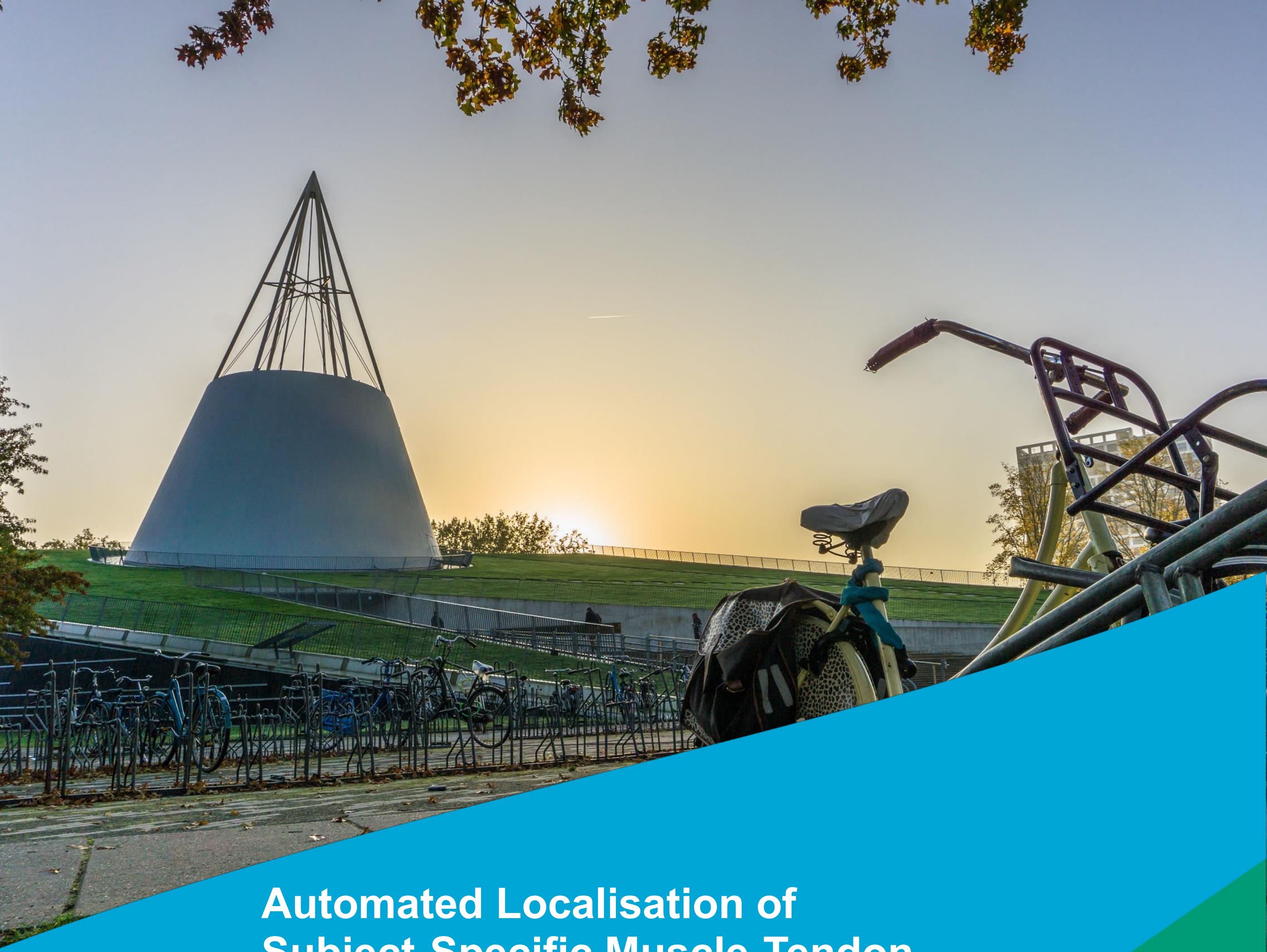

Automated Localisation of Subject-Specific Muscle-Tendon Paths of the Lower Limbs Using nnU-Net on Magnetic Resonance Images

Author: Christiaan Andrès Gabriël van Straaten

Date: 30/01/2025

Automated Localisation of Subject-Specific Muscle-Tendon Paths of the Lower Limbs Using nnU-Net on Magnetic Resonance Images

By

Christiaan A.G. van Straaten

In partial fulfilment of the requirements for the degree of:

Master of Science
in Biomedical Engineering
Track Neuromusculoskeletal Biomechanics

at the Delft University of Technology,
to be defended publicly on [Thursday February 6, 2025 at 10:45 AM.]

Thesis committee:	Prof. dr. ir. J. Harlaar,	TU Delft & Erasmus MC, Supervisor, Chair
	Dr. M.G.H. Wesseling,	TU Delft & Erasmus MC, Daily Supervisor
	Dr. J. Hirvasniemi,	TU Delft & Erasmus MC, Daily Supervisor
	Dr. ir. E. van der Kruk,	TU Delft

Automated Localisation of Subject-Specific Muscle-Tendon Paths of the Lower Limbs Using nnU-Net on Magnetic Resonance Images

Christiaan A.G. van Straaten (4728408)

January 30, 2025

Abstract

Muscle-tendon paths vary between individuals, and musculoskeletal models show sensitivity to these variations when estimating muscle and joint forces. Accurate estimations of these internal forces allow for a more comprehensive approach to researching debilitating musculoskeletal pathologies such as osteoarthritis. However, defining subject-specific muscle-tendon paths is labour-intensive and requires expert knowledge, which is not as repeatable. Therefore, in this study, the accuracy of determining subject-specific muscle points and volumes based on lower limb Magnetic Resonance (MR) scans with the nnU-net is evaluated. Two models are trained using the open-source pipeline, referred to as the point and volume model. The volume model aims to segment muscle-tendon volumes and is trained on the open-source augmented dataset of Henson et al. (2023). The point model localises attachment and via points describing muscle action lines for a subset of relevant muscles of the volume model. Since U-net is not designed for predicting points, a workaround is introduced by creating cubes around the points as labels. The 3D volume model scored a median Dice Similarity Coefficient of 92.7 % and shows some generalisation capability. The 3D point model scored a median Euclidean error of 5.1 mm. Compared to intra-operator variability for attachment points, this approach yields lower and more repeatable accuracy without required manual intervention. The nnU-net pipeline is capable of producing accurate models that can define subject-specific muscle-tendon paths based on MR scans.

1 Introduction

Muscle and joint reaction forces are critical measures for understanding both normal and pathological human movement, as they reflect the internal loading of the musculoskeletal system. The only currently accessible method of determining these in vivo is by using an inverse or forward approach in combination with a musculoskeletal model (MSM) and an optimisation algorithm. The MSM is required to detail geometric and dynamic relations of muscles and bones, and the optimisation is used to solve muscle and/or trajectory redundancy problems. In research, these MSMs are predominately used in gait investigations. Accurate estimations of these forces are valuable for understanding joint load-related pathologies such as osteoarthritis (OA) [1]. Despite their potential, MSMs remain underutilised in clinical practice [2, 3].

Before MSMs can be readily adopted in the clinic, muscle-tendon and skeletal parameters of the model need to be set appropriately, then, these models are ready to be verified and validated to ensure accurate estimations [3]. Subject-specific models aim to determine model parameters that most accurately reflect the subject's mechanical properties. Whereas in certain research settings proof of concept could be sufficient [4], in clinical scenarios, it is often desirable to have biomechanical estimations of individual patients. For example, in a group of knee OA patients, the most effective gait alteration for reducing knee load varied depending on whether the model was personalised [5]. Furthermore, including subject-specificity in a gait simulation has been shown to result in an average difference in the second hip contact force peak of 47% of body weight compared to a generic MSM [6].

Evidence shows that muscle-tendon paths vary per individual [7, 8] and estimations using MSMs such as muscle [8, 9, 10] and joint forces [11] are considerably sensitive to changes in these parameters. An anatomical study of the femur observed that muscle-tendon attachment sites of the gluteus maximus and rectus femoris vary in such a way that their resulting moment arms around the joint have a standard deviation of 65% [7]. Another study investigated the effect of natural anatomic variability of muscle paths on muscle forces during gait [8]. The authors noted that 10 muscle attachments of a generic musculoskeletal model [12] fell outside of the anatomic variability. Perturbing the muscle paths with natural amplitude does affect muscle forces significantly. Especially, perturbing the psoas resulted in large changes in muscle force ranging up to 230 N. Others also investigating gait found that perturbations of 1 cm in muscle paths result in significant muscle force differences [9]. Sensitivity to muscle path parameters is not exclusive to lower limb models; perturbing the quadratus lumborum attachment site in the Twente Spine Model can alter the disc shear forces by 353% [11].

In the current state-of-the-art multi-body MSMs, which are typically used for gait investigations, muscle forces are conceptualised as bound vectors [13, 8, 6]. Such a conceptualisation is mechanically justified when using enough bound vectors located at the centroid of different directional muscle fibre groups of the muscle-tendon [14]. The paths of muscle-tendons are typically mapped out by a chain of straight lines where the attachment points define the origin and insertion and the via points define the trajectory in between these points. Via points are used to model the muscle-tendon unit when it wraps around a bony contour or when the fascia restricts its path. This approach will be referred to as the poly-line approach. These bound vectors act about a point which is typically chosen as the rotational point of the joint [13]. The resulting moment can be calculated by crossing the vector from the rotational point to any point on the line of action of the bound vector with the bound vector [15]. From this definition, it is evident that both the muscle attachment location and direction of the muscle-tendon at the location of attachment are crucial to accurately calculate muscle joint torque.

New ways of modelling muscle-tendons have been proposed that require fewer assumptions than the poly-line approach. These will be referred to as the complex modelling approach. Blemker and Delp introduced a finite element model of a muscle-tendon [16]. They simulated geometrically more complex muscle-tendons such as the gluteus maximus, gluteus medius, psoas, and iliacus in hip movement around each anatomical axis. Complex geometry was achieved by mapping different pennation types to volumetric muscle-tendon data derived from Magnetic Resonance (MR) scans. Unfortunately, the computational expense of the method withholds it from being simulations that are already computationally expensive, such as a static optimisation using inverse kinematics of roughly 90 muscle-tendons, which is typical for gait investigations in the clinic. Recently, a volume-based model was proposed that addresses the same issues only with less computational expense. The model automatically creates a predefined number of fibres and a predefined number of straight lines to model each fibre based on muscle-tendon volumes and attachment sites [17]. For the same simulation as in Blemker and Delp's experiment [16], it took less than a minute to compute the simulation compared to, at the time, 5 to 10 CPU hours. Both these methods require volumetric data of the muscle-tendons to overcome the muscle-tendon modelling limitations of the poly-line approach.

Independent of which muscle-tendon modelling approach is used, a fully automated muscle point and volume identification technique could lower or eliminate the required labour and be more repeatable than human expertise. Currently, the gold standard for defining muscle points or muscle volumes is to let an expert annotate an MR scan manually. It takes an expert three to four hours to define muscle points of a multi-body lower limb model with 34 muscles per leg [18]. Experts showed an intra- and inter-rater variability of 6.9 and 5.6 mm while correcting predicted attachment points, respectively [18]. Another study showed that inter-operator variability can cause up to 64% of deviation in peak muscle force in pathological ankle and foot simulations [10]. In the case of semantic segmentation, i.e. determining muscle-tendon volumes, it can take up to 40 hours to manually define 18 muscle-tendons per subject bilaterally [19]. Ground truths in this field also suffer from repeatability issues. For instance, out of the 35 lower limb muscle segmentations defined by a single operator, only 23 were included because they exhibited less than 10% variation in volume across all three runs [20].

In the literature, various automated methods have been proposed that can approximate muscle-tendon paths (Figure 1). There are three main muscle-tendon path descriptors that are produced by these methods: muscle-tendon attachments defined by points, lines, and/or planes, muscle-tendon centroid paths often defined by a series of points connected with line segments (poly-line approach), and muscle-tendon masks which are always defined as point clouds. Typically, each muscle-tendon approximation corresponds with a specific way of modelling muscle-tendons geometrically. Scaling based on palpable bony landmarks is generally viewed as inaccurate [21, 22, 23] and requires no imaging modalities. Wesseling et al. (2019) found that compared to scaling based on palpable bony landmarks, non-rigid deformation based on bone shape reduced the average error in muscle points by 21%, resulting in an Euclidean error of 17.3 mm [23]. Only Scheys et al. (2009) used an approach that could be classified as a segmentation technique to estimate muscle points [18]. By applying a non-rigid registration algorithm to an MR atlas with muscle points to a new MR image, it achieved a median Euclidean error of 6.1 mm. This is roughly 65% less error than the non-rigid deformation based on bone shape method [23]. The segmentation-like approach [18] is considered to be superior compared to scaling based on bone shape [23]. Both techniques rely on a labelled atlas, so their accuracy may be influenced by how well the atlas matches the subject's characteristics.

