512x3 image.

## Training the Proposed Model Architecture

Training data set consisted of CT image stacks of the three slices and manually contoured skeletal muscle at L3. To increase the number of training data, we performed data augmentation 'on the fly'. Specifically, we performed horizontal flip, vertical flip and addition of Gaussian noise. The use of Gaussian noise is a technique to improve generalization ability of the trained model, implicitly assuming that CT images can be degraded with a Gaussian noise component. It is noteworthy that other augmentation techniques – cropping, rotation, random translation and elastic transformation – were deliberately omitted in the final model as they did not improve performance. The loss function was a combination of Dice loss (28) and focal loss (weighted cross entropy loss) (29) and model weights were optimized using "Adam" (31) optimization technique during training. The network was trained up to 300 epochs. The best model was selected based on the model's accuracy, which was measured by Dice score, on the validation set at each epoch. We performed 5-fold cross validation (CV) and retained the best model for each fold (**Figure S2**). The entire cross-validation model training process took ~14 hours.

## Output of the Model

The output of the model is a probability map of the image pixels. The probabilities give the model's confidence on predicting each pixel being inside or outside of skeletal muscle area. The output, which is 256x256x3, is scaled up to match the original image dimensions of 512x512x3. Pixels with a probability above 0.3



**FIGURE 1** | 2.5D U-Net like architecture used in the current model.
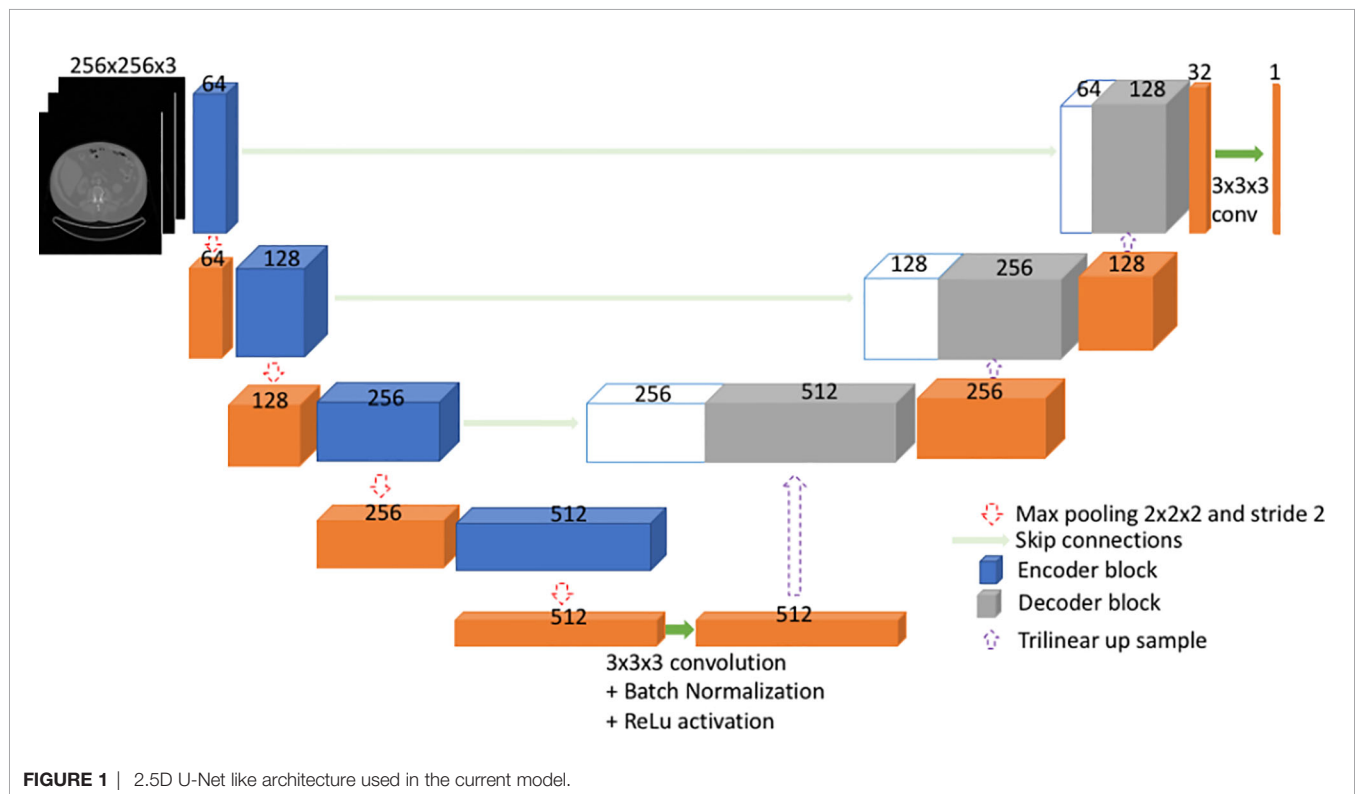
were included in the resultant segmentation. The threshold of 0.3 was chosen empirically based on inclusiveness and absolute percentage error; 0.5 was also compared but was too stringent and led to suboptimal results. The segmentation output of the model was compared with ground truth segmentation using Dice score, absolute percentage error (APE) in skeletal muscle area and SMI. Further, the model was compared with an existing 2D deep learning segmentation model (16). The AutoMATiCA model had been trained on diagnostic quality CT scans with (mean ± SD) tube current of 338 ± 123 mA, which was substantially higher than the tube current in the validation set images (mean ± SD) 158 ± 52 mA.

