Machine learning-driven Heckmatt grading in facioscapulohumeral muscular dystrophy: A novel pathway for musculoskeletal ultrasound analysis

Clinical Neurophysiology 172 (2025) 61–69

Contents lists available at ScienceDirect

Clinical Neurophysiology

journal homepage: www.elsevier.com/locate/clinph

Machine learning-driven Heckmatt grading in facioscapulohumeral muscular dystrophy: A novel pathway for musculoskeletal ultrasound analysis

Francesco Marzola [a,*], Nens van Alfen [b], Jonne Doorduin [c], Kristen M. Meiburger [a]

[a] Biolab, Polito[BIO]Med Lab, Department of Electronics and Telecommunications, Politecnico di Torino, Turin Italy
[b] Department of Neurology, Clinical Neuromuscular Imaging Group, Donders Institute for Brain Cognition and Behavior, Radboud University Medical Center, Nijmegen, The Netherlands
[c] Department of Intensive Care Medicine, Radboud University Medical Center, Nijmegen, The Netherlands

ARTICLE INFO

Keywords:
Muscle ultrasound
Machine learning
Muscle segmentation
Heckmatt grading
Neuromuscular disease diagnosis

ABSTRACT

Objective: This study introduces a machine learning approach to automate muscle ultrasound analysis, aiming to improve objectivity and efficiency in segmentation, classification, and Heckmatt grading.
Methods: We analyzed a dataset of 25,005 B-mode images from 290 participants (110 FSHD patients) acquired using a single Esaote ultrasound scanner with a standardized protocol. Manual segmentation and Heckmatt grading by experienced observers served as ground truth. K-Net was utilized for simultaneous muscle segmentation and classification. Heckmatt scoring was approached with texture analysis, using a modified scale with three classes (Normal, Uncertain, Abnormal). Radiomics features were extracted using PyRadiomics and automatic scoring was performed using XGBoost, incorporating explainability through SHAP analysis.
Results: K-Net demonstrated high accuracy in skeletal muscle classification and segmentation, with Intersection over Union ranging from 73.40 to 74.03 across folds. Heckmatt's grading achieved an Area Under Curve of 0.95, 0.87, and 0.97 for classes Normal, Uncertain, and Abnormal. SHAP analysis highlighted histogram-based features as critical for visual scoring.
Conclusion: This study proposes and validates an automatic pipeline for muscle ultrasound analysis, leveraging machine learning for segmentation, classification, and quantitative Heckmatt grading.
Significance: Automating the visual assessment of muscle ultrasound images improves the objectivity and efficiency of muscle ultrasound, supporting clinical decision-making.

1. Introduction

Muscle ultrasound (MUS) using B-mode imaging has emerged as a valuable tool to study the morphology and composition of muscle. Its capacity to evaluate various properties of muscle tissue, not limited to echogenicity but also including muscle texture changes, atrophy, vascularization, and elasticity, is beneficial in diagnosing neuromuscular diseases (Mah and van Alfen, 2018). Furthermore, it offers the advantage of visualizing dynamic muscle movements like fasciculations and fibrillations, enhancing its sensitivity in detecting conditions such as amyotrophic lateral sclerosis (Arts et al., 2012). The assessment of ultrasound images has traditionally involved visual or semi-quantitative analysis, exemplified by the Heckmatt scale (Heckmatt et al., 1982), which classifies muscles based on their appearance to infer muscle condition. An example of this classification is available in Fig. 1. Although visual grading is time-efficient as it can be performed online and offline during muscle ultrasound scanning, it still requires considerable time and expertise as its effectiveness is critically dependent on observer experience. Variations in normal muscle appearance between different muscle groups with sex, age, BMI, and across different ultrasound platforms make it difficult to define a precise transition from normal to abnormal (Wijntjes et al., 2022). Quantitative muscle

Abbreviations: MUS, Muscle Ultrasound; FSHD, Facioscapulohumeral Muscular Dystrophy; QMUS, Quantitative Muscle Ultrasound; EI, Echointensity; ROI, Region of Interest; CSA, Cross-Sectional Area; SHAP, Shapley Additive exPlanations; IoU, Intersection over Union.