Segmentation approaches can be divided into segmentation with explicit algorithms and segmentation using supervised deep learning (Figure 1). Supervised deep learning methods have been shown to achieve a Dice Similarity Coefficient (DSC) that is 4.5% higher than that of a multi-atlas registration approach, which is considered the state-of-the-art among explicit algorithms. [24]. There are many deep learning models capable of segmenting muscle volumes, but none are open-source and able to segment muscles around the hip, knee and ankle [25, 26, 27]. Recently, Henson et al. created a publicly available augmented dataset of lower limb muscles by using their non-linear deformable image registration model [20]. By augmenting the images, it was shown that the range of muscle volumes was increased. As the authors suggested, this makes for a good opportunity to train a deep learning model to segment lower limb muscles. Finally, a recent study demonstrated that a segmentation deep learning model (U-net) could be adapted for point prediction with high accuracy on X-ray scans [28]. Possibly, a similar conversion could be done to predict muscle points on MR images.

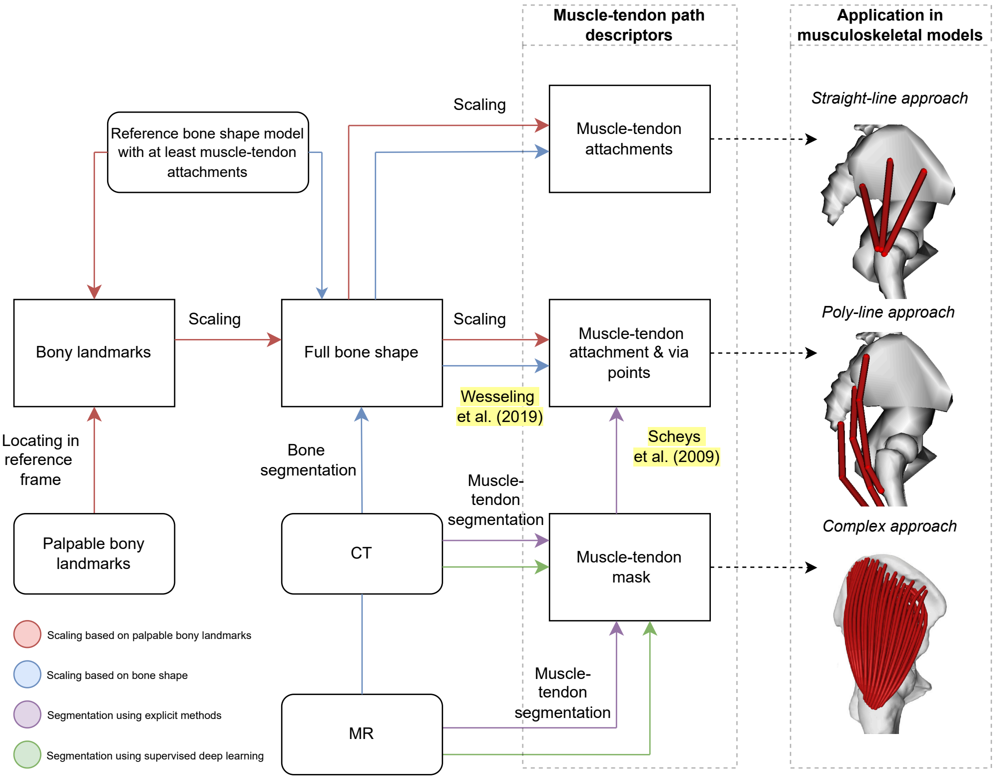
Muscle-tendon path descriptors
Application in musculoskeletal models
Straight-line approach
Poly-line approach
Complex approach
Reference bone shape model with at least muscle-tendon attachments
Scaling
Muscle-tendon attachments
Bony landmarks
Scaling
Full bone shape
Scaling
Muscle-tendon attachment & via points
Wesseling et al. (2019)
Scheys et al. (2009)
Locating in reference frame
Bone segmentation
Muscle-tendon segmentation
Palpable bony landmarks
CT
Muscle-tendon mask
MR
Muscle-tendon segmentation
Scaling based on palpable bony landmarks
Scaling based on bone shape
Segmentation using explicit methods
Segmentation using supervised deep learning

Figure 1: Overview of flow of information of the four groups of techniques to identify subject-specific muscle-tendon paths in the literature. The rectangles with sharp edges represent anatomical properties. The rectangles with rounded edges represent inputs. The arrows show the flow of information and the colours of the arrows indicate which categorical technique is used, where red, blue, purple, and green indicate scaling based on bony landmarks, scaling based on bone shape, muscle segmentation with explicit algorithms, and muscle segmentation with supervised deep learning, respectively. Views of straight- and poly-line approaches are screenshots of a gait model created in OpenSim and used under the Apache 2.0 License. See reference [29] for full details. The complex approach view is adapted from Figure 2 in Modenese and Kohout, Automated Generation of Three-Dimensional Complex Muscle Geometries for Use in Personalised Musculoskeletal Models, licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). See reference [17] for full details.

In the domain of biomedical semantic image segmentation, the U-net architecture is currently considered to be state-of-the-art. It was introduced by Ronneberger et al. during the International Symposium on Biomedical Imaging cell tracking challenge in 2015 [30]. In terms of the union of intersections, the model achieved a score of 92%, compared to the second-best score of 83%. Realising high performance using the U-net is, however, not straightforward. It requires expert knowledge to correctly implement and train the model. Key decisions include choices about pre-processing, architecture configuration, post-processing, hyperparameter initialisation, and data augmentation, all of which depend on the dataset and available computational resources. To address this limitation, the nnU-Net, a self-configuring pipeline, was introduced [31]. By analysing the 'fingerprint' of a raw dataset, rule-based decision making, and empirically derived fixed parameters, it achieves state-of-the-art results, outperforming other models on 23 public datasets.

nnU-Net appears to be a promising tool for both muscle point prediction and muscle volume segmentation. This study aims to develop and evaluate automated methods for accurately identifying muscle points and volumes by using the nnU-Net. Two models are examined: the ’point model’, which identifies muscle points to directly define lines of action, and the ’volume model’, which segments muscle-tendon volumes. The central question addressed is: How accurately can muscle points and volumes be defined by nnU-net based on MR scans of the lower limbs?

2 Methods

2.1 Datasets

2.1.1 Point Model

The dataset comprised 12 bilateral lower limbs MR imaging scans from a healthy control group (n = 7) and a hip OA patient group (n = 5). Muscle points were defined by an expert annotator. The MR dataset used in this study has been previously employed in [32, 6, 33]. Among the control group, four participants were aged between 45 and 60 years, and two participants were between 20 and 30 years. The mean age of the OA group was 54 ± 8.6 [32]. Scans of the hip, femur, knee, and tibia scans were made per participant. The voxel sizes (coronal × sagittal × and axial) of the hip, femur, knee, and tibia were 0.93 mm × 1.00 mm × 0.93 mm, 0.94 mm × 2.00 mm × 0.94 mm, 0.93 mm × 1.00 mm × 0.93 mm, and 0.93 mm × 1.00 mm × 0.93 mm, respectively. The only exceptions were found in the scans of control participants C2, C3, and C4. For C2, the hip scan had voxel sizes of 0.88 mm × 1.00 mm × 0.88 mm, and the knee scan had the same voxel sizes of 0.88 mm × 1.00 mm × 0.88 mm. In contrast, the knee scans for C3 and C4 had voxel sizes of 0.84 mm × 1.00 mm × 0.84 mm. For the tibia scans, the voxel sizes were 0.91 mm × 1.00 mm × 0.91 mm for C2, 0.84 × 1.00 mm × 0.84 mm for C3, and 0.88 mm × 1.00 mm × 0.88 mm for C4. All images were acquired using T1-weighted spin echo sequences on a Philips Ingenia 3.0T. The combined superior-inferior range of the scans extended from the most distal part of the toes to the level of vertebrae L5 up to L2.

2.1.2 Volume Model

The first dataset was retrieved from Henson et al. (2023) and contained augmented images and labels [20]. The labels include 37 volume segmentations of the following muscles: adductor brevis, adductor longus, adductor magnus, biceps femoris caput brevis, biceps femoris caput longum, extensor digitorum longus, extensor hallucis longus, flexor digitorum longus, flexor hallucis longus, gastrocnemius lateralis, gastrocnemius medialis, gemellus superior, gluteus maximus, gluteus medius, gluteus minimus, gracilis, iliacus, obturator externus, obturator internus, pectineus, peroneus brevis, peroneus longus, piriformis, popliteus, psoas, quadratus femoris, rectus femoris, sartorius, semimembranosus, semitendinosus, soleus, tensor fasciae latae, tibialis anterior, tibialis posterior, vastus intermedius, vastus lateralis, and vastus medialis. The images included the full lower limbs of 11 post-menopausal women (mean age 69 ± 7 years, mean weight 66.9 ± 7.7 kg, mean height 159 ± 3 cm) without movement limitations [34]. All scans were acquired during a hospital visit on a 1.5 T Magnetom Avanto scanner using T1-weighted sequences. Imaging parameters included an echo time of 2.59 ms, a repetition time of 7.64 ms, and a flip angle of 10°. Voxel dimensions were 1.1 × 1.1 × 5.0 mm for long bones and 1.1 × 1.1 × 3.0 mm for joint regions. The augmented dataset consists of 69 combined lower limb scans based on the original 11. The second dataset consists of a lower limb image of an anonymous participant with voxel sizes of 0.98 × 0.98 × 1.5 mm. No label was available for this image. An axial fast-spoiled gradient-echo sequence (LavaFlex, GE) with an isotropic resolution of R1.5 was used. All scans of this participant were acquired on a 3T GE PET/MRI hybrid system.

2.2 Pre-processing

2.2.1 Point Model

In Figure 2 the workflow of the preprocessing is presented. The hip, femur, knee and shank MR scans of each participant were pre-processed in the following order: intensity rescaled if necessary, cropped, resampled if necessary, merged into one image, and split into left and right. Intensity rescaling was only applied to the femur scan of C16 because the distribution of intensities far exceeded the other scans. This femur scan has been scaled with the maximum value found in the other scans of the participant. Cropping was done in the superior-inferior direction to remove the planes containing only zeros. If necessary, the scans were resampled to match the voxel sizes of the hip using a fourth order spline interpolation. The scans were merged into a single image by using the affine matrices of the local and target scans to map each voxel in the local scan to its corresponding voxel in the target scan. In the case of overlap, intensity values were averaged using the inverse of the absolute product of the diagonal terms of the original affine matrix.

In this dataset, muscle point locations are specified in the scanner's reference frame. These points were defined in a combined image that was manually aligned by only using translations. The following steps were taken to achieve a fully automated alignment that accommodates rotations while preserving the original voxel index for each muscle point. The annotated voxel of each muscle point was located by using the original alignment and the scan's affines, subsequently, these voxel indices were transformed back to the combined image's voxel space in the same way the individual scans are aligned. The only exception was C16, where the affine matrices were missing in the header, so the original alignment was used.

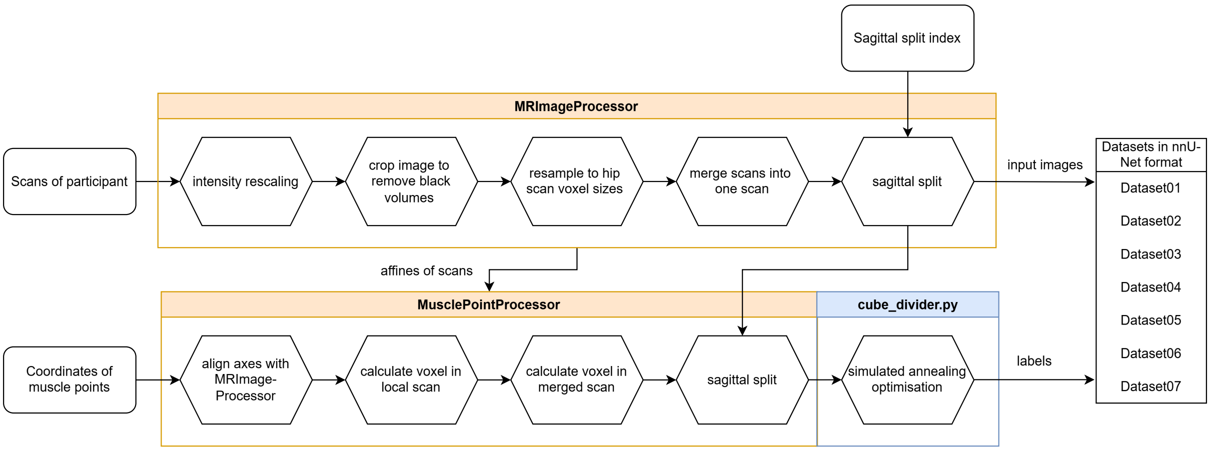
Sagittal split index
MRImageProcessor
Scans of participant
intensity rescaling
crop image to remove black volumes
resample to hip scan voxel sizes
merge scans into one scan
sagittal split
input images
Datasets in nnU-Net format
Dataset01
Dataset02
Dataset03
Dataset04
Dataset05
Dataset06
Dataset07
affines of scans
MusclePointProcessor
cube_divider.py
Coordinates of muscle points
align axes with MRImage-Processor
calculate voxel in local scan
calculate voxel in merged scan
sagittal split
simulated annealing optimisation
labels

Figure 2: Overview of the conceptual preprocessing workflow for the point model. Orange containers represent objects used and blue containers represent scripts. The code can be found on https://gitlab.tudelft.nl/clinical-biomechanical-lab.