## RESULTS

### Inter-Observer Variation

Strong agreement between both observers was achieved with a mean ± SD Dice score 0.96 ± 0.02 and mean ± SD absolute

percentage difference in muscle area of 2.9 ± 2.5% for the 40 images with both observers' contours.

### Cross-Validation Performance on Cohort 1

The resultant network consists of a 2.5D U-Net trained using focal loss and Dice loss for 300 epochs. The training and validation performance of Cohort 1 in terms of Dice score and network loss for the five CVs are given in **Figure S3** and **Figure S4** respectively. Average performance during CV is given in **Figure 2**. For each CV, the results from best model on validation set are given in **Table 2** and **Supplementary Table 2**. The results in terms of Dice score and percentage/absolute percentage error between manual and automatic contours are given. All CV folds show similar performance.

### Ensemble Learning Outperforms Individual Learning

For each image in the test set, five different probability maps were predicted using the five models from 5-fold CV. Each image took
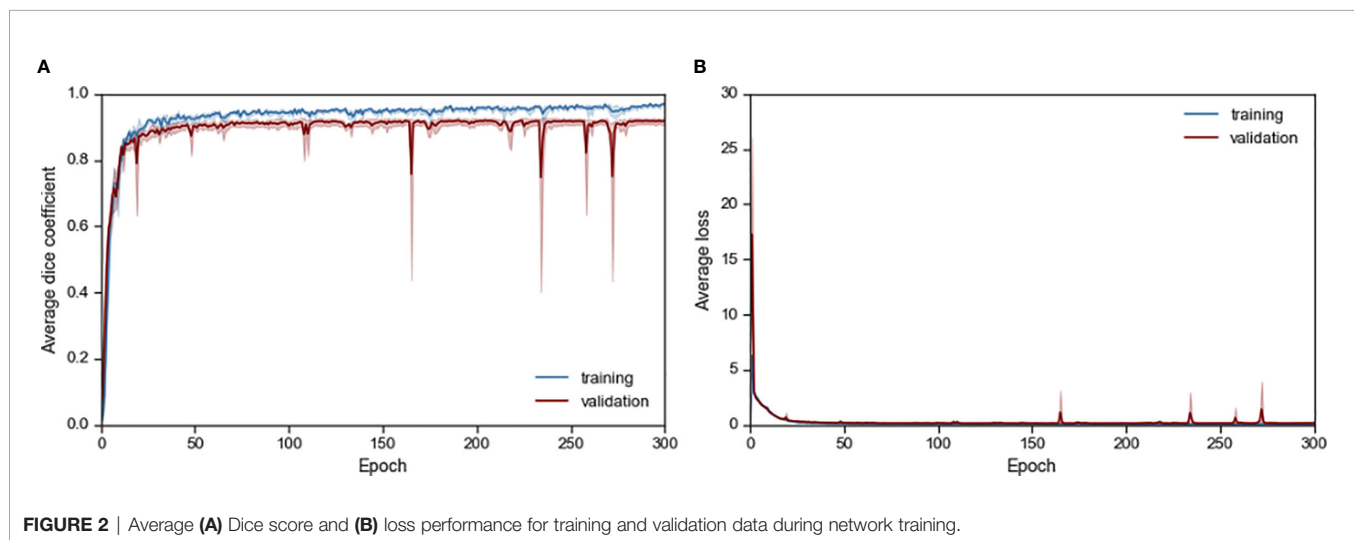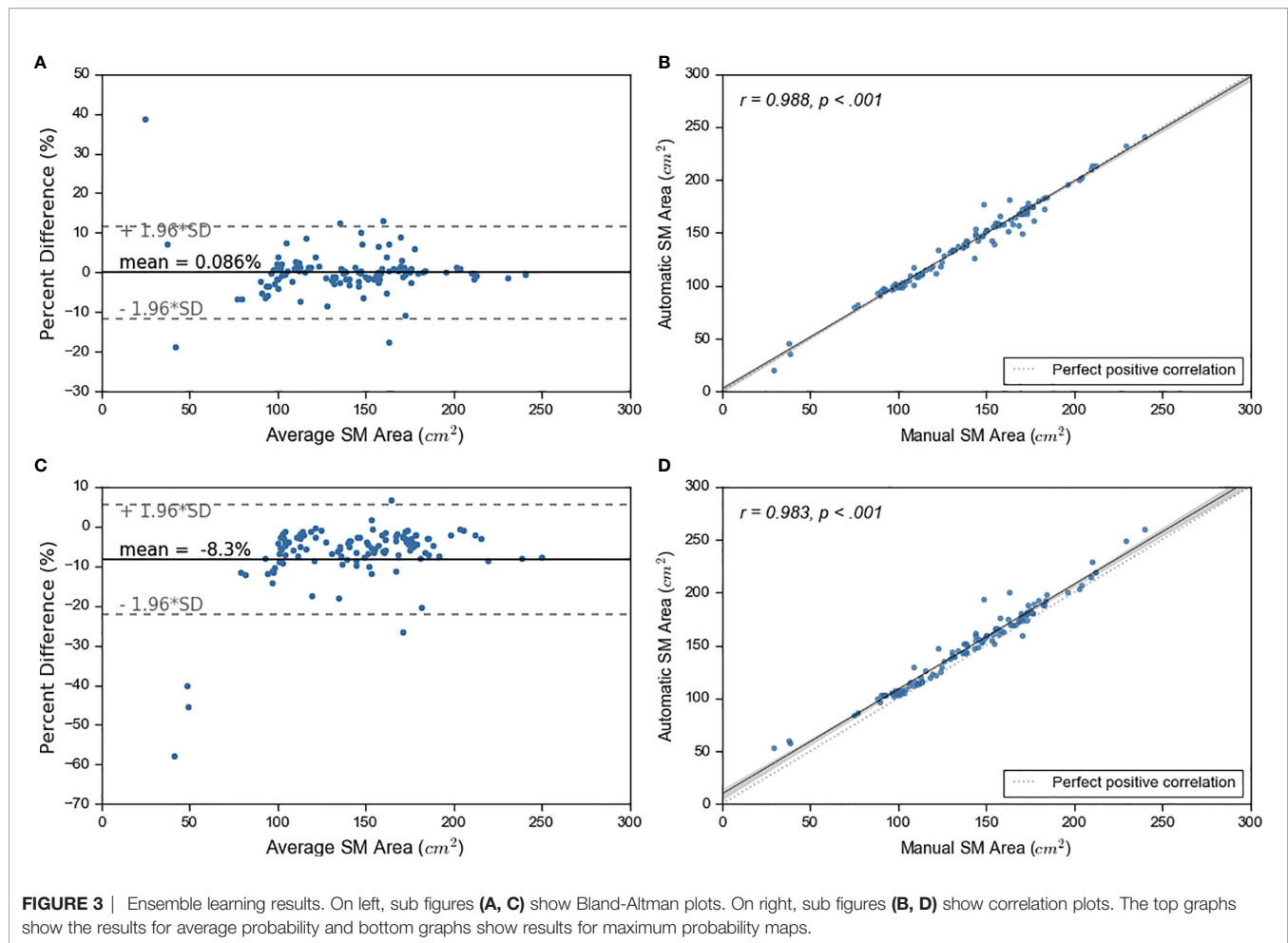