* Corresponding author at: Politecnico di Torino, corso Duca degli Abruzzi 24, 10129, Torino, Italy.
E-mail addresses: Francesco.marzola@polito.it (F. Marzola), nens.vanalfen@radboudumc.nl (N. van Alfen), jonne.doorduin@radboudumc.nl (J. Doorduin), kristen.meiburger@polito.it (K.M. Meiburger).

https://doi.org/10.1016/j.clinph.2025.01.016
Accepted 31 January 2025
Available online 6 February 2025

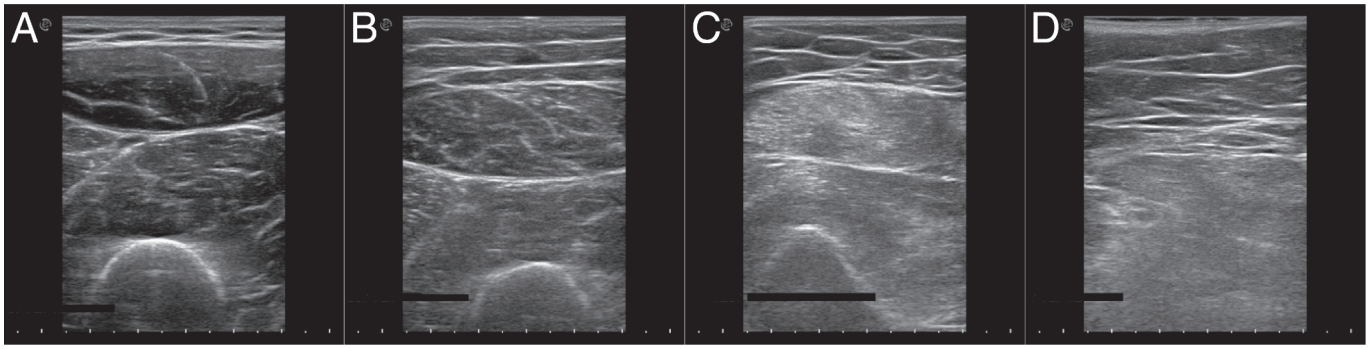

Fig. 1. Heckmatt grades. Examples from Rectus Femoris muscle. A) Heckmatt score 1(Normal), Normal muscle. B) Heckmatt score 2 (Uncertain), increased muscle echo intensity with distinct bone echo. C) Heckmatt score 3(Abnormal), marked increased muscle echo intensity with a reduced bone echo. D) Heckmatt score 4 (Abnormal), very strong muscle echo and complete loss of bone echo.

ultrasound (QMUS) methods bring a more objective examination of muscle tissue, expanding beyond the capabilities of conventional ultrasound. It is often reported in terms of z-scores, calculated by measuring the echointensity (EI) inside the cross-sectional area of each muscle and comparing the value with a cohort of reference subjects (van Alfen and Mah, 2018). The clinical adoption of QMUS is challenged by the dependency on the specifics of ultrasound machines for calibration and reference values and on subjects' characteristics for result interpretation, which complicates standardization and comparison of results across different devices and settings (Nijboer-Oosterveld et al., 2011; Pillen and van Alfen, 2015). As evidenced in (Vincenten et al., 2024), QMUS brings information complementary to Heckmatt grading. While both visual and quantitative analysis can track the early stage of conditions like facioscapulohumeral muscular dystrophy (FSHD), the Heckmatt score is more suitable for differentiating late-stage dystrophic muscles with severe fat replacement. This discrepancy shows differently between muscles, showing an agreement between the z-scores and Heckmatt grades ranging from 82.2 % for the tibialis anterior to 100 % for the biceps brachii.

Notably, advancements in QMUS include texture analysis, incorporating first-order features such as muscle EI and more complex analyses based on higher-order texture features (Paris and Mourtzakis, 2021). First-order features of ultrasound images reflect muscle composition, indicating the proportions of adipose and connective tissue infiltration