Since U-Net is a volume segmentation model and point prediction is desired, a workaround has been implemented by placing 35-voxel cubes around the muscle points, inspired by [28]. After inference, the centroid of the segmentation is calculated, to obtain a point. Beforehand, it was explored which cube sizes led to steady improvements in pseudo DSCs while training for roughly 50 to 100 epochs. Because nnU-Net only allows for a unique label map and the cubes overlap given the muscle points, a subset of relevant muscle points of a typical lower limb model was selected. The following muscles with a number of points have been included: tensor fasciae latae (TFL_1-4), iliacus (iliacus_1-5), psoas (psoas_1-5), long head and short head of the biceps femoris (bi_fem_lh/sh_1-3), gastrocnemius lateral head (gas_lat_1-2), gastrocnemius medial head (gas_med_1-2), gluteus maximus (glut_max1-3_1-4), gluteus medius (glut_med1-3_1-2), rectus femoris (rectus_fem_1-2), semimembranosus (semimem_1-3), semitendinosus (semiten_1-5), vastus intermedius (vas_int_1-3), vastus lateralis (vas_lat_1-3), vastus medialis (vas_med_1-3). The final number in the shorthand notation increases with the distal position of the point (Figure 3). This subset consists of 61 points that are most contributing in terms of support [35], knee stability [36, 37], knee loading [36, 37], and that have the largest effect on other muscle forces [8] during gait. These points were divided over seven datasets, and thereby models, to further avoid label overlap. A simulated annealing optimisation was implemented to find a division of muscle points over the datasets to avoid overlap in all instances (Figure 2). Although the point model is not one model, but seven, it will still be referred to as one. Each sample consisted of an MR image capturing a unilateral view of a single leg (Figure 4).

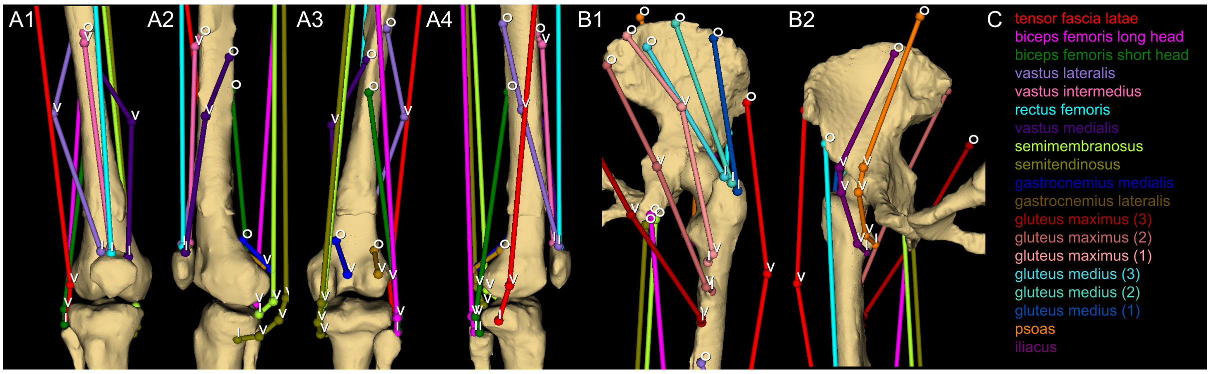

Figure 3: Muscle points defining muscle action lines, where O, V, and I represent the origin, via, and insertion points, respectively. The views illustrate the approximate locations of muscle points relative to the bones. A1–A4 show an anterior, medial, posterior, and lateral view of the knee. B1–B2 show the hip joint. C provides the legend for the muscle action lines.

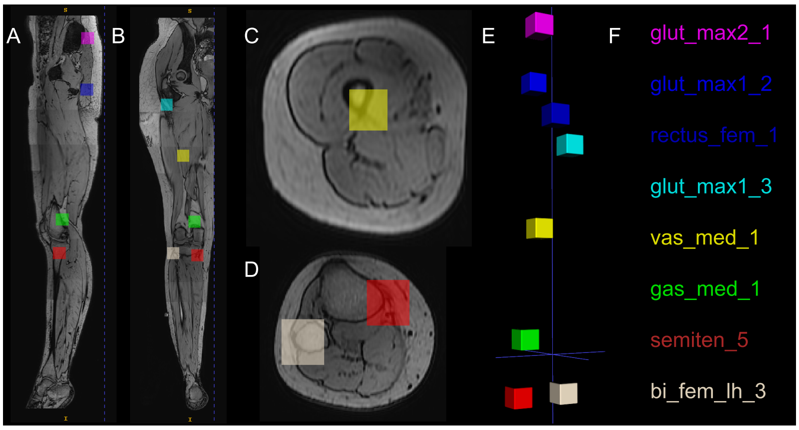

Figure 4: Example of a sample out of the seven datasets used to train the point model. The pictures include a sagittal view (A), a coronal view (B), axial views at the thigh (C) and shank (D), and a 3D representation of the muscle point labels (E). The legend for the muscle point labels is provided in (F).

2.2.2 Volume Model

The DICOM files of both datasets were converted to NifTi files using the NiBabel library [38]. The original class values of the labels of the first dataset were roughly evenly distributed across a byte (0-255) [20]. These values were then replaced with consecutive integers, as required by the nnU-Net format. Before performing this replacement, a unique mapping between the original values and the new class integers was created by iterating over all class values in the labels. This mapping was used to ensure consistent meaning of class values across samples. For both datasets, all samples consisted of an MR image capturing a unilateral view of a single leg (Figure 5).

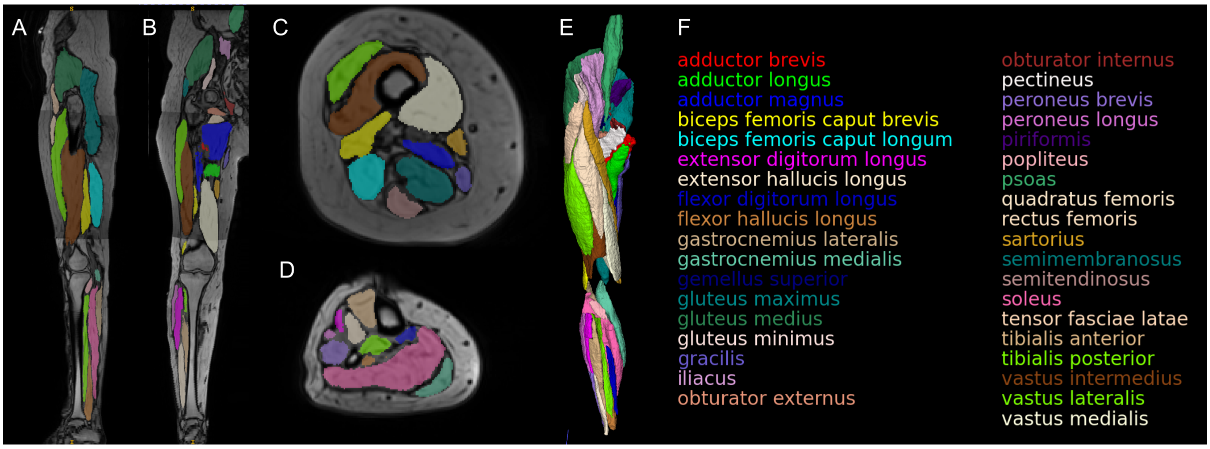

Figure 5: Example of sample out the Henson et al. dataset [20] used to train the volume model. The pictures include a sagittal view (A), a coronal view (B), an axial view at the thigh (C), an axial view at the shank (D), and a 3D representation of the muscle volume segmentations (E). The legend for the muscle segmentation labels is shown in (F).

2.3 U-net Models

The point and volume U-net models were created by the nnU-Net pipeline [31]. Both models have 2D and 3D configurations, and the volume model includes an additional ensemble prediction. An ensemble configuration was not created for the point model because nnU-Net requires training on five folds. To reduce GPU usage due to the seven point models, only a single fold was trained for each dataset instead of the default five. The pipeline was configured with a specified VRAM capacity of 80 GB. Training has been performed on the supercomputer DelftBlue [39] using the NVIDIA A100 Tensor Core GPU with 80 GB of VRAM. The volume and point models have been trained for 250 and 500 epochs per fold, respectively. For the remaining specifications, the default was used for both models.

2.4 Validation

2.4.1 Point Model

The point model was validated by comparing its results on the test set against the annotations provided by an experienced muscle point annotator, which served as the ground truth. Model accuracy was assessed using both the Euclidean Error (EE) and the absolute distance error along each axis. For all spatial errors, the median is reported with the interquartile range (IQR) in parentheses. The relation between the Dice Similarity Coefficient (DSC) and EE was also evaluated to validate the cube-segmentation approach to perform point prediction. A custom randomised 80-20 train-test split was implemented for both the 2D and 3D configurations. By randomly sampling from both populations in proportion to their occurrences, both populations are equally represented in each split.

2.4.2 Volume Model

The performance of the volume model was evaluated by comparing its predictions on the test set with the ground truth labels, using the Dice Similarity Coefficient (DSC) as a measure of accuracy. The median DSC is reported with the IQR in parentheses. The default five-fold cross-validation setup of nnU-Net was applied to the first dataset. Inference on the second dataset was performed for a qualitative evaluation to better understand the models performance, as validation with the first dataset is limited by the augmented relation between test and train samples.

3 Results

3.1 Point Model

Figure 6 shows a prediction with the 3D point model and the ground truth. The locations of the predicted segmentations are similar to those of the ground truth. The shapes of the predictions are roughly cube-like, although they deviate in some instances, such as at the third point of the vastus lateralis (Figure 6C). The first via point of the TFL is missing in the prediction (Figure 6D).

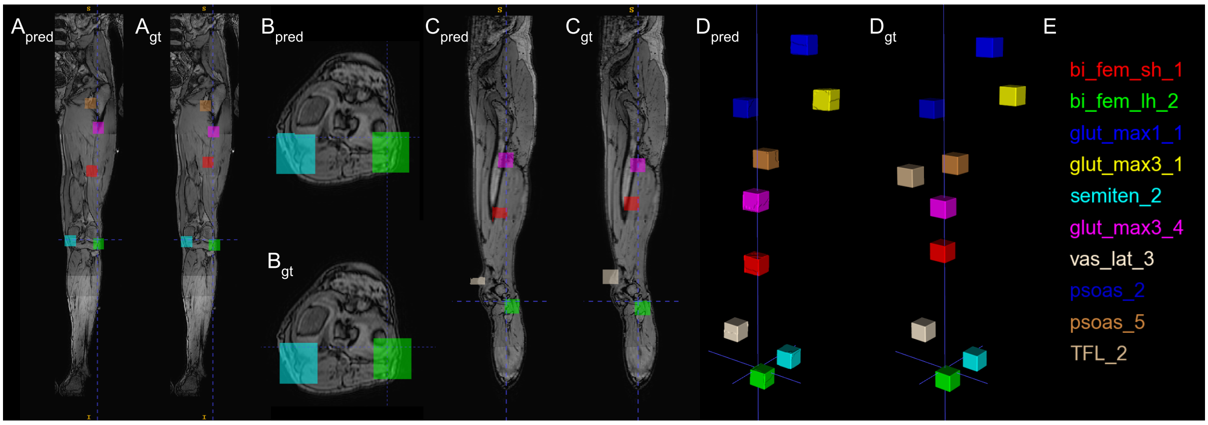

Figure 6: Predictions (pred) and ground truths (gt) of the 3D point model on C14 (L) out of the validation set. A, B, C, and D show a coronal, axial, sagittal, and 3D view, respectively. E shows a legend of the muscle point label.