**FIGURE 2** | Average **(A)** Dice score and **(B)** loss performance for training and validation data during network training.

**TABLE 2** | Validation results for cross validation (CV) in terms of mean and standard deviation (SD) of performance measure.

| | CV 1 | CV 2 | CV 3 | CV 4 | CV 5 |
|---|---|---|---|---|---|
| **Training** | | | | | |
| **No. of patients** | 52 | 53 | 53 | 53 | 53 |
| **Total scans** | 112 | 118 | 122 | 117 | 119 |
| **Validation** | | | | | |
| **No. of patients** | 14 | 13 | 13 | 13 | 13 |
| **Total scans** | 35 | 29 | 25 | 30 | 28 |
| **Mean skeletal muscle area (cm²)** | | | | | |
| (mean ± SD) | 122.37 ± 34.84 | 138.75 ± 33.33 | 148.54 ± 34.34 | 141.15 ± 35.38 | 130.63 ± 43.44 |
| [Range] | 42.05 − 178.35 | 83.91 − 186.35 | 86.55 − 213.48 | 80.87 − 226.47 | 76.81 − 205.03 |
| **Dice score** | | | | | |
| (mean ± SD) | 0.90 ± 0.07 | 0.91 ± 0.06 | 0.93 ± 0.03 | 0.93 ± 0.02 | 0.90 ± 0.04 |
| **Percentage Error (PE) - %** | | | | | |
| (mean ± SD) | 0.09 ± 9.31 | -4.73 ± 4.11 | -2.28 ± 3.31 | -2.79 ± 4.22 | -6.81 ± 5.98 |
| **Absolute Percentage Error (APE) - %** | | | | | |
| (mean ± SD) | 4.67 ± 8.02 | 4.78 ± 4.04 | 3.50 ± 1.90 | 3.88 ± 3.21 | 6.81 ± 5.98 |

**FIGURE 3** | Ensemble learning results. On left, sub figures **(A, C)** show Bland-Altman plots. On right, sub figures **(B, D)** show correlation plots. The top graphs show the results for average probability and bottom graphs show results for maximum probability maps.