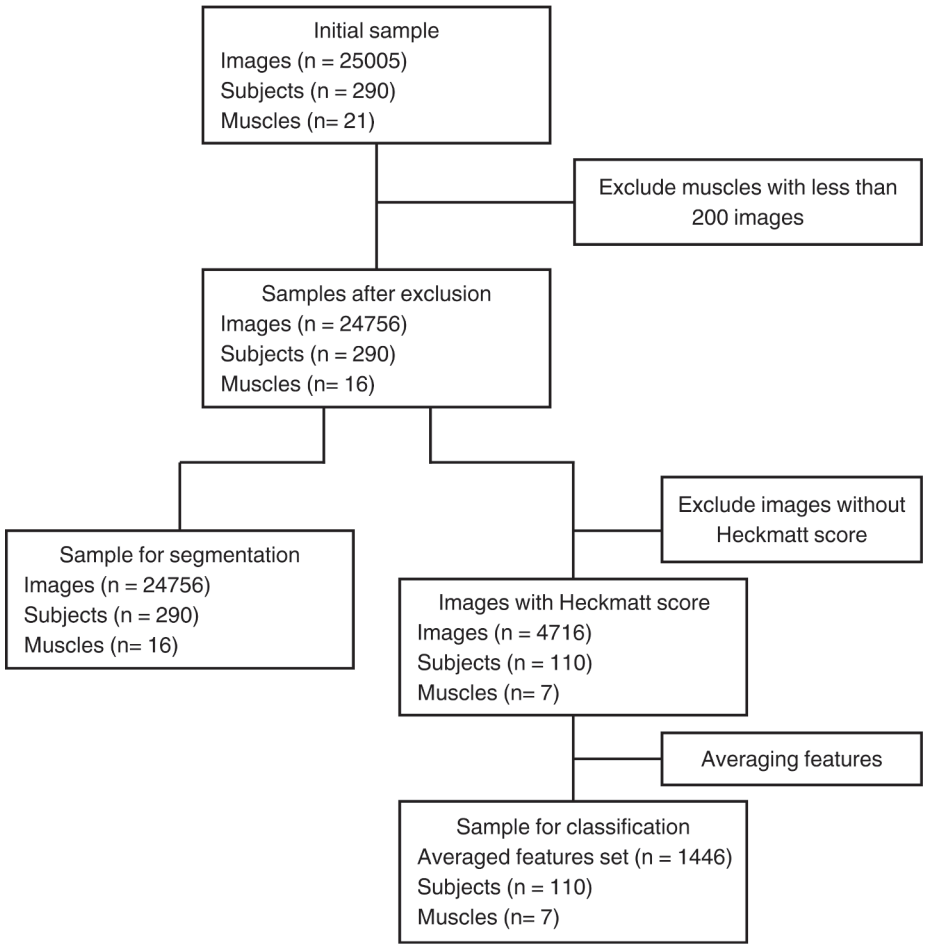

Fig. 2. Sample selection flow diagram.

within the muscle. These features focus on the intensity values of individual pixels, neglecting the spatial relationships between the pixels (Sahinis and Kellis, 2023). Techniques based on second-order texture features, such as the Haralick (Haralick et al., 1973) or Galloway (Galloway, 1975) features enable a deeper evaluation of muscle tissue by analyzing the spatial variation in pixel intensities to identify patterns and pixel interrelationships in an ultrasound image and are more robust to changes in dynamic range (Seoni et al., 2023).

The efficacy of higher-order features in distinguishing between muscles of different architectural characteristics and their potential in monitoring disease diagnosis and progression have been highlighted in several studies (Martínez-Payá et al., 2018; Molinari et al., 2015; Watanabe et al., 2017). However, the lack of a standardized extraction method for these features limits the comparability of results between different studies (Paris and Mourtzakis, 2021). Hence using standardized extraction methods like Pyradiomics (Van Griethuysen et al., 2017) is important to quantitatively characterize muscle phenotypes (Mirón-Mombiela et al., 2023).

To perform a quantitative analysis of muscle ultrasound, the first step is to define the region of interest (ROI) as to where to measure image features. Automating this segmentation step could reduce technicians' and physicians' workloads, and reduce variability related to the different readers for large-scale studies. The unique challenges of muscle segmentation, including heterogeneous textures, variable fiber orientations, and dynamic changes, necessitate adaptable algorithms specific to each application and muscle group. Previous works focused on a few muscle groups and normal muscles (Caresio et al., 2017; Marzola et al., 2021; Salvi et al., 2019), making them not feasible in clinical settings where a global assessment of the subject with sometimes severely distorted muscle architecture is required. For diseases like FSHD, different muscles can be affected at different rates between patients, hence automated segmentation models have to deal with multiple muscles and conditions.

This paper aims to develop an automatic pipeline for segmenting and classifying 16 muscles in transverse B-mode ultrasound images and to quantitatively assess muscle composition. We hypothesize that integrating deep learning segmentation with radiomics-based classification can enhance the Heckmatt score's objectivity and explainability. This work could support clinicians by providing quantitative and reproducible muscle evaluations, thus advancing the diagnostic capabilities of MUS in neuromuscular disease analysis. The code and data used in this study are openly accessible at https://doi.org/10.17632/yzg86vb895.1.

2. Materials and methods

The dataset constructed for this retrospective study encompasses 25,005 B-mode images acquired from 290 participants (180 healthy, 110 from an FSHD cohort) across 21 distinct muscles, with three acquisitions for each muscle. The muscle ultrasound procedures were carried out using an Esaote MyLabTwice ultrasound scanner (Esaote SpA, Genoa, Italy) equipped with a LA533 3–13 MHz linear transducer with an axial resolution of approximately 0.2 mm. To ensure uniformity in the data collection process, system presets were consistently maintained for all measurements. A precise description of the acquisition process is available in (Mah and van Alfen, 2018). Data from the FSHD patients has been used in (Vincenten et al., 2024).

The original dataset was cleaned to include only muscles represented by at least 200 images to mitigate dataset imbalance. As a result, images of 5 muscles were excluded resulting in a total of 24,756 images of 16 muscles that were used for developing the muscle segmentation and classification model. A subset was extracted to develop the quantitative radiomics-based approach for Heckmatt scoring (n = 1466 images acquired on 110 FSHD patients). Fig. 2 shows the flow diagram describing data management. This study uses images collected as part of regular clinical practice and anonymized before sharing among participating centers, hence does not require additional ethical approval.