The EEs for the 3D and 2D point models are roughly six times higher for C16 compared to the other samples (Table 1). At least one scan of C16 is assumed to be an outlier due to technical imaging errors. As a result, this sample will be excluded from further analyses and conclusions related to the point model. Detailed results including C16 are provided in the Appendix in Figure 11 and 12.

Table 1: Spatial errors (mm) of the point model per sample in the validation set.

	2D								3D							
	Euclidean		Coronal		Sagittal		Axial		Euclidean		Coronal		Sagittal		Axial	
Sample	MED	IQR	MED	IQR	MED	IQR	MED	IQR	MED	IQR	MED	IQR	MED	IQR	MED	IQR
C16 (L)	**34.1**	**9.1**	29.3	8.0	13.5	11.6	7.8	12.1	35.1	12.2	29.3	9.0	12.1	13.3	8.8	14.0
C1 (R)	5.4	4.3	1.9	3.0	1.8	2.8	3.1	3.7	**4.2**	**4.0**	1.9	2.0	2.8	2.8	3.0	2.8
C14 (L)	7.3	4.6	2.8	3.0	4.7	2.8	4.0	6.5	**5.3**	**5.6**	1.9	2.0	3.7	1.9	2.0	4.7
OA1 (L)	**6.4**	**5.8**	2.8	2.0	2.8	3.0	3.0	4.7	6.6	5.3	2.8	2.0	2.8	3.7	2.0	4.7
OA2 (R)	5.8	3.5	1.9	3.0	2.8	2.8	2.0	3.7	**5.1**	**3.2**	1.9	2.0	2.8	2.8	3.0	3.7

Bold numbers indicate the lowest overall Euclidean Error per participant between the 2D and 3D configuration. L and R in the sample name refer to the unilateral side from the perspective of the participant.

The pooled median for the 2D and 3D point model configurations are 6.3 (4.5) and 5.1 (4.3) mm, respectively (Table 2). The 3D model outperforms the 2D model most of the time when grouping based on individual samples (Table 1), point type, and anatomical regions (Table 2). The median EE difference between via and attachment points is small in both configurations. Out of all the anatomical regions, the femur region has the highest EE in both configurations. The 2D EE is 30% lower than the 3D EE in the femur region. In both configurations of this anatomical region, the largest amount of error is along the sagittal axis and the highest variability is along the axial axis.

Table 2: Spatial errors (mm) of the point model per category.

	2D								3D							
	Euclidean		Coronal		Sagittal		Axial		Euclidean		Coronal		Sagittal		Axial	
Category	MED	IQR	MED	IQR	MED	IQR	MED	IQR	MED	IQR	MED	IQR	MED	IQR	MED	IQR
att. (38)	6.4	4.6	2.8	3.0	2.8	3.0	3.0	3.7	**4.9**	**3.8**	1.9	2.0	2.8	2.8	2.1	3.7
via (23)	6.3	4.5	1.9	3.0	2.8	2.8	3.1	4.7	**5.1**	**4.3**	1.9	2.0	2.8	2.8	2.1	4.4
hip (35)	6.9	5.0	2.8	3.0	3.7	3.7	3.9	3.8	**5.7**	**4.6**	1.9	2.0	3.7	2.8	3.0	4.7
femur (5)	**7.5**	**4.9**	0.9	2.0	4.7	1.9	2.2	6.5	10.7	7.8	1.9	2.0	5.6	2.3	2.0	7.9
knee (10)	**4.5**	**3.5**	2.8	3.0	1.8	1.9	3.5	2.3	4.7	3.2	2.8	2.0	1.8	3.7	2.5	1.9
tibia (11)	4.5	3.7	1.9	2.0	2.8	1.9	2.0	3.7	**3.6**	**2.1**	1.9	2.0	1.9	0.9	2.0	2.8
overall (61)	6.3	4.5	1.9	3.0	2.8	2.8	3.1	4.7	**5.1**	**4.3**	1.9	2.0	2.8	2.8	2.1	4.4

Bold numbers indicate the lowest overall Euclidean Error between the 2D and 3D configuration. Attachments are referred to as 'att.' in the table. The numbers in parentheses next to the categories represent the average number of points per participant.

The different mode configurations performed in most muscle points similarly with the exception of the vas_lat_2, vas_med_2, TFL_2, and psoas_1 (Figure 7). The vas_lat_2, vas_med_2, and psoas_1 favour the 2D model. Only the TFL_2 shows a favour towards the 3D model, mostly in terms of range. Both model configurations show a large range and moderately high median for the bi_fem_sh_1. The high maximum of the bi_fem_sh_1 results from C14 (L). In three instances, via points were not predicted. The 3D model failed to segment vas_lat_2 in OA1 (L) and TFL_2 in C14 (L; Figure 6). The 2D model failed to segment TFL_2 in OA1 (L).

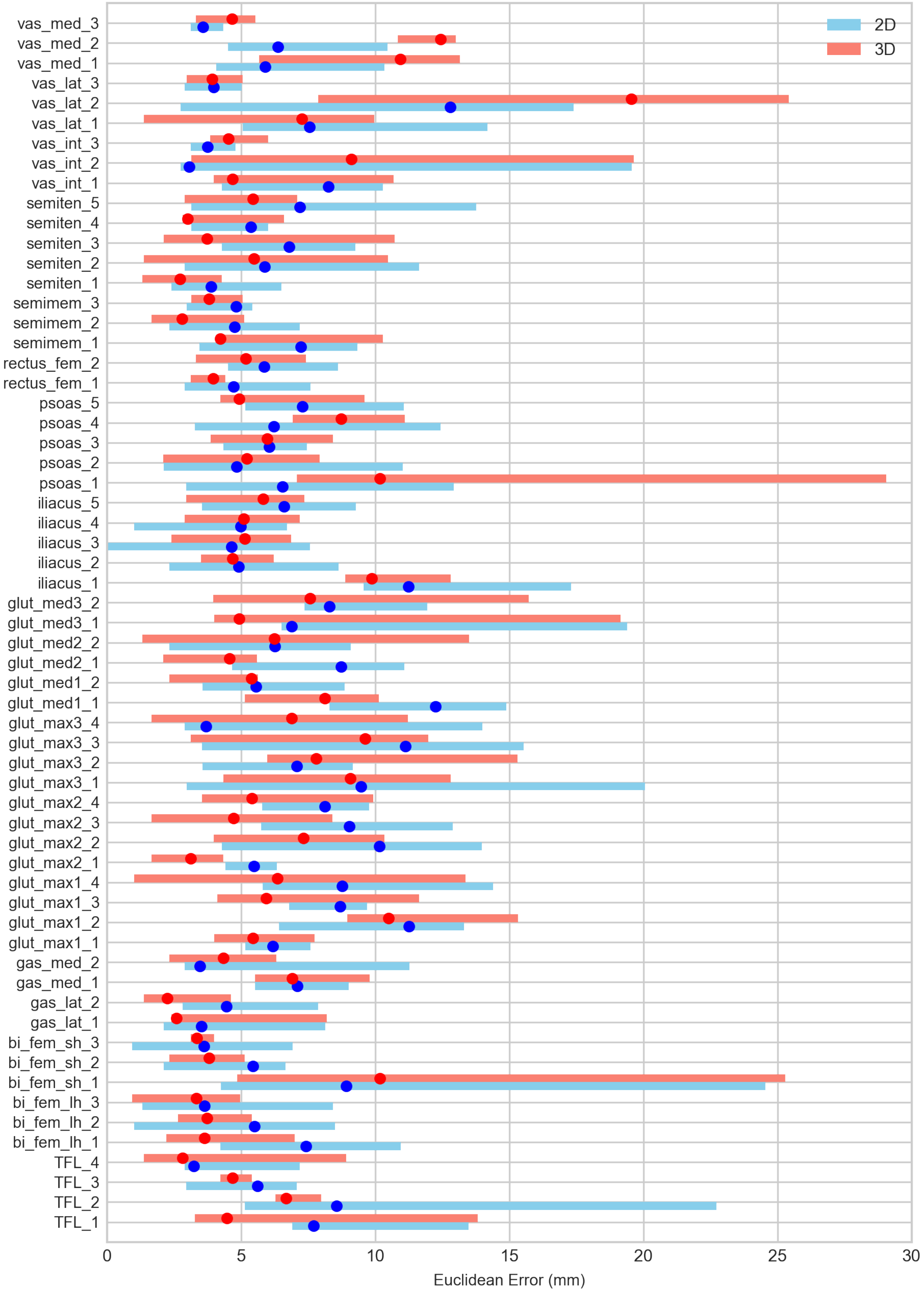

Figure 7: Median Euclidean errors per muscle point (circles) with minimum and maximum range (bars).

The pooled median DSC of the 2D and 3D point model configurations are 71.9 (16.4) and 74.5 (18.3)%, respectively. Figure 8 presents the relation between the DSC and EE for all muscle points. Generally, the higher the DSC, the lower the EE. The variance of EE also decreases with a higher DSC.

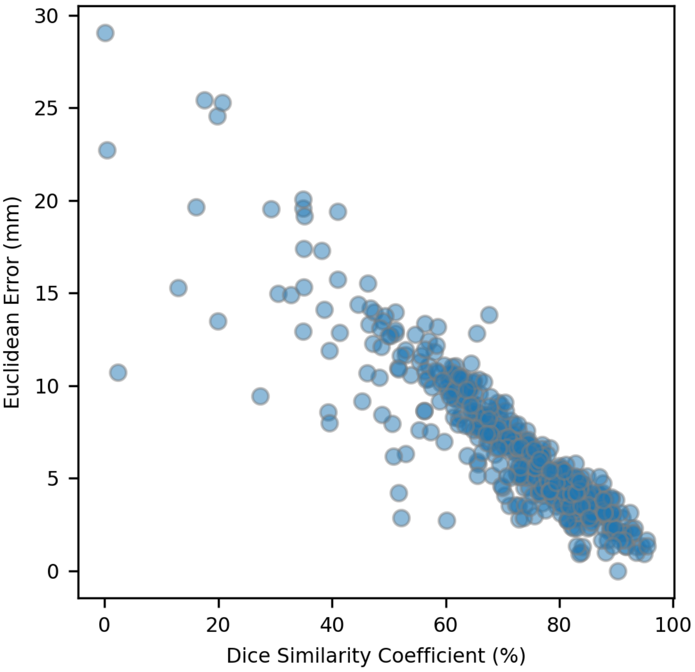

Figure 8: Every DSC plotted against the corresponding Euclidean error.

3.2 Volume Model

Figure 9 shows a segmentation with the 3D volume model and the ground truth. Overall, the prediction and ground truth segmentation look very similar. Both the prediction and ground truth show a lack of labelled voxels in the medial midway thigh area (Figure 9A). In both prediction and ground truth, misplaced small fractions of the vastus lateralis can be observed (posterior to the brown vastus intermedius in Figure 9C).

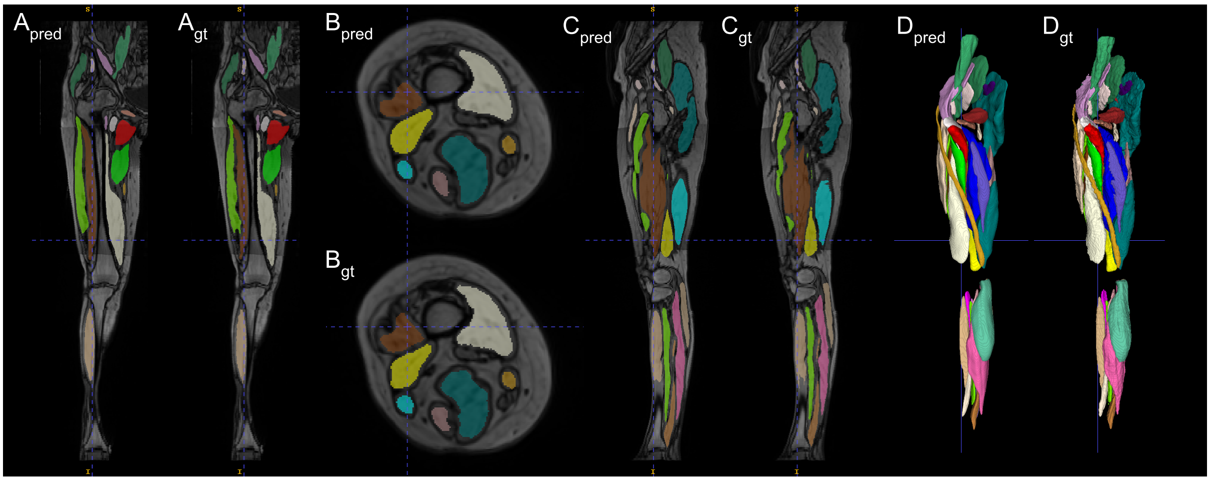

Figure 9: Predictions (A, C, E, and G) and ground truths (B, D, F, and H) of the volume model of sample number 13. A, B, C, and D show a coronal, axial, sagittal, and 3D view, respectively. The legend of the segmentations can be found in Figure 5.