~0.4 seconds to go through the five models. Then, the final probability map for a test image was calculated by combining the probabilities of the five probability maps. Any pixel with probability greater or equal to 0.3 was classified as positive (i.e. belongs to skeletal muscle) and others as negative (i.e. outside of skeletal muscle).

We experimented with two approaches on how to combine the outputs from the five models: 1) taking the average probability and 2) taking the maximum probability. We compared the SM area from both approaches to the manual SM area (**Figure 3**). Superior performance was achieved by calculating average probability maps. Based on average probability, the skeletal muscle area for the test set ranged from 19.76 cm$^2$ to 241.04 cm$^2$ (with mean ± standard deviation of 138.88 cm$^2$ ± 38.27 cm$^2$). The majority of test cases (n=92, 79%), were within ±5% error between manual and automated contours. 57 (49%) cases showed an error within ± 1% (**Figure 3**, **Figure S5**).

## Accuracy of Segmentation of Skeletal Muscle Using the Deep Learning Model

The accuracy of the model was calculated by Dice score and difference in area between manual and automated contours. The

prediction accuracies are given in **Table 3** for 2D U-Net, single 2.5D U-Net model and the ensemble of 2.5D U-Net models (**Supplementary Table 3**). Improvements between the 2D U-Net and 2.5D U-Net was limited to classification of the liver-muscle interface; where there was no clear border between the liver and adjacent muscle on the L3 slice, the 2D U-Net was unable to accurately define the skeletal muscle. Use of the 2.5D U-Net improved this. The qualitative results are shown in **Figure 4**.

**TABLE 3** | Performance on test data set using 2.5D/2D U-Net.

| Performance measure | Median | Mean ± Std. dev. | Range |
|---|---|---|---|
| **Based on 5-fold cross validation (2.5D U-Net)** | | | |
| Dice score | 0.94 | 0.92 ± 0.06 | 0.57 – 0.97 |
| APE (%) | 1.46 | 3.09 ± 4.52 | 0.02 – 32.49 |
| **Based on single model with 0.1 validation set (2.5D U-Net)** | | | |
| Dice score | 0.94 | 0.92 ± 0.07 | 0.41 – 0.97 |
| APE (%) | 2.80 | 4.86 ± 7.62 | 0.04 – 54.28 |
| **Based on single model with 0.1 validation set (2D U-Net)** | | | |
| Dice score | 0.94 | 0.92 ± 0.05 | 0.57 – 0.97 |
| APE (%) | 3.82 | 5.16 ± 5.64 | 0.03 – 46.55 |
| **AutoMATiCA results** | | | |
| APE (%) | 4.0 | 4.98 ± 6.78 | 0.20 – 53.68 |
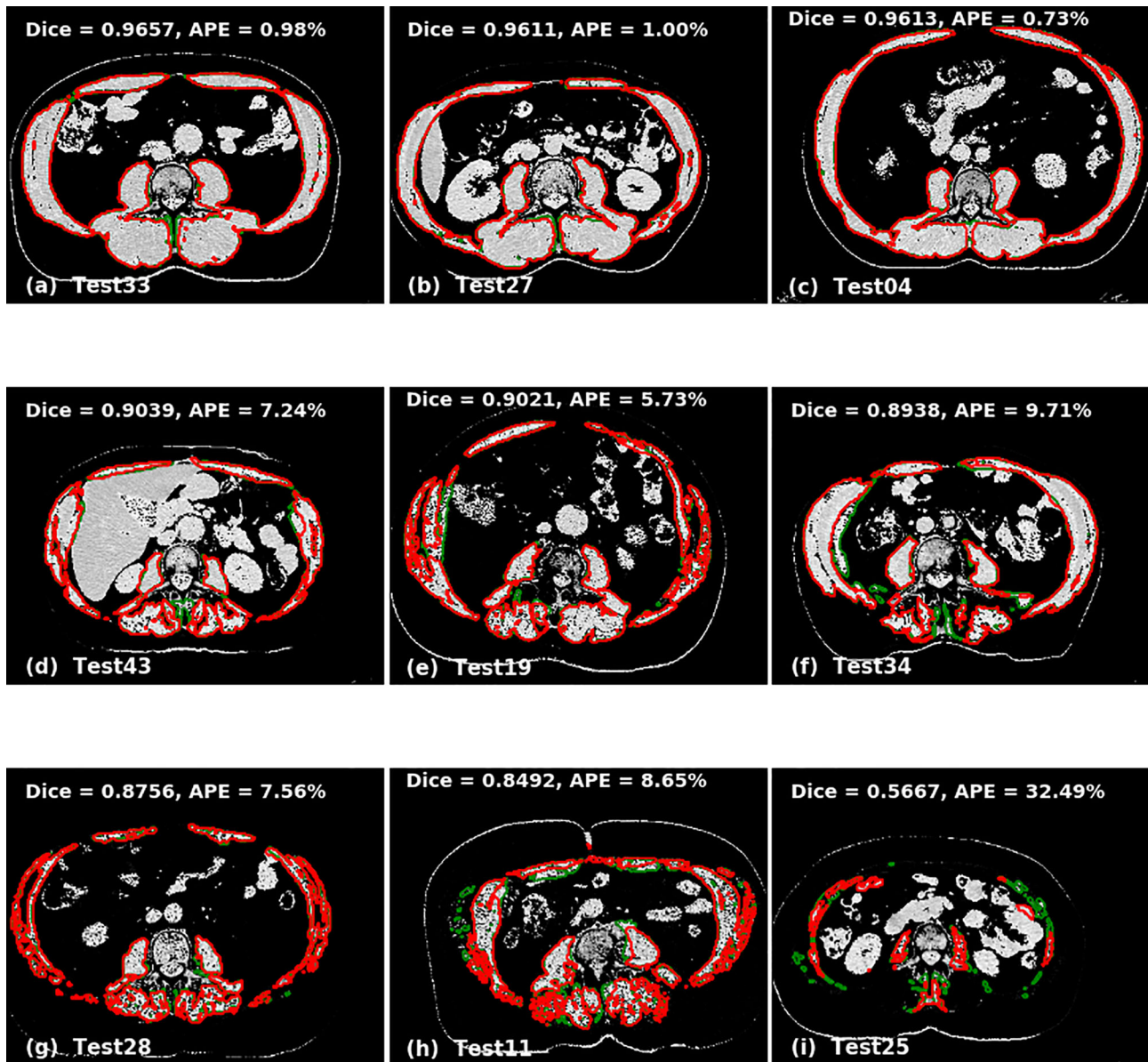
**FIGURE 4** | Qualitative performance of the model on Cohort 2. Red represents deep learning contours and green represents manual contours. Panels **(A–C)** in the top panel show three of the best performing cases (Dice ± 0.96). Panels **(D–F)** in the middle panel show the cases with average performance (Dice ± 0.90) and Panels **(G–I)** in the bottom panel show cases with lowest performance (Dice ± 0.88).