Table 1
Subjects' characteristics. Values are displayed as mean +/- standard deviation.

Characteristic	Value
Total count	110
Sex (0: Male, 1: Female)	58 (0) 52 (1)
Age (years)	52.71 +/- 14.25
Weight (kg)	79.77 +/- 13.57
Height (m)	1.76 +/- 0.09
BMI	25.82 +/- 3.91
Heckmatt class	502 (Normal), 532 (Uncertain), 432 (Abnormal)

2.1. Manual evaluation and clinical data

For the muscle segmentation and classification task, the images were manually segmented and labeled during routine clinical assessments by neurodiagnostic technicians and used as the ground truth for the deep learning approach. A strict acquisition protocol was followed to ensure consistency across operators.

For the quantitative radiomics-based strategy, the Heckmatt semi-quantitative grading was performed by an experienced observer for the 110 patients of the FSHD cohort and 7 different muscles. Characteristics of these subjects are provided in Table 1. The evaluation of the muscle ultrasound images involved a visual inspection using the 4-point Heckmatt scale, without prior knowledge of the subjects' clinical history, EMG, earlier ultrasound/EMG examinations, or laboratory findings. As quantitative analysis, these images were manually segmented, and the gray level was measured by averaging the pixel intensities. The measurements were then compared against expected values derived from the subjects' age, sex, and BMI to calculate a z-score.

2.2. Deep learning simultaneous segmentation and classification

The muscle segmentation task becomes more challenging for automatic algorithms when different muscles are present in the dataset, due to the diverse textures and orientations of muscle fibers, as well as the heterogeneity in their surrounding anatomy. To overcome these challenges, a two-step approach using the K-Net architecture was employed (Salvi et al., 2024; Zhang et al., 2021). First, a multi-class K-Net identifies the muscle group present in the image, enabling automated labeling for large datasets where muscle labels might be missing. This global classification step ensures scalability and autonomy in retrospective analyses. Second, we fine-tuned a muscle-specific K-Net for each identified muscle to refine segmentation, reducing variability caused by different muscle properties and further improving accuracy.

This setup was chosen since the K-Net architecture dynamically modifies conditional kernels according to the category of each identified object (i.e., muscle) within the image. This adaptation facilitates more accurate mask generation since each kernel can be tuned on a specific muscle. A graphical representation of the key components of the chosen architecture available in Fig. 3. The first network outputs the pixel-by-pixel classification distinguishing between the background and each muscle group for a total of 17 classes. The muscle group in the image is therefore known and the second network outputs a refined binary mask of the muscle. The final output is the classification of the muscle group and the binary mask of the detected muscle.

For the networks training and validation, the dataset was divided into five cross-validation folds. Images were normalized to the [-1,1] range and real-time augmentations were employed. The real-time augmentations included random rotations (up to ± 15 degrees), horizontal and vertical flips, scaling (between 0.9 and 1.1), and brightness adjustments (±10 %). For the first network, each run consisted of 80,000 iterations, utilizing the AdamW optimizer and a MultiStep Learning Rate learning strategy. The batch size was 2 with an image resolution of 512x512. For the fine-tuning of the second network, only 20,000 iterations were performed with the same parameters. The CrossEntropyLoss was selected for loss computation, assigning weights to each class to