Figure 10 shows an inference result from the second dataset sample. Compared to the inference result in Figure 9, this one appears less densely annotated. Specifically, the adductor brevis, adductor longus, flexor digitorum longus, extensor hallucis longus, and semimembranosus muscles are not fully segmented. The segmentations for the gluteus maximus, gluteus medius, vastus intermedius, sartorius, soleus, tibialis anterior, gastrocnemius lateralis, and gastrocnemius medialis are more complete compared to the others. The relative positions of the larger segmentations are anatomically plausible. The only exception is a small portion of the vastus medialis segmentation, which incorrectly overlaps with the caudal part of the femur (Figure 10A).

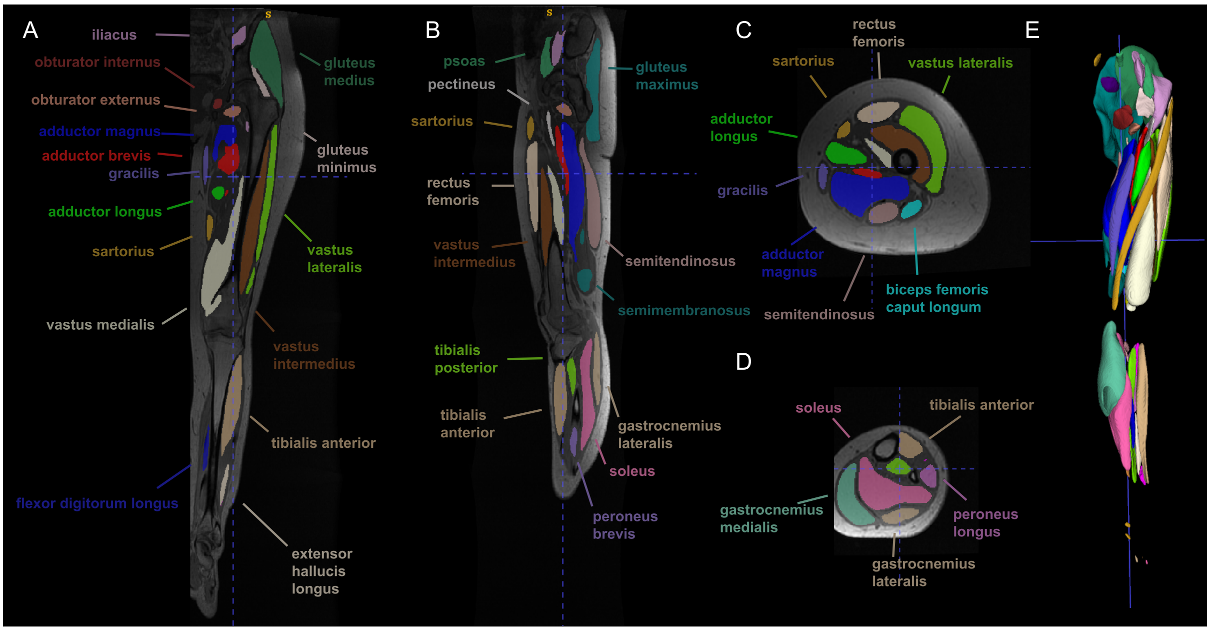

Figure 10: Inference result of the second dataset, shown for qualitative purposes. A coronal (A), sagittal (B), axial (C-D), and 3D view (E) are presented, where the upper axial slice is located at the thigh (C) and the bottom one at the shank (D). The location of the axial thigh slice is shown in A, B, and E by the blue dotted lines. The full legend of the labels can be found in Figure 5.

The 3D configuration of the volume model scores the highest median DSC of 92.7 (4.4)% (Table 3). The quadratus femoris was not segmented in 11 cases for the 2D configuration, 12 cases for the 3D configuration, and 12 cases for the ensemble configuration.

Table 3: DSC values (%) for each muscle for the different volume model configurations.

	2D		3D		Ensemble	
muscle	MED	IQR	MED	IQR	MED	IQR
adductor brevis	90.1	3.5	**92.2**	**2.0**	91.5	2.9
adductor longus	92.8	1.9	**94.2**	**1.8**	93.6	1.8
adductor magnus	95.2	1.1	**96.0**	**0.8**	95.9	0.9
biceps femoris caput brevis	92.9	4.6	93.8	3.0	**93.8**	**3.8**
biceps femoris caput longum	95.1	1.1	**95.7**	**0.8**	95.6	0.7
extensor digitorum longus	87.3	5.5	**90.4**	**4.5**	89.0	5.1
extensor hallucis longus	90.0	3.7	90.3	2.8	**90.7**	**3.3**
flexor digitorum longus	87.5	4.6	87.9	3.6	**88.2**	**4.3**
flexor hallucis longus	89.0	4.1	89.2	3.6	**89.7**	**3.8**
gastrocnemius lateralis	90.5	3.4	**91.6**	**1.8**	91.5	3.0
gastrocnemius medialis	95.2	1.9	95.7	1.2	**95.7**	**1.5**
gemellus superior	83.8	8.3	**87.2**	**4.5**	85.8	7.3
gluteus maximus	96.1	1.2	96.1	1.1	**96.4**	**1.1**
gluteus medius	95.0	1.1	95.2	0.7	**95.5**	**0.8**
gluteus minimus	89.8	3.4	90.3	2.8	**90.7**	**2.6**
gracilis	87.9	5.5	**89.0**	**4.9**	89.0	5.5
iliacus	92.6	1.6	93.2	1.3	**93.4**	**1.3**
obturator externus	89.0	4.1	**92.1**	**2.5**	90.9	2.8
obturator internus	90.4	4.1	91.4	2.8	**91.5**	**3.7**
pectineus	90.8	2.9	**92.8**	**1.8**	92.3	2.4
peroneus brevis	88.8	4.7	88.5	4.5	**89.2**	**4.5**
peroneus longus	90.4	3.5	**91.9**	**2.5**	91.5	2.8
piriformis	91.2	2.9	91.9	2.0	**92.2**	**2.6**
popliteus	85.9	8.3	**88.4**	**5.1**	87.4	7.3
psoas	92.1	3.0	**93.0**	**2.6**	92.9	2.8
quadratus femoris	84.4	11.8	**86.3**	**6.1**	86.2	8.9
rectus femoris	94.4	1.4	**94.7**	**1.0**	94.7	1.0
sartorius	90.4	5.0	91.6	3.5	**91.7**	**4.2**
semimembranosus	94.2	2.0	**95.0**	**1.3**	94.9	1.5
semitendinosus	93.5	2.1	**94.5**	**1.4**	94.3	1.5
soleus	93.8	2.3	94.2	2.0	**94.3**	**2.0**
tensor fasciae latae	91.9	2.1	92.5	1.8	**92.5**	**1.9**
tibialis anterior	93.8	1.5	94.1	1.5	**94.2**	**1.3**
tibialis posterior	91.8	2.3	92.0	2.2	**92.4**	**2.3**
vastus intermedius	93.2	2.2	**93.9**	**2.2**	93.8	2.0
vastus lateralis	94.3	1.6	**95.0**	**1.6**	94.8	1.5
vastus medialis	95.0	1.4	**95.8**	**1.1**	95.7	1.1
overall	91.8	5.2	**92.7**	**4.4**	92.6	4.6

The bold numbers indicate the DSC with the highest median for each muscle. In all three model instances, there were cases where no pixels were segmented for the quadratus femoris.

4 Discussion

4.1 Point Model

The cube-segmentation approach results in accurate muscle point predictions in the lower limbs, with an overall median EE of the 3D configuration of 5.1 (4.3) mm. Compared to Scheys et al., this study scores a 16% lower EE overall [18]. However, it must be mentioned that in this study only the relevant points of the lower limbs are selected with only a few near the tibia and none at the ankle. Moreover, the EE of the point model is smaller in terms of median and IQR compared to the inter- and intra-rater operator attachment point variability of 5.6 (10.7) and 6.9 (7.7) mm, respectively [18]. Assuming that the operator-variability is the same for the dataset in this paper, including via points, this suggests that the 3D point model is indistinguishable from human expertise in terms of EE in most cases.

The biggest limitation of this study is the dependency between the train and test set. The model is trained and validated on opposing leg scans of the same individual. To assess the extent to which this dependency biased the results, the same operator labelled another sample of the point model dataset from the healthy group; C0 (L). This sample was not used previously because the ground truth was missing, and the scans lacked the superior part of the pelvis (vertebral L5 is not visible). The points that should be placed on the superior part of the pelvis (glut_med1-3_1, glut_max1_1, psoas_1, and iliacus_1) are excluded from the following results. The point model scored a median EE of 8.9 (5.6) and 6.8 (5.8) mm for the 2D and 3D configurations, respectively. Both EEs exceed the respective pooled medians and IQR sizes of the validation set. Only the median and total range of the 3D configuration fall within the IQR and total range of the validation set. The EE distributions of the validation set and the new sample show a lot of overlap (Figure 13). Apart from gaining insight into the possible dependency bias, this sample also introduces challenges specific to its unique characteristics. First, the participant had a relatively low body fat percentage which could complicate muscle volume segmentation [19] and possibly point prediction as well. Only two other participants were similar in terms of body fat percentage, where one of which is present in the validation set: C14 (L: Figure 6). C14 (L) scored the highest 2D EE and second highest 3D EE. This supports the notion that the low body fat percentage of the new sample also affected the results. Second, there was approximately an eight-year gap between the labelling of the validation set and the new sample, which undoubtedly introduced variability to the ground truth muscle point locations. It is assumed that the dependency of the validation set on the train set led to a slight underestimation of the true EE. Given the result of the new sample, the 3D configuration still scored a lower median and IQR compared to intra-operator attachment point variability.

The 3D configuration outperforms the 2D configuration in terms of DSC and EE in most cases (including the new sample C0L), but the significance of this difference remains uncertain. A statistical test was not performed due to the small validation set, which limits the likelihood of finding significant differences. The high error observed in the sagittal direction of the femur region suggests that point prediction performance may decrease with voxel size. Both configurations were affected by this, though the 3D configuration appeared more sensitive to it. The high error variability of the origin of the psoas in the 3D configuration might indicate that the varying anatomical range withholds the model from learning general patterns. In three instances, via points were not predicted, which may indicate that via points are more challenging to localise than attachment points. However, the difference in EE medians between the two point types is small, and the point type with the lowest error varies across configurations. The implicit nature of deep learning models makes it difficult to pinpoint the causes of deviating results.

Unlike in typical deep learning studies, where the cost function consists of the metric the model is evaluated with, this study performed an indirect approach with the point model. The standard cost function of nnU-Net was used to train the point model which consists of an equal weighting combination of the cross-entropy loss and DSC term. Figure 8 shows that, for this study, when the DSC increases, the EE decreases. The DSC is often used to assess segmentation performance, therefore, it could be assumed that this approach is reasonable to estimate points when good segmentation performance is expected. This approach is appealing in its current form because it does not require any modifications to the nnU-Net. However, this comes with a high computational cost. This approach needs multiple models to avoid overlap of cubes which scales the training computations needed linearly. With which factor it scales, depends on how close the points are relative to each other and how large the cubes are. It is observed in this study, that a sufficiently large cube is needed for the cube-segmentation approach to work. A cube with an edge length of 11 voxels resulted in poor increases of pseudo DSCs within 100 epochs. Therefore, adjusting the architecture of U-Net to allow for overlap between cubes would be beneficial. Isensee, the founder of the nnU-Net, stated that this could be done by changing the activation function of the output layer to a binary sigmoid function and the cost function's cross-entropy term to a binary one [40]. Additionally, a thresholding method needs to be applied. In this way, the voxels are no longer mutually exclusive. It was also noted that this would likely be incorporated in the next version of nnU-Net.

Each point model shows a higher final training loss than final validation loss for both configurations (see example in Figure 16). This points towards successful training. However, there are still likely three factors that diminish performance. First, it was assumed that the left portion of C16 images is not closely related to the other images. The other sagittal half was present in the training set. This probably stagnated the progress during the training and thereby performance. Second, for a deep learning model, a dataset of 24 cases is rather small. Third, it can be seen that after around 100 epochs in the 3D configuration, the training loss converges and the validation loss has a slight decreasing trend (Figure 16). This suggests that more epochs could increase performance. This is not as surprising since 1000 epochs are the default for nnU-Net.