Top, middle and bottom rows show representative contours from best performing, less than average performing and least performing cases respectively. **Figure 4I** shows a scan with a very low skeletal muscle area compared to all other images in training and testing sets, and this resulted in a very low Dice score. For the 20 scans in the test data set that had contours from both observers, the inter-observer mean ± SD of the dice scores was 0.96 ± 0.02 and the mean ± SD of the absolute percentage difference in muscle area was 2.7 ± 2.5%. For the same 20 scans, the mean ± SD of the dice scores

was 0.93 ± 0.03 and the mean ± SD of the absolute percentage error in muscle area was 3.8 ± 2.8%, showing slightly inferior performance compared with inter-observer variation.

The AutoMATiCA model was applied to all images in the validation image set. The Dice score was not able to be computed as AutoMATiCA does not export segmentation, only the muscle area measurement and a merged image file. The mean ± SD APE for AutoMATiCA was 5.0 ± 6.8%, compared with 3.1 ± 4.5% for the current model.

## Using Model Output to Predict Sarcopenia

Skeletal muscle index (SMI) at L3, which is calculated by dividing skeletal muscle area by the height of patient squared, is a well-known surrogate for sarcopenia in cancer. In the test set, 35 patients out of 42 had recorded height and weight information, which we used to calculate SMI on each of their images. Then, we compared the SMI values based on manual and automated SM area for these patients (**Figure S6**, **Supplementary Table 3**). Finally, we use these SMI values to classify scans into sarcopenic and non-sarcopenic groups (**Figure 5**). Sarcopenic patients were classified based on reference values from Martin et al. (32). Sarcopenia was defined as SMI $< 43$ cm$^2$/m$^2$ in men with a body mass index (BMI) $< 24.9$ kg/m$^2$ and $< 53$ cm$^2$/m$^2$ in men with a BMI $> 25$ kg/m$^2$; and $< 41$ cm$^2$/m$^2$ in women of any BMI. Out of 94 CT scans in the validation set for which we had the required clinical data, 85 were correctly classified as sarcopenic or not by automatic SM contours. The positive and negative predictive values shown in **Figure 5** resulted in a sensitivity of 84% and a specificity of 95%.

## DISCUSSION

In this study, we present an ensemble model of 2.5D CNNs to automatically segment skeletal muscle on low quality CTs acquired in PET/CT studies, and investigate its qualitative and quantitative accuracy and precision to measure SM area and detect sarcopenia in NSCLC patients. It is widely recognized that robust measures of skeletal muscle mass are often challenging to use in clinical practice due to cost, time and the training required (33). This results in a tendency to use subjective assessments which have demonstrated inaccuracy in identifying sarcopenia (34).