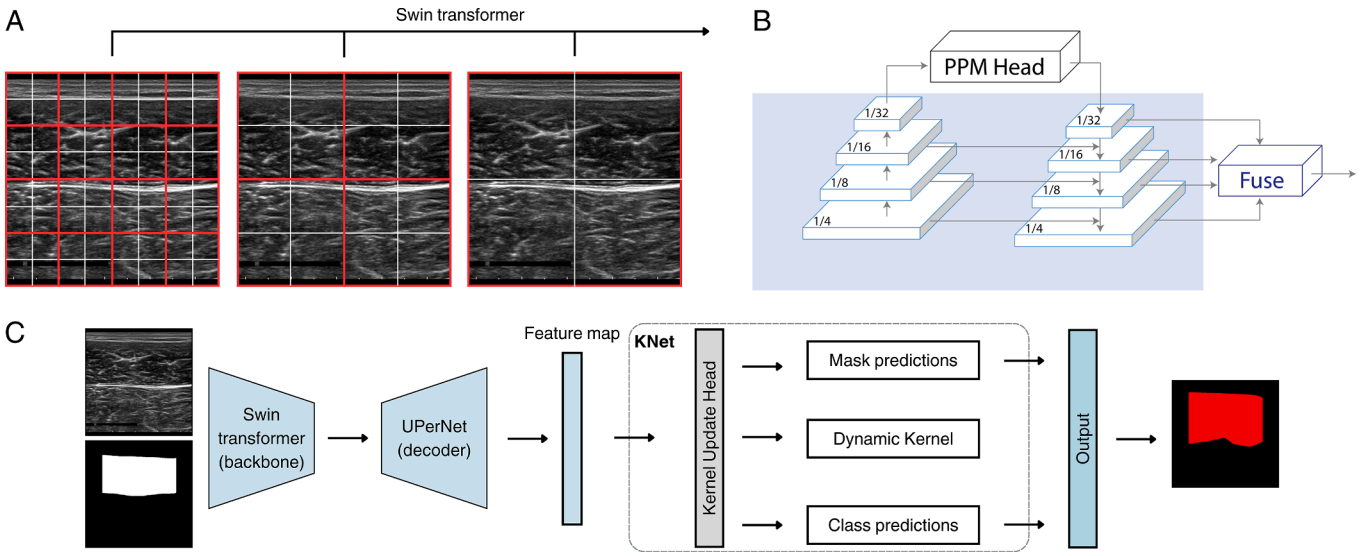

Fig. 3. Segmentation framework used in this work. (A) Features are extracted using a Swin transformer as the backbone. (B) Multiscale feature aggregation is performed by UPerNet (Unified Perceptual Parsing for Scene Understanding) using a Pyramid Pooling Module (PPM). (C) The overall architecture of the K-Net. During the inference process, the network generates a softmax output that indicates the probability of each pixel belonging to one of the muscle or background classes.

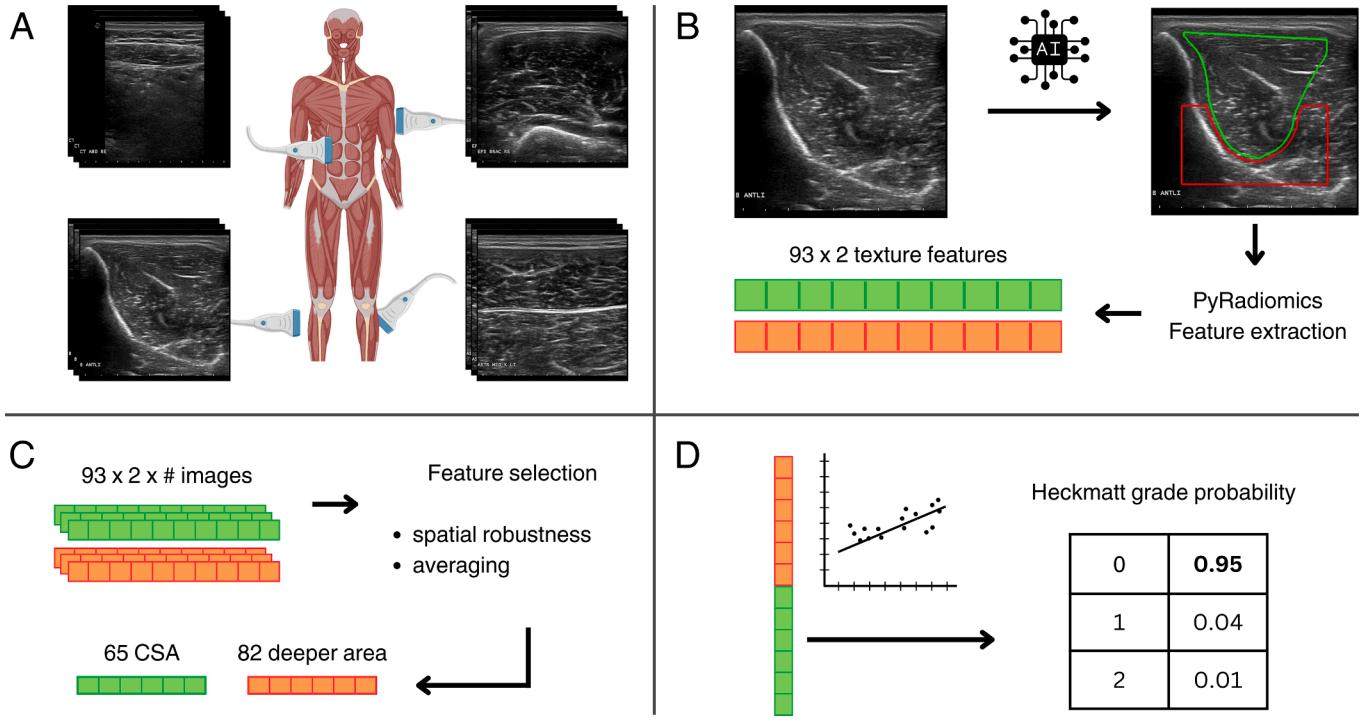

Fig. 4. Project pipeline. A) Three bilateral acquisitions for each muscle. Examples from Rectus Abdominis, Biceps Brachii, Tibialis Anterior, and Gastrocnemius Medialis B) The segmentation networks perform muscle segmentation and classification, features are extracted from the Cross Sectional Area (CSA) in green, and the area deeper to the muscle in red, using PyRadiomics. C) Feature dimensionality is reduced considering spatial robustness. The remaining features are averaged. D) The classification model predicts the probability of the modified Heckmatt class.

address class imbalance. These weights were calibrated inversely proportional to the pixel frequency of each class using a logarithmic scale to lessen the effects of drastic frequency disparity. During training, the mean Intersection over Union (mIoU) was measured performing inference on the validation set every 8000 iterations (2000 during finetuning). The best-performing model was saved.