The image alignment process introduced some artefacts at the shank in the case of C1. These artefacts include a few black lines in the shank region (Figure 14A) and a 'blocky' appearance at primarily the posterior end of the shank (Figure 14B). Further investigation revealed that, when overlapping voxels were not averaged, bright lines also appeared (Figure 14C-D). These artefacts are caused by the voxel spaces having different orientations, which leads to misalignment between the voxels in the two spaces using the current approach. As the misalignment errors accumulate and necessary rounding to integers is applied, they eventually result in either bright or black voxels. The contracting drift along the posterior-anterior axis results in periodic skipping of values, which appear as black voxels. Conversely, the expanding drift along the superior-inferior axis causes a periodic double-intensity value, appearing as bright voxels. This issue can be resolved by reorienting the MR data of the local scan to match the orientation of the target scan beforehand and updating the local affine matrix accordingly. The off-diagonal terms in the rotation and scaling portion of the transformation matrix are relatively small, approximately one-hundredth the size of the diagonal terms. Therefore, it is assumed that these artefacts had minimal impact on the images and point accuracy of the prediction. This is evident in the relatively low accuracy error for both configurations in the case of C1 (R).

Not including points around the shank is a limitation of the dataset, as the Achilles tendon insertion exhibits the highest overall sensitivity to other muscle forces during gait [9]. Actually, not including muscle points at the shank might have underestimated the overall error for muscle points in the lower limbs. It could be assumed that it is more challenging to accurately define points around the ankle because of the observed variability of the posture. This adds extra complexity to the problem for the model to account for. Although the authors did not attribute it primarily to the lower signal-to-noise ratio, it was observed that registration methods to estimate muscle points showed lower accuracy in the more caudal parts of the scans [18].

In this study, the EE is primarily used to validate the point model, which does not directly indicate the effect on moment arm lengths. Not every muscle point error along each axis contributes equally to moment arm length errors. For example, the maximum EE of the origin of the short head of the biceps femoris was high (Figure 7). The largest part of the EE was found in the axial axis which roughly aligns with the direction of the action line and the distance with the next muscle point is quite large. In turn, the resulting moment arm length errors are expected to be small because the change in muscle action line direction is small. It is also possible for a small muscle point error to cause a relatively large change in moment arm length. This is the case because in some instances attachment and via points are positioned close together, such as in the psoas, iliacus and semimembranosus. This increases the directional change of muscle action lines due to a muscle point error, thereby affecting the moment arm length. Future research will focus on the effect of the reported errors in muscle points on the moment arm length errors.

4.2 Volume Model

The 3D volume model has been shown to be able to segment lower limb muscles accurately in the validation set (median DSC of 92.7 %). A beneficial result is that the DSCs of muscles not well represented by a poly-line approach (iliacus, psoas, gluteus maximus and gluteus medius) score relatively high (average DSC of 94.5%). However, drawing conclusions based on this result is not sound because of the dependency of the train and test data on each other. This is inherently the case because of the augmentation process and cannot be resolved because no further details are provided regarding the relation between the augmented and original data. The inference of the sample of the second dataset shows that the volume model is capable of generalising to other data so far that the relative positions of the segmentations are correct. This sample is different from the original dataset in scanner, settings, and scan post-processing. This suggests that the model could be used for lower limb muscle segmentation to speed up the labelling of similar datasets.

The volume model not predicting the quadratus femoris at all instances, can be explained by the fact that twelve labels did not contain this class. It was found that in these cases, the segmentation of the adductor magnus covered the actual quadratus femoris voxels (Figure 15). It is likely that multiple original labels contain this error and during the augmentation process, it has propagated. In the study itself [20], not all segmentations were used because some muscles showed low repeatability in terms of volume. For the same muscle, not more than 10% variation in volume was allowed. Out of the two muscles, only the quadratus femoris was excluded. Since the adductor magnus is typically larger than the quadratus femoris, it could be the case that this volume criteria is inadequate to detect such a problem. Furthermore, out of all the relative muscle volumes of the augmented data, the adductor magnus showed the largest relative range in volume. This can be explained by the fact that in some instances, the quadratus volume was added to it, and in others, subtracted from it. This further supports the notion that this is the case.

At first glance, high DSC coefficients show great promise for an automated complex approach pipeline to model muscle-tendons (Figure 1). However, muscle segmentation studies do often not specify what is regarded as the boundaries of a muscle. This is evident in the fact that the segmentation volume variability is often larger than 10% [20]. This makes comparing volume models based on expert opinion ground truths between studies less valid. Furthermore, it has been shown that for the iliopsoas and gluteus medius, the intra-operator segmentation variability increases when approaching the attachment sites in MR images [41]. It is likely that errors towards the attachments in complex muscle models will also be the most sensitive to changes in muscle force during gait, as is the case in the poly-line approach [9]. This further diminishes the utility of the DSC for biomechanical purposes. In this study, the labels of the volume dataset did not include all tendons crossing the knee joint (Figure 5). Reporting the proportion of segmentation coverage along the direction of a straight line from the origin to the insertion could aid in standardisation that is relevant for creating subject-specific MSMs. Such a technique has been implemented in [41]. This could be another potential use case for muscle attachment prediction models like the point model in this study.

Typically, volume models are used for the complex approach, however, they could also be used for the poly-line approach. In fact, the poly-line approach or also known as the centroid line approach, used volumes to estimate centroids when it was introduced [42]. This could improve the repeatability issues in defining muscle points by eliminating subjective decisions from the process. The automated technique proposed by Modenese and Kohout allows the user to define the level of discretisation of a muscle, i.e. picking the poly-line or complex approach or anywhere in between [17]. This technique could serve as a tool to investigate the effect of the discretisation of a muscle-tendon per individual in a systematic manner on any desired outcome variable. For the iliacus, psoas, gluteus maximus and gluteus medius it has been shown that most of the time the moment arms of the poly-line approach fall mostly within the range of the complex approach over a physiological hip range of motion [17]. However, how these differences affect desirable internal loading parameters during gait has not yet been studied. This is especially relevant because when using the poly-line approach, it is known that the perturbing the gluteus medius, iliacus, and psoas results in the largest changes in muscle forces during gait [8].

Muscle volume models can offer some more benefits in the scope of subject-specific MSMs apart from defining muscle paths. Firstly, when muscle fibre paths and the volume are known, as is typically the case in the complex modelling approach, the physiological cross-sectional area can be calculated. This variable is assumed to highly correlate with the maximal isometric force of a muscle which has also shown sensitivity to muscle force estimations [43]. Secondly, the current point model has been shown to not always define via points reliably. Since both models take MR scans as input, it takes minimal extra manual labour to run an inference of the volume model. Given the muscle volumes, a centroid approximation can be done to determine the location of the missing via point or it could be checked if the point at hand is classified as the correct corresponding muscle class. Since muscle volume segmentations often contain the most segmented voxels in the central parts of the muscle and only via point predictions were absent, such a strategy could function. Of course, checking the output of the volume is also possible with the point model. Both propositions and models in this paper could be integrated into the Musculoskeletal Atlas Project Client to minimise the required labour to include subject-specificity [44].

5 Conclusion

This study introduces two models designed to automatically identify muscle points and volumes, offering researchers and clinicians tools to estimate internal loading measures of the musculoskeletal system in vivo. The 3D muscle point model shows comparable accuracy to intra-operator variability of attachment points. The cube-segmentation approach used is accessible and requires limited deep learning expertise. It is likely that with more coherent data, this approach could attain higher accuracy. Future research will include a moment arm length error analysis. The muscle volume model could be used as a starting point to label other datasets that are similar to the current data it is tested and trained on. A validation with independent samples of the training data is required to evaluate the model's performance properly. Different use cases have been described where these models could support or progress following computational biomechanical research. This study demonstrated that the nnU-Net can produce accurate models that can automatically define subject-specific muscle-tendon paths.

References

[1] S. I. Sulsky, L. Carlton, F. Bochmann, R. Ellegast, U. Glitsch, B. Hartmann, D. Pallapies, D. Seidel, and Y. Sun, "Epidemiological Evidence for Work Load as a Risk Factor for Osteoarthritis of the Hip: A Systematic Review," *PLoS ONE*, vol. 7, p. e31521, 2 2012.

[2] S. H. Smith, R. J. Coppack, A. J. van den Bogert, A. N. Bennett, and A. M. Bull, "Review of musculoskeletal modelling in a clinical setting: Current use in rehabilitation design, surgical decision making and healthcare interventions," 3 2021.

[3] J. L. Hicks, T. K. Uchida, A. Seth, A. Rajagopal, and S. L. Delp, "Is My Model Good Enough? Best Practices for Verification and Validation of Musculoskeletal Models and Simulations of Movement," 2 2015.

[4] M. F. Bobbert, D. A. Kistemaker, M. A. Vaz, and M. Ackermann, "Searching for strategies to reduce the mechanical demands of the sit-to-stand task with a muscle-actuated optimal control model," *Clinical Biomechanics*, vol. 37, pp. 83–90, 8 2016.

[5] C. M. Dzialo, M. Mannisi, K. S. Halonen, M. de Zee, J. Woodburn, and M. S. Andersen, "Gait alteration strategies for knee osteoarthritis: a comparison of joint loading via generic and patient-specific musculoskeletal model scaling techniques," *International Biomechanics*, vol. 6, pp. 54–65, 1 2019.

[6] M. Wesseling, F. De Groote, L. Bosmans, W. Bartels, C. Meyer, K. Desloovere, and I. Jonkers, "Subject-specific geometrical detail rather than cost function formulation affects hip loading calculation*," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 19, pp. 1475–1488, 10 2016.

[7] G. N. Duda, D. Brand, S. Freitag, W. Lierse, and E. Schneider, "Variability of femoral muscle attachments," *Journal of Biomechanics*, vol. 29, pp. 1185–1190, 9 1996.

[8] L. Bosmans, G. Valente, M. Wesseling, A. Van Campen, F. De Groote, J. De Schutter, and I. Jonkers, "Sensitivity of predicted muscle forces during gait to anatomical variability in musculotendon geometry," *Journal of Biomechanics*, vol. 48, pp. 2116–2123, 7 2015.

[9] V. Carbone, M. M. van der Krogt, H. F. Koopman, and N. Verdonschot, "Sensitivity of subject-specific models to errors in musculo-skeletal geometry," *Journal of Biomechanics*, vol. 45, pp. 2476–2480, 9 2012.

[10] I. Hannah, E. Montefiori, L. Modenese, J. Prinold, M. Viceconti, and C. Mazzà, "Sensitivity of a juvenile subject-specific musculoskeletal model of the ankle joint to the variability of operator-dependent input," *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of engineering in medicine*, vol. 231, pp. 415–422, 5 2017.

[11] R. Bayoglu, O. Guldeniz, N. Verdonschot, B. Koopman, and J. Homminga, "Sensitivity of muscle and intervertebral disc force computations to variations in muscle attachment sites," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 22, pp. 1135–1143, 10 2019.

[12] S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, A. Habib, C. T. John, E. Guendelman, and D. G. Thelen, "OpenSim: Open-source software to create and analyze dynamic simulations of movement," *IEEE Transactions on Biomedical Engineering*, vol. 54, pp. 1940–1950, 11 2007.

[13] S. L. Delp, J. P. Loan, M. G. Hoy, F. E. Zajac, E. L. Topp, and J. M. Rosen, "An interactive graphics-based model of the lower extremity to study orthopaedic surgical procedures," *IEEE transactions on bio-medical engineering*, vol. 37, no. 8, pp. 757–767, 1990.

[14] F. van der Helm and R. Veenbaas, "Modelling the mechanical effect of muscles with large attachment sites: Application to the shoulder mechanism," *Journal of Biomechanics*, vol. 24, pp. 1151–1163, 1 1991.

[15] T. R. Kane and D. A. Levinson, "Dynamics, theory and applications," p. 379, 1985.

[16] S. S. Blemker and S. L. Delp, "Three-Dimensional Representation of Complex Muscle Architectures and Geometries," *Annals of Biomedical Engineering*, vol. 33, pp. 661–673, May 2005.

[17] L. Modenese and J. Kohout, "Automated Generation of Three-Dimensional Complex Muscle Geometries for Use in Personalised Musculoskeletal Models," *Annals of Biomedical Engineering*, vol. 48, pp. 1793–1804, 6 2020.

[18] L. Scheys, D. Loeckx, A. Spaepen, P. Suetens, and I. Jonkers, "Atlas-based non-rigid image registration to automatically define line-of-action muscle models: A validation study," *Journal of Biomechanics*, vol. 42, pp. 565–572, 3 2009.