**FIGURE 5** | Sarcopenia classification based on manual and automatic SM area.

The variation between the model and manual segmentation was similar to that measured between multiple observers; Perthen et al. (35) quantified inter-observer variation for L3 skeletal muscle area as delineated by three radiologists. The mean absolute difference between any two observers was up to 2.69 cm$^2$. In our data set we achieved a mean absolute difference between two observers of 3.55 cm$^2$ and a mean absolute difference between the manual and automated contours of 3.69 cm$^2$. Further, the mean Dice score between our two observers was 0.96, compared with the mean dice score for our automated contours with manual contours of 0.92. We achieved a sensitivity of 84% and specificity of 95% when classifying sarcopenic patients using automatic contours and using Alberta protocol based manual diagnosis as ground truths. These results indicate our model has the potential to facilitate large scale robust assessment of skeletal muscle from low quality CT scans in the research setting, as well as clinical practice to support early identification and intervention.

Despite promising results, our study has several limitations. In some cases, with very low SM area, the model tends to misclassify other organs as belonging to skeletal muscle. These results suggest that the CNN has not been trained with images that fully represent the diversity and heterogeneity of SM area. To potentially overcome this problem, the proposed model can be retrained with new images as they are being acquired. We also observed limited benefit of data augmentation apart from flipping and addition of Gaussian noise, which may suggest limited variability in the validation set. Further improvement for use in external data sets may be achieved with increased variability in the training image acquisition and reconstruction parameters, and inclusion of images from a wider range of institutions. However, we observed that these limitations did not impact the ability to provide correct sarcopenia classification. Further, these can be improved by incorporating user interaction to correct mislabeled sections of CT. A further systematic difference between the manual and automated segmentation occurred when the patient was scanned with arms down; the model mis-classified portions of the arm at the L3 level, as there were no patients in the training set who were scanned with arms down.

Our approach is only trained and validated on attenuation correction quality CTs, specifically those that were obtained as part of PET/CT studies since these will typically contain L3 in the scan range. Our model shows improved results on our validation data set compared with a 2D AI model trained specifically on higher quality diagnostic CT scans. This suggests a domain specific training is likely required for widespread applicability of such models. Further, our model was a 2.5D model, which may provide further improvements over the 2D model, in particular with organ-muscle interfaces that may not be visible in the selected slice for analysis. Specific image normalization methods and model parameter tuning are needed to extend our method to other modalities, including diagnostic quality CTs and magnetic resonance imaging (MRI). Potential improvement may be achieved through use of higher quality diagnostic CT scans.

Deep learning-based methods are showing the potential to reliably automate a number of rudimentary pattern recognition
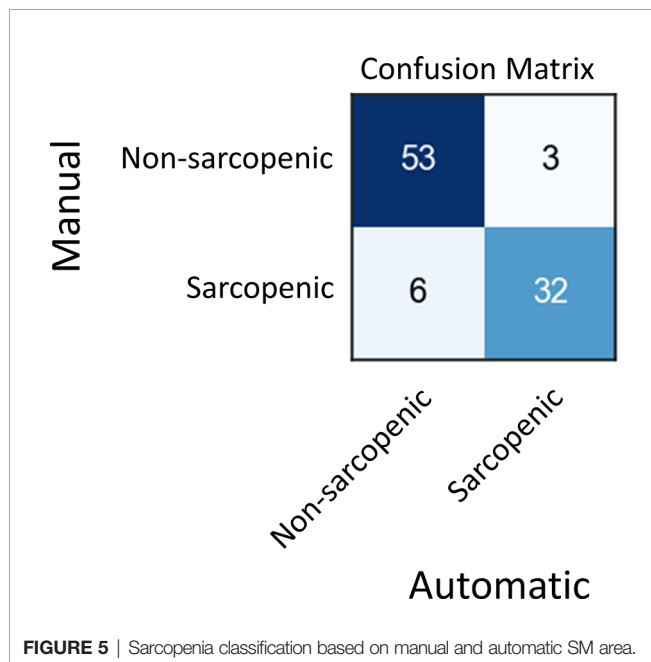
tasks. If coupled with other methods to localize to the appropriate L3 slice (36), there is a pathway to fully automate these measures for any patient receiving CT imaging. It is foreseeable that being able to track trends in body composition would have implications in management in a number of chronic diseases, which are regularly monitored through volumetric CT or MR imaging.

## CONCLUSION

We present an automated method to delineate skeletal muscle area at L3 region of attenuation correction CT scans acquired as part of PET/CT studies for patients with NSCLC. The proposed method can be used to classify sarcopenia with minimal manual intervention, which may be an efficient method in large studies. Further, the model can be potentially used in clinical practice to identify early sarcopenia in patients with lung cancer.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Peter MacCallum Cancer Centre. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.580806/full#supplementary-material

## REFERENCES

1. Fearon K, Strasser F, Anker SD, Bosaeus I, Bruera E, Fainsinger RL, et al. Definition and classification of cancer cachexia: An international consensus. *Lancet Oncol Lancet Oncol* (2011) 12:489–95. doi: 10.1016/S1470-2045(10)70218-7

2. Cederholm T, Jensen GL, Correia MITD, Gonzalez MC, Fukushima R, Higashiguchi T, et al. GLIM criteria for the diagnosis of malnutrition – A consensus report from the global clinical nutrition community. *Clin Nutr* (2019) 38(1):1–9. doi: 10.1016/j.clnu.2019.02.033