2.3. Post-processing and muscle classification

Post-processing of the first network output includes three steps: initially, the convex hull for all detected objects is generated. When objects of different classes overlap with an IoU greater than 0.6, they are merged, favoring the class that occupies the largest area. Then, only the largest connected component for each class is retained, eliminating smaller disconnected objects. The final step determines the predominant

Table 2
Segmentation and classification metrics by muscle. Values are displayed as mean +/- standard deviation. IOU (Intersection over Union).

Muscle type	Classification accuracy (%)	Segmentation failures (%)	IOU	Precision	Recall	Support
Biceps brachii	99.54	0	0.85 +/- 0.11	0.89 +/- 0.11	0.95 +/- 0.08	1735
Deltoideus	99.81	0,29	0.88 +/- 0.09	0.90 +/- 0.09	0.97 +/- 0.07	1029
Depressor anguli oris	83.83	16,67	0.57 +/- 0.28	0.63 +/- 0.30	0.75 +/- 0.34	1258
Digastricus	98.22	1,38	0.78 +/- 0.13	0.84 +/- 0.13	0.92 +/- 0.13	1440
Gastrocnemius medial head	99.66	0,74	0.91 +/- 0.10	0.94 +/- 0.09	0.97 +/- 0.10	1747
Geniohyoideus	98.64	1,23	0.77 +/- 0.13	0.84 +/- 0.14	0.90 +/- 0.14	738
Masseter	99.59	0,41	0.87 +/- 0.09	0.90 +/- 0.10	0.96 +/- 0.07	1478
Mentalis	98.95	17,93	0.45 +/- 0.23	0.55 +/- 0.28	0.64 +/- 0.31	1243
Orbicularis oris	75.84	25,21	0.44 +/- 0.28	0.52 +/- 0.33	0.62 +/- 0.37	1584
Rectus abdominis	98.96	0,69	0.85 +/- 0.11	0.89 +/- 0.11	0.95 +/- 0.10	1732
Rectus femoris	97.93	0,29	0.88 +/- 0.10	0.90 +/- 0.10	0.97 +/- 0.05	1737
Temporalis	99.8	2,58	0.77 +/- 0.22	0.81 +/- 0.22	0.93 +/- 0.14	1471
Tibialis anterior	99.88	0,23	0.78 +/- 0.15	0.83 +/- 0.15	0.93 +/- 0.09	1736
Trapezius	99.87	0,66	0.80 +/- 0.14	0.86 +/- 0.13	0.93 +/- 0.12	1522
Vastus lateralis	98.61	0,58	0.88 +/- 0.12	0.91 +/- 0.11	0.96 +/- 0.10	1725
Zygomaticus	99.73	32,89	0.40 +/- 0.30	0.47 +/- 0.34	0.58 +/- 0.40	2571

label within the refined predictions, assigning to the image the class of the most prevalent label in the segmentation. Consequently, the network's output is a classification of the image based on the dominant label and the segmentation maps for one or more muscles. For the second network, only the elimination of all the objects smaller than the largest connected component is applied.

2.4. Radiomics-based quantitative Heckmatt scoring

For the quantitative Heckmatt scoring strategy, texture analysis is performed utilizing PyRadiomics after the muscle is segmented. The analysis is firstly done on the Cross-Sectional Area (CSA) delineated by the segmentation map, extracting 93 texture features across various groups: first-order, Gray Level Co-occurrence Matrix (GLCM) (Haralick et al., 1973), Gray-Level Run-Length Matrix (GLRLM) (Galloway, 1975), Gray-Level Size-Zone Matrix (GLSZM) (Thibault et al., 2009), Gray-Level Dependence Matrix (GLDM) (Sun and Wee, 1982), and Neighbouring Gray-Tone Difference Matrix (NGTDM) (Amadasun and King, 1989). During manual Heckmatt scoring, the physician looks for visual clues outside the CSA, such as the presence of bone or other structures (Heckmatt et al., 1982). Therefore, the deeper part of the image to the CSA is also considered and a separate feature extraction is performed in this area. A visualization of this process is shown in Fig. 4B with the CSA of the muscle in green and the deeper part in red. The deeper part is defined in its top limit by the centroid of the CSA lowered by 10 % of image height and by the CSA, while the lateral and lower limits are defined by the 15 % of image width and height.