[19] L. Piecuch, V. G. Duque, A. Sarcher, E. Hollville, A. Nordez, G. Rabita, G. Guilhem, and D. Mateus, "Muscle volume quantification: guiding transformers with anatomical priors," 10 2023.

[20] W. H. Henson, C. Mazzá, and E. DallAra, "Deformable image registration based on single or multi-atlas methods for automatic muscle segmentation and the generation of augmented imaging datasets," *PLoS ONE*, vol. 18, 3 2023.

[21] L. Scheys, A. Van Campenhout, A. Spaepen, P. Suetens, and I. Jonkers, "Personalized MR-based musculoskeletal models compared to rescaled generic models in the presence of increased femoral anteversion: Effect on hip moment arm lengths," *Gait and Posture*, vol. 28, no. 3, pp. 358–365, 2008.

[22] L. Scheys, K. Desloovere, P. Suetens, and I. Jonkers, "Level of subject-specific detail in musculoskeletal models affects hip moment arm length calculation during gait in pediatric subjects with increased femoral anteversion," *Journal of Biomechanics*, vol. 44, pp. 1346–1353, 4 2011.

[23] M. Wesseling, L. Bosmans, C. Van Dijck, J. Vander Sloten, R. Wirix-Speetjens, and I. Jonkers, "Non-rigid deformation to include subject-specific detail in musculoskeletal models of CP children with proximal femoral deformity and its effect on muscle and contact forces during gait," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 22, pp. 376–385, 3 2019.

[24] Y. Hiasa, Y. Otake, M. Takao, T. Ogawa, N. Sugano, and Y. Sato, "PUBLISHED IN IEEE TRANSACTIONS ON MEDICAL IMAGING, MONTH 20XX 1 Automated Muscle Segmentation from Clinical CT using Bayesian U-Net for Personalized Musculoskeletal Modeling," tech. rep., 2019.

[25] A. Agosti, E. Shaqiri, M. Paoletti, F. Solazzo, N. Bergsland, G. Colelli, G. Savini, S. I. Muzic, F. Santini, X. Deligianni, L. Diamanti, M. Monforte, G. Tasca, E. Ricci, S. Bastianello, and A. Pichiecchio, "Deep learning for automatic segmentation of thigh and leg muscles," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 35, pp. 467–483, 6 2022.

[26] R. Ni, C. H. Meyer, S. S. Blemker, J. M. Hart, and X. Feng, "Automatic segmentation of all lower limb muscles from high-resolution magnetic resonance imaging using a cascaded three-dimensional deep convolutional neural network," *Journal of Medical Imaging*, vol. 6, p. 1, 12 2019.

[27] C. Yan, J. J. Lu, K. Chen, L. Wang, H. Lu, L. Yu, M. Sun, and J. Xu, "Scale- and Slice-aware Net (S2aNet) for 3D segmentation of organs and musculoskeletal structures in pelvic MRI," *Magnetic Resonance in Medicine*, vol. 87, pp. 431–445, 1 2022.

[28] E. Goutham, S. Vasamsetti, P. Kishore, and H. Sardana, "Automatic localization of landmarks in cephalometric images via modified U-Net," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–6, July 2019.

[29] A. Seth, M. Sherman, J. A. Reinbolt, and S. L. Delp, "OpenSim: A musculoskeletal modeling and simulation framework for *in silico* investigations and exchange," *Procedia IUTAM*, vol. 2, pp. 212–232, Jan. 2011.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, Springer Verlag, 2015.

[31] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, Feb. 2021.

[32] M. Wesseling, F. De Groote, C. Meyer, K. Corten, J.-P. Simon, K. Desloovere, and I. Jonkers, "Subject-specific musculoskeletal modelling in patients before and after total hip arthroplasty," vol. 19, no. 15, pp. 1683–1691.

[33] M. Wesseling, S. V. Rossom, i. jonkers, and C. R. Henak, "Subject-specific geometry affects acetabular contact pressure during gait more than subject-specific loading patterns," vol. 22, no. 16, pp. 1323–1333.

[34] E. Montefiori, B. M. Kalkman, W. H. Henson, M. A. Paggiosi, E. V. McCloskey, and C. Mazzà, "MRI-based anatomical characterisation of lower-limb muscles in older women," *PLOS ONE*, vol. 15, p. e0242973, Dec. 2020.

[35] F. C. Anderson and M. G. Pandy, "Individual muscle contributions to support in normal walking," *Gait & Posture*, vol. 17, pp. 159–169, Apr. 2003.

[36] C. Winby, D. Lloyd, T. Besier, and T. Kirk, "Muscle and external load contribution to knee joint contact loads during normal gait," *Journal of Biomechanics*, vol. 42, pp. 2294–2300, Oct. 2009.

[37] K. B. Shelburne, M. R. Torry, and M. G. Pandy, "Contributions of muscles, ligaments, and the ground-reaction force to tibiofemoral joint loading during normal gait," *Journal of Orthopaedic Research*, vol. 24, no. 10, pp. 1983–1990, 2006.

[38] M. Brett, C. J. Markiewicz, M. Hanke, M.-A. Côté, B. Cipollini, D. Papadopoulos Orfanos, P. McCarthy, D. Jarecka, C. P. Cheng, E. Larson, Y. O. Halchenko, M. Cottaar, S. Ghosh, D. Wassermann, S. Gerhard, G. R. Lee, Z. Baratz, B. Moloney, H.-T. Wang, E. Kastman, J. Kaczmarzyk, R. Guidotti, J. Daniel, O. Duek, A. Rokem, M. Scheltienne, C. Madison, A. Sólon, F. C. Morency, M. Goncalves, R. Markello, C. Riddell, C. Burns, J. Millman, A. Gramfort, J. Leppäkangas, J. J. van den Bosch, R. D. Vincent, H. Braun, K. Subramaniam, A. Van, J. H. Legarreta, K. J. Gorgolewski, P. R. Raamana, J. Klug, R. Vos de Wael, B. N. Nichols, E. M. Baker, S. Koudoro, S. Hayashi, B. Pinsard, C. Haselgrove, M. Hymers, O. Esteban, F. Pérez-García, G. Becq, J. Dockès, N. N. Oosterhof, B. Amirbekian, H. Christian, I. Nimmo-Smith, L. Nguyen, P. Suter, S. Reddigari, S. St-Jean, E. Panfilov, E. Garyfallidis, G. Varoquaux, J. Newton, K. S. Hahn, L. Waller, O. P. Hinds, Sandro, B. Fauber, B. Dewey, F. Perez, J. Roberts, J.-B. Poline, J. Stutters, K. Jordan, M. Cieslak, M. E. Moreno, T. Hrnčiar, V. Haenel, Y. Schwartz, B. C. Darwin, B. Thirion, C. Gauthier, I. Solovey, I. Gonzalez, J. Palasubramaniam, J. Lecher, K. Leinweber, K. Raktivan, M. Calábková, P. Fischer, P. Gervais, S. Gadde, T. Ballinger, T. Roos, V. R. Reddam, and freec84, "Nipy/nibabel: 5.3.1." Zenodo, Oct. 2024.

[39] Delft High Performance Computing Centre (DHPC), "DelftBlue Supercomputer (Phase 2)." https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2, 2024.

[40] "Guide to using nnU-Net with overlapping labels · Issue #653 · MIC-DKFZ/nnUNet." https://github.com/MIC-DKFZ/nnUNet/issues/653.

[41] G. Davico, F. Bottin, A. Di Martino, V. Castafaro, F. Baruffaldi, C. Faldini, and M. Viceconti, "Intra-operator Repeatability of Manual Segmentations of the Hip Muscles on Clinical Magnetic Resonance Images," *Journal of Digital Imaging*, vol. 36, pp. 143–152, 2 2023.

[42] R. H. Jensen and D. T. Davy, "An investigation of muscle lines of action about the hip: A centroid line approach vs the straight line approach," *Journal of Biomechanics*, vol. 8, pp. 103–110, Mar. 1975.

[43] V. Carbone, M. M. van der Krogt, H. F. J. M. Koopman, and N. Verdonschot, "Sensitivity of subject-specific models to Hill muscle–tendon model parameters in simulations of gait," *Journal of Biomechanics*, vol. 49, pp. 1953–1960, June 2016.

[44] J. Zhang, H. Sorby, J. Clement, C. D. L. Thomas, P. Hunter, P. Nielsen, D. Lloyd, M. Taylor, and T. Besier, "The MAP Client: User-Friendly Musculoskeletal Modelling Workflows," in *Biomedical Simulation* (F. Bello and S. Cotin, eds.), (Cham), pp. 182–192, Springer International Publishing, 2014.

6 Appendix

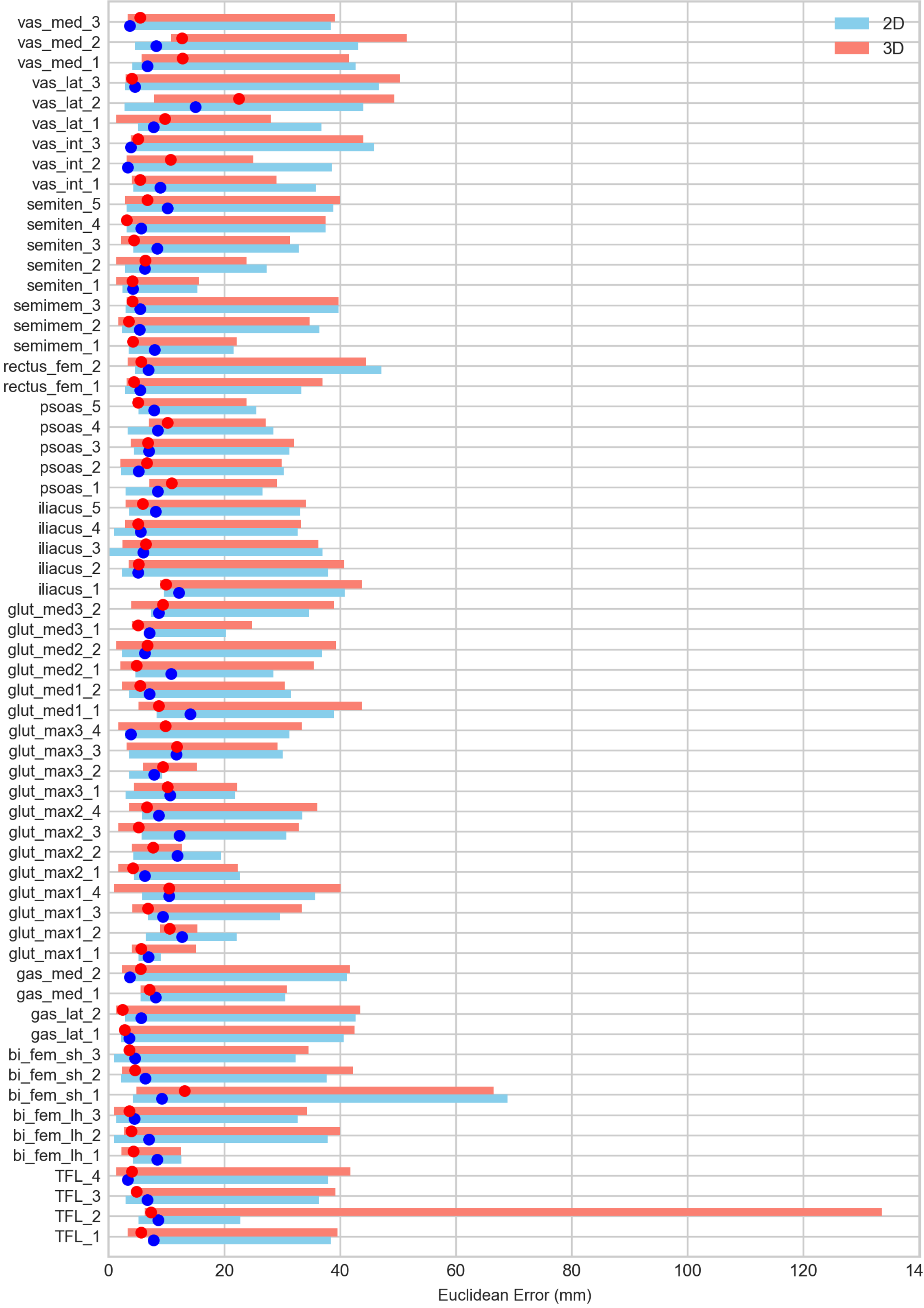

Figure 11: Median Euclidean errors per muscle point (circles) with minimum and maximum range (bars) while including C16.