3. Moses AWG, Slater C, Preston T, Barber MD, Fearon KCH. Reduced total energy expenditure and physical activity in cachectic patients with pancreatic cancer can be modulated by an energy and protein dense oral supplement enriched with n-3 fatty acids. *Br J Cancer* (2004) 90(5):996–1002. doi: 10.1038/sj.bjc.6601620

4. Arends J, Bachmann P, Baracos V, Barthelemy N, Bertz H, Bozzetti F, et al. ESPEN Guideline ESPEN guidelines on nutrition in cancer patients *. *Clin Nutr* (2016) 4:11–48. doi: 10.1016/j.clnu.2016.07.015

5. Bril SI, Wendrich AW, Swartz JE, Wegner I, Pameijer F, Smid EJ, et al. Interobserver agreement of skeletal muscle mass measurement on head and neck CT imaging at the level of the third cervical vertebra. *Eur Arch Oto-Rhino-Laryngol* (2019) 276(3):1175–82. doi: 10.1007/s00405-019-05307-w

6. van Vugt JLA, Levolger S, de Bruin RWF, van Rosmalen J, Metselaar HJ, IJzermans JNM. Systematic Review and Meta-Analysis of the Impact of Computed Tomography-Assessed Skeletal Muscle Mass on Outcome in Patients Awaiting or Undergoing Liver Transplantation. *Am J Transplant* (2016) 16(8):2277–92. doi: 10.1111/ajt.13732

7. Levolger S, Van Vugt JLA, De Bruin RWF, IJzermans JNM. Systematic review of sarcopenia in patients operated on for gastrointestinal and hepatopancreatobiliary malignancies. *Br J Surg* (2015) 102:1448–58. doi: 10.1002/bjs.9893

8. Miyamoto Y, Baba Y, Sakamoto Y, Ohuchi M, Tokunaga R, Kurashige J, et al. Sarcopenia is a Negative Prognostic Factor After Curative Resection of Colorectal Cancer. *Ann Surg Oncol* (2015) 22(8):2663–8. doi: 10.1245/s10434-014-4281-6

9. Prado CMM, Lieff JR, Mccargar LJ, Reiman T, Sawyer MB, Martin L, et al. Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: a population-based study. *Lancet Onco* (2008) 9(7):629–35. http://oncology.thelancet.com.

10. Mourtzakis M, Prado CMM, Lieffers JR, Reiman T, McCargar LJ, Baracos VE. A practical and precise approach to quantification of body composition in cancer patients using computed tomography images acquired during routine care. *Appl Physiol Nutr Metab* (2008) 33(5):997–1006. doi: 10.1139/H08-075

11. Shen W, Punyanitya M, Wang Z, Gallagher D, St-Onge MP, Albu J, et al. Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image. *J Appl Physiol [Internet]* (2004) 97(6):2333–8. doi: 10.1152/japplphysiol.00744.2004

12. Prado CMM, Baracos VE, McCargar LJ, Reiman T, Mourtzakis M, Tonkin K, et al. Sarcopenia as a Determinant of Chemotherapy Toxicity and Time to Tumor Progression in Metastatic Breast Cancer Patients Receiving Capecitabine Treatment. *Clin Cancer Res [Internet]* (2009) 15(8):2920–6. doi: 10.1158/1078-0432.CCR-08-2242

13. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyère O, Cederholm T, et al. Sarcopenia: Revised European consensus on definition and diagnosis. *Age Ageing* (2019) 48(1):16–31. doi: 10.1093/ageing/afz046

14. Cardenas CE, McCarroll RE, Court LE, Elgohari BA, Elhalawani H, Fuller CD, et al. Deep Learning Algorithm for Auto-Delineation of High-Risk Oropharyngeal Clinical Target Volumes With Built-In Dice Similarity Coefficient Parameter Optimization Function. *Int J Radiat Oncol Biol Phys [Internet]* (2018) 101(2):468–78. doi: 10.1016/j.ijrobp.2018.01.114

15. Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother Oncol* (2018) 126(2):312–7. doi: 10.1016/j.radonc.2017.11.012

16. Paris MT, Tandon P, Heyland DK, Furberg H, Premji T, Low G, et al. Automated body composition analysis of clinically acquired computed

tomography scans using neural networks. *Clin Nutri* (2020) 39(10):3049–55. doi: 10.1016/j.clnu.2020.01.008

17. Dabiri S, Popuri K, Cespedes Feliciano EM, Caan BJ, Baracos VE, Faisal Beg M. Muscle segmentation in axial computed tomography (CT) images at the lumbar (L3) and thoracic (T4) levels for body composition analysis. *Comput Med Imaging Graph* (2019) 75:47–55. doi: 10.1016/j.compmedimag.2019.04.007

18. Burns JE, Yao J, Chalhoub D, Chen JJ, Summers RM. A Machine Learning Algorithm to Estimate Sarcopenia on Abdominal CT. *Acad Radiol* (2020) 27 (3):311–20. doi: 10.1016/j.acra.2019.03.011

19. Podoloff DA, Advani RH, Allred C, Benson Iii B, Brown E, Burstein HJ, et al. JNCCN NCCN Task Force Report: PET/CT Scanning in Cancer. *J Natl Compr Cancer Netw* (2007) 5(Supplement 1).