An XGBoost classifier was trained to classify the images into three classes, as a modification of the original Heckmatt score that has more equal distances between categories (Lagarde et al., 2023; Wijntjes et al., 2024): Normal (Heckmatt 1, 502 images), Uncertain (Heckmatt 2, 532 images), and Abnormal (Heckmatt 3 + 4, 432 images). Features were measured for each subject-muscle-side combination for 1466 entries averaging the value on the three acquisitions. Features exceeding 50 % variability relative to their average were labeled as "high variability." Only those identified as high variability in less than 10 % of the dataset (146 total) were retained. This selection ensures that the features reliably reflect muscle composition, unaffected by small changes in measurement conditions like insonation angle, muscle movement, or ROI placement. This resulted in 65 features for the CSA and 82 for the deeper part of the image.

Heckmatt scoring was modeled on this set of features as a classification task using an XGBoost model (Chen and Guestrin, 2016). The classifier was trained using 10-fold cross-validation stratified by the Heckmatt score and grouped by subject. On the training folds, features were standardized, and model-based feature selection was applied. The pipeline was developed using the scikit-learn library (Pedregosa et al., 2011). Explanatory analysis was performed using Shapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017). A graphical representation of the whole pipeline is illustrated in Fig. 4.

2.5. Performance validation

Muscle classification and segmentation were assessed on the union of the 5 held-out folds, each evaluated using the model trained on the respective training set. The Intersection over Union (IoU) was used to assess segmentation and Precision and Recall were analyzed to discern the model's tendency to under- or over-segment. To evaluate the muscle classification accuracy, a confusion matrix and its associated metrics were calculated.

The Heckmatt scoring performance was assessed by computing the predictions on the held-out folds at the end of each fold repetition. To test the accuracy in predicting each class, ROC curves in One vs Other (OvO) and One vs Rest (OvR) configurations were calculated. Accuracy and Cohen's Kappa scores were measured by selecting the class with the highest probability. An ablation study to assess the contribution of measuring texture features on the lower part of the image was performed. Finally, the correlation of human and computerized Heckmatt scores with manual z-scores was checked using Spearman-ranked correlation analysis to assess that the two scoring techniques follow the same relationship with QMUS.

3. Results

The segmentation outcomes vary among different types of muscles. As shown in Table 2, larger skeletal muscles that typically feature distinct fascial structures exhibit superior segmentation accuracy compared to the (much) smaller facial muscles. Recall is consistently higher than precision for all the muscles, showing how the network tends to over-segment the CSA, favoring having fewer false negative pixels at the cost of having more false positives. Facial muscles remained challenging. Fine-tuning muscle-specific K-Nets improved IoU by ~ 5 % for such difficult cases compared to a single multi-class model alone, confirming the benefit of muscle-specific refinement. Comparison between multi-class and muscle specific models is available in supplementary materials, Fig. S1.

Segmentation failures were defined as muscles with an IoU lower than 0.2 and the only muscles with more than 5 % of failures were facial muscles. In Fig. S2 the robustness of segmentation performance to sampling is shown. The accuracy of the muscle group classification task is available in Table 2. In Supplementary material, Fig. S3 shows the normalized confusion matrix. For all the muscles the developed method showed an impressive performance, except for the depressor anguli oris and the orbicularis oris. These two muscles were often confused by the