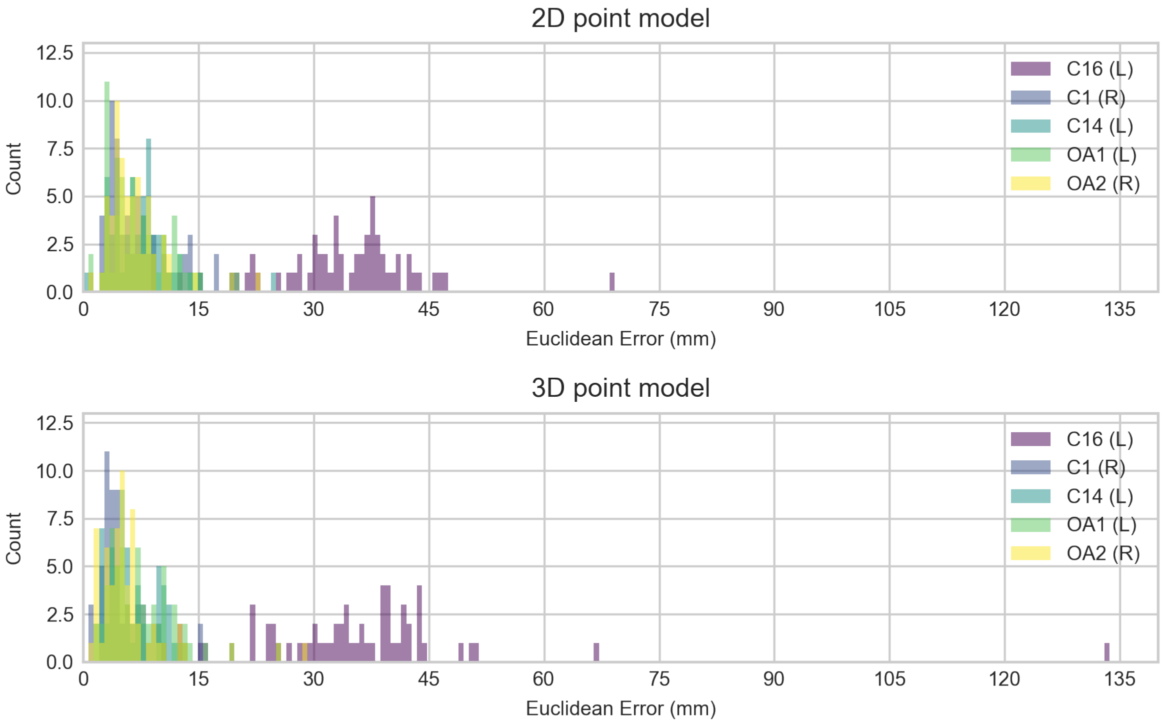

Figure 12: Distributions of Euclidean errors of the 2D (above) and 3D (below) point model on the original validation set.

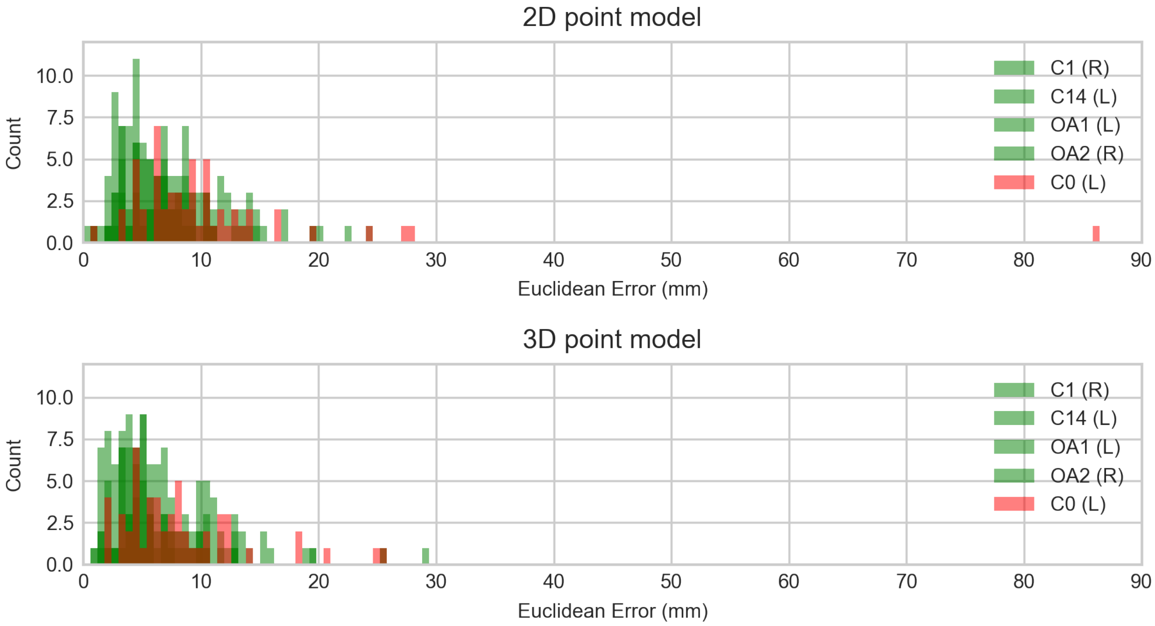

Figure 13: Distributions of Euclidean errors of the 2D (above) and 3D (below) point model on the validation set (green) and the new sample (red).

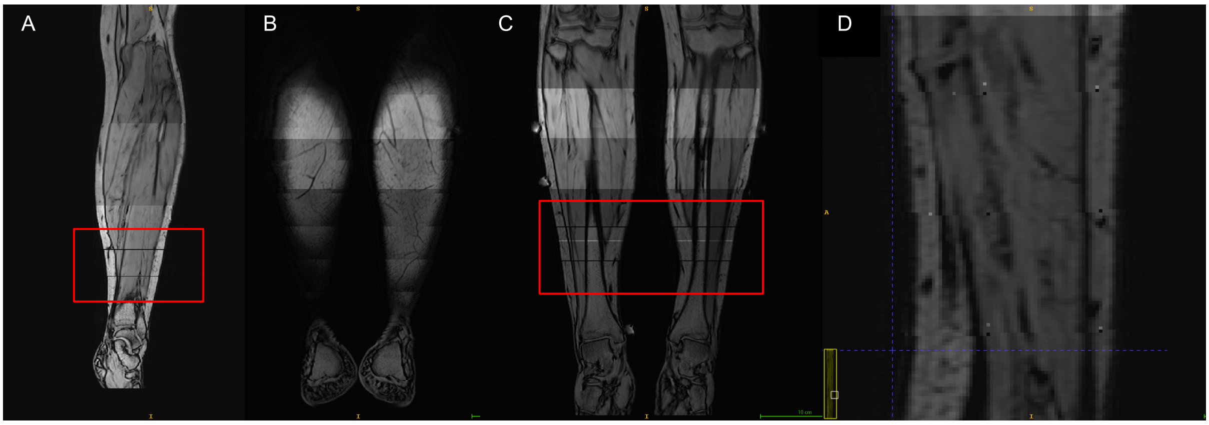

Figure 14: Observed data artefacts after pre-processing of C1 out of the point model dataset. A and B show black lines and 'blockiness' in a coronal view of the shank after pre-processing. C shows the combined image without weighing individual scans and shows also bright lines. D shows a sagittal view of the shank where bright and black voxel can be seen. In B and C, both legs are shown for illustration purposes, unilateral samples were used for training and validation.

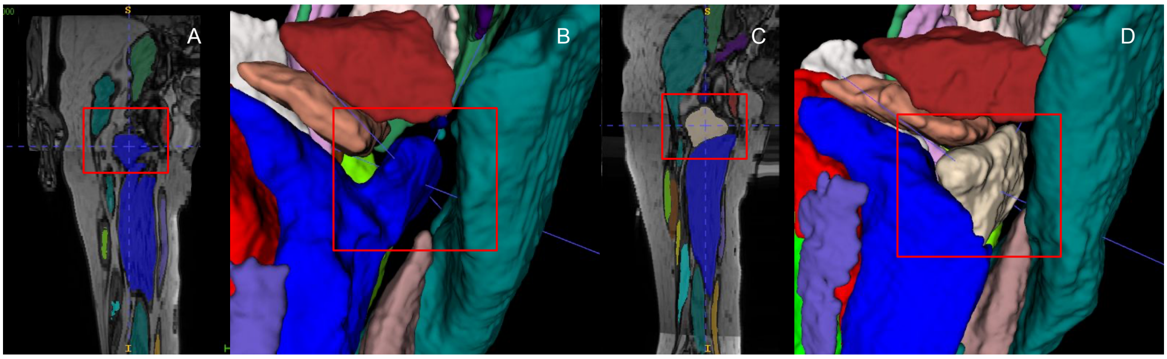

Figure 15: Erroneous segmentation of quadratus femoris in Henson et al (2023) dataset [20]. A and B show sample 1 and C and D show sample 2. A and C show a coronal view and B and D show a 3D view.

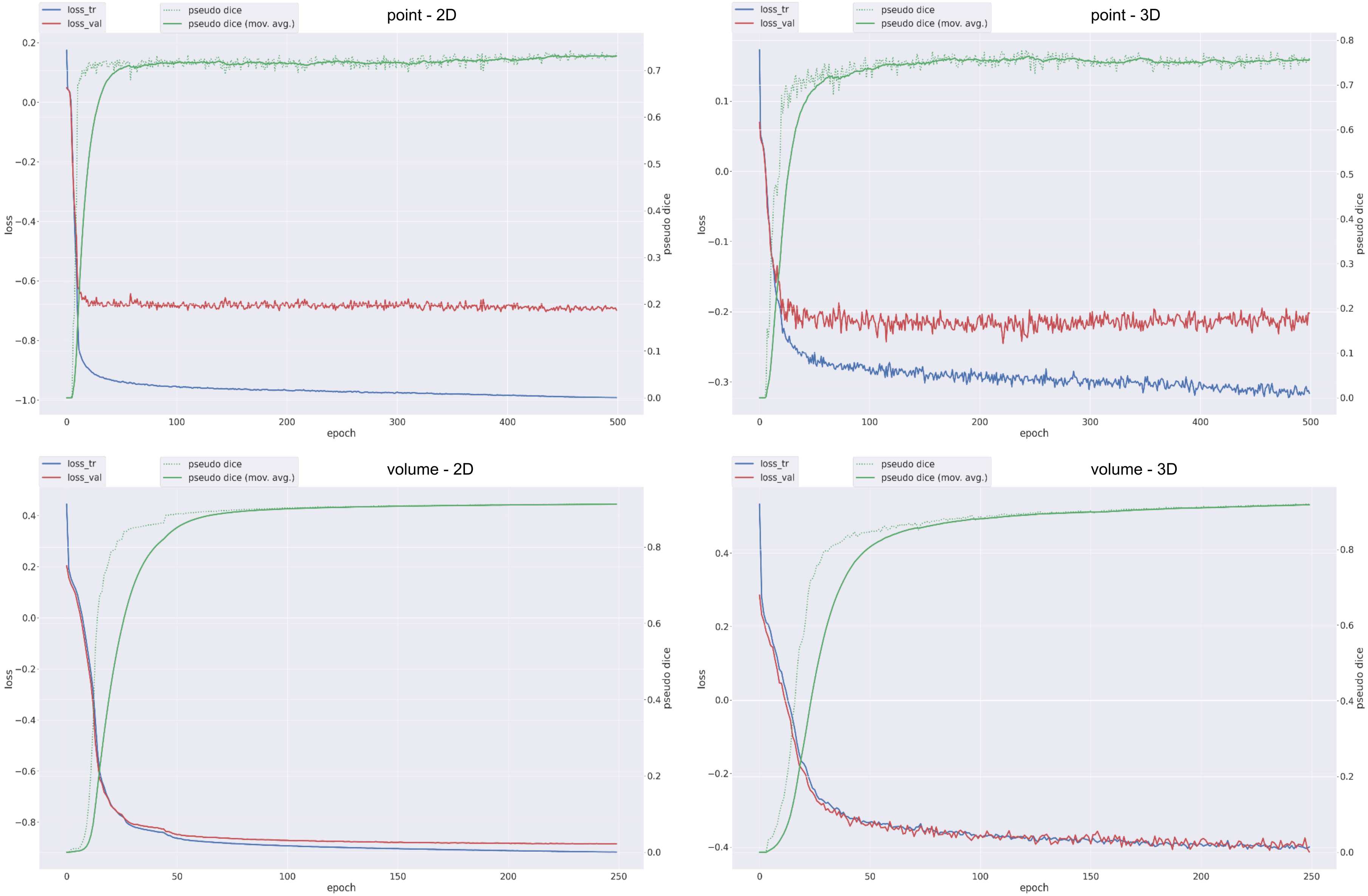

Figure 16: Example training progression curves of the point and volume model. All graphs show the progressions of the first fold.