20. Yararbas U, Avci NC, Yeniay L, Argon AM. The value of 18F-FDG PET/CT imaging in breast cancer staging. *Bosn J Basic Med Sci* (2018) 18(1):72–9. doi: 10.17305/bjbms.2017.2179

21. Silvestri GA, Gonzalez AV, Jantz MA, Margolis ML, Gould MK, Tanoue LT, et al. Diagnosis and management of lung cancer, 3rd ed: accp guidelines. *Chest* (2013) 143:e211S–50S. doi: 10.1378/chest.12-2355

22. Wong TZ, Paulson EK, Nelson RC, Patz EF, Coleman RE. Practical Approach to Diagnostic CT Combined with PET. *Am J Roentgenol* (2007) 188(3):622–9. doi: 10.2214/AJR.06.0813

23. Blodgett TM, Casagranda B, Townsend DW, Meltzer CC. Issues, controversies, and clinical utility of combined PET/CT imaging: what is the interpreting physician facing? *AJR Am J Roentgenol AJR Am J Roentgenol* (2005) 184. doi: 10.2214/ajr.184.5_supplement.0184s138

24. Siva S, Callahan J, Kron T, Martin OA, MacManus MP, Ball DL, et al. A prospective observational study of Gallium-68 ventilation and perfusion PET/CT during and after radiotherapy in patients with non-small cell lung cancer. *BMC Cancer [Internet]* (2014) 14(1):740. doi: 10.1186/1471-2407-14-740

25. Everitt S, Ball D, Hicks RJ, Callahan J, Plumridge N, Trinh J, et al. Prospective Study of Serial Imaging Comparing Fluorodeoxyglucose Positron Emission Tomography (PET) and Fluorothymidine PET During Radical Chemoradiation for Non-Small Cell Lung Cancer: Reduction of Detectable Proliferation Associated With Worse Survival. *Int J Radiat Oncol Biol Phys* (2017) 99(4):947–55. doi: 10.1016/j.ijrobp.2017.07.035

26. Kiss N, Beraldo J, Everitt S. Early Skeletal Muscle Loss in Non-Small Cell Lung Cancer Patients Receiving Chemoradiation and Relationship to Survival. *Support Care Cancer* (2019) 27(7):2657–64. doi: 10.1007/s00520-018-4563-9

27. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *Lecture Notes in Computer Science*. Munich: Springer (2015). p. 234–41.

28. Milletari F, Navab N, Ahmadi S-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: *Proc - 2016 4th Int Conf 3D Vision, 3DV 2016*, Institute of Electrical and Electronics Engineers Inc. (2016). p. 565–71. Available at: http://arxiv.org/abs/1606.04797.

29. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. *IEEE Trans Pattern Anal Mach Intell* (2017) 42(2):318–27. doi: 10.1109/TPAMI.2018.2858826

30. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer (2016). p. 424–32.

31. Kingma DP, Ba JL. (2015). Adam: A method for stochastic optimization3rd International Conference on Learning Representations, ICLR 2015, in: *International Conference on Learning Representations, ICLR*, Conference Track Proceedings. Available at: https://arxiv.org/abs/1412.6980v9.

32. Martin L, Birdsell L, MacDonald N, Reiman T, Clandinin MT, McCargar LJ, et al. Cancer cachexia in the age of obesity: skeletal muscle depletion is a powerful prognostic factor, independent of body mass index. *J Clin Oncol* (2013) 31(12):1539–47. doi: 10.1200/JCO.2012.45.2722

33. Price KL, Earthman CP. Update on body composition tools in clinical settings: computed tomography, ultrasound, and bioimpedance applications for assessment and monitoring. *Eur J Clin Nutr* (2019) 73:187–93. Nature Publishing Group. doi: 10.1038/s41430-018-0360-2

34. Sheean PM, Peterson SJ, Gomez Perez S, Troy KL, Patel A, Sclamberg JS, et al. The prevalence of sarcopenia in patients with respiratory failure classified as normally nourished using computed tomography and subjective global assessment. *J Parenter Enter Nutr* (2014) 38(7):873–9. doi: 10.1177/0148607113500308

35. Perthen JE, Ali T, McCulloch D, Navidi M, Phillips AW, Sinclair RCF, et al. Intra- and interobserver variability in skeletal muscle measurements using computed tomography images. *Eur J Radiol* (2018) 109:142–6. doi: 10.1016/j.ejrad.2018.10.031

36. Chen Y, Gao Y, Li K, Zhao L, Zhao J. Vertebrae identification and localization utilizing fully convolutional networks and a hidden markov model. *IEEE Trans Med Imaging* (2020) 39(2):387–99. doi: 10.1109/TMI.2019.2927289