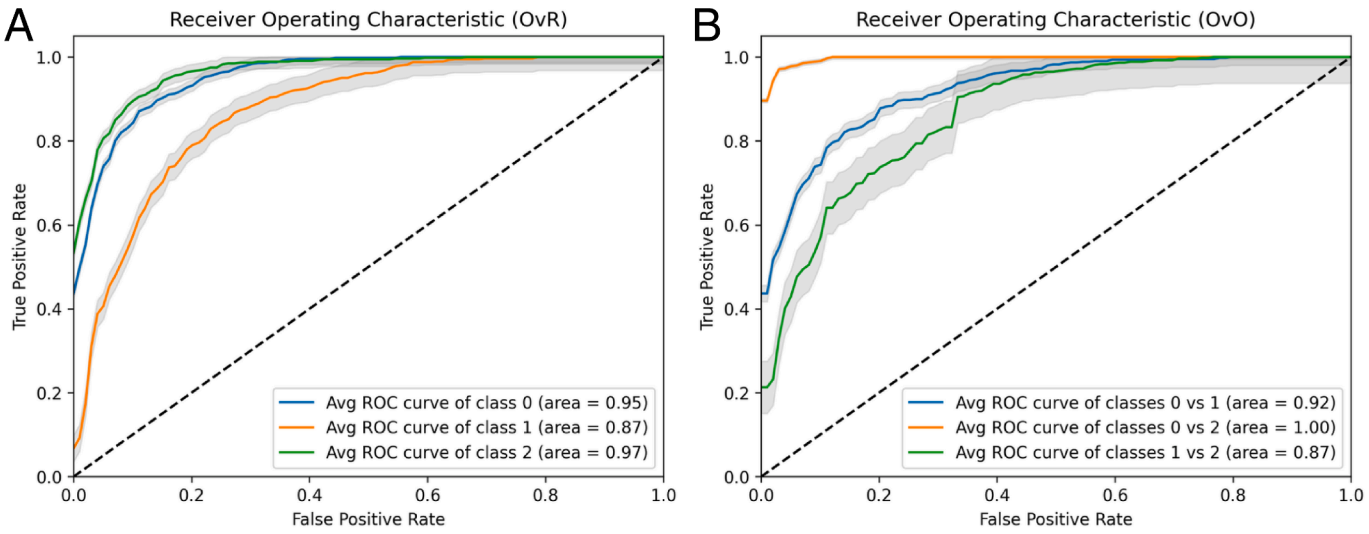

Fig. 5. A) ROC One-vs-Rest approach. B) ROC One-vs-One approach. Class 0: Normal, Class 1: Uncertain, Class 2: Abnormal.

Table 3
Ablation Study on feature extraction location. The number of features refers to the features selected during the fitting of the Heckmatt classification model. In the first row, the number in the round bracket is the number of features measured inside the CSA (Cross Sectional Area). Lower indicates the features extracted from the part of the image deeper with respect to CSA. AUC (Area Under Curve).

				AUC (One-vs-One)			AUC (One-vs-Rest)		
	# features	Accuracy	Kappa	01	02	12	0	1	2
CSA + Lower	74 (33)	0.80	0.71	0.92	1.00	0.87	0.95	0.87	0.97
CSA	33	0.78	0.67	0.91	1.00	0.85	0.95	0.85	0.96
Lower	41	0.70	0.54	0.83	0.97	0.79	0.89	0.78	0.93

Table 4
Spearman correlation between echointensity in the cross-sectional area of the muscle expressed as z-score and manual or automatic Heckmatt grading.

	Biceps brachii	Gastrocnemius medialis	Rectus abdominis	Rectus femoris	Tibialis anterior	Trapezoid	Vastus lateralis
Manual	0.79	0.69	0.49	0.76	0.78	0.66	0.71
Automatic	0.81	0.75	0.43	0.80	0.79	0.70	0.72

segmentation network, which can be understood from the similar appearance of both these muscles and the surrounding anatomy. Across folds, the weighted classification accuracy average ranged from 0.963 to 0.968, showing great robustness to dataset sampling. Automatic labeling ensures scalability for large datasets or retrospective analyses where manual labels might be missing or incorrect.

The Heckmatt grading performance by the algorithm is reported in Fig. 5, highlighting that the model's uncertainty lies mostly in discerning between consecutive classes. This directly corresponds to the model's capability to accurately capture the ordered and semi-quantitative nature of the Heckmatt scale, as evidenced by the minimal misclassification between non-consecutive classes ($AUC_{Normal-Abnormal} > 0.99$). ROC curves show very robust performances in evaluating the extreme classes ($AUC_{Normal} = 0.95$, $AUC_{Uncertain} = 0.87$, $AUC_{Abnormal} = 0.97$) and slightly better accuracy in discerning class Normal to Uncertain (AUC = 0.92) than Uncertain to Abnormal (AUC = 0.87). Table 3 shows how adding features from the area below the CSA slightly improves the classification performance ($Kappa_{CSA} = 0.67$, $Kappa_{CSA+LOW} = 0.71$).

The correlation between EI values expressed as Z-scores and Heckmatt scores for each muscle group shows similar values using manual and automatic scoring (Table 4). In Fig. 6, the SHAP analysis quantifies the most important feature's impact on class predictions using Beeswarm plots, with single decision plots further evaluating key features' effects on individual images. For all the classes histogram-based features are among the most important ones. In particular, median EI is the most important feature for assigning classes Normal and Uncertain aligning with conclusions from (Vincenten et al., 2024) where EI is shown to describe early stages of muscle degeneration.

4. Discussion

In this paper, a dataset with more than 25,000 muscle ultrasound images with annotations enabled the development of a deep learning model with high segmentation and classification accuracy for more than 10 different skeletal muscles with diverse muscle conditions. This opens the way for the use of automatic segmentation algorithms for studying the differential involvement of each muscle in muscular dystrophy disease progression.

While the standardized acquisition protocol developed in Radboudumc center further facilitates muscle recognition, we found that segmenting facial muscles posed a significantly greater challenge compared to automated image processing of the larger skeletal muscles. The smaller size and complex entanglement of facial muscles with their surrounding tissue often made them indistinguishable and more difficult to classify and segment accurately. From other work on FSHD (Wijntjes et al., 2024) we know that humans also have low interoperator reproducibility visually scoring the facial muscles and that only manual identification, segmentation, and quantified grayscale analysis produced acceptable results. Fine-tuned segmentation models for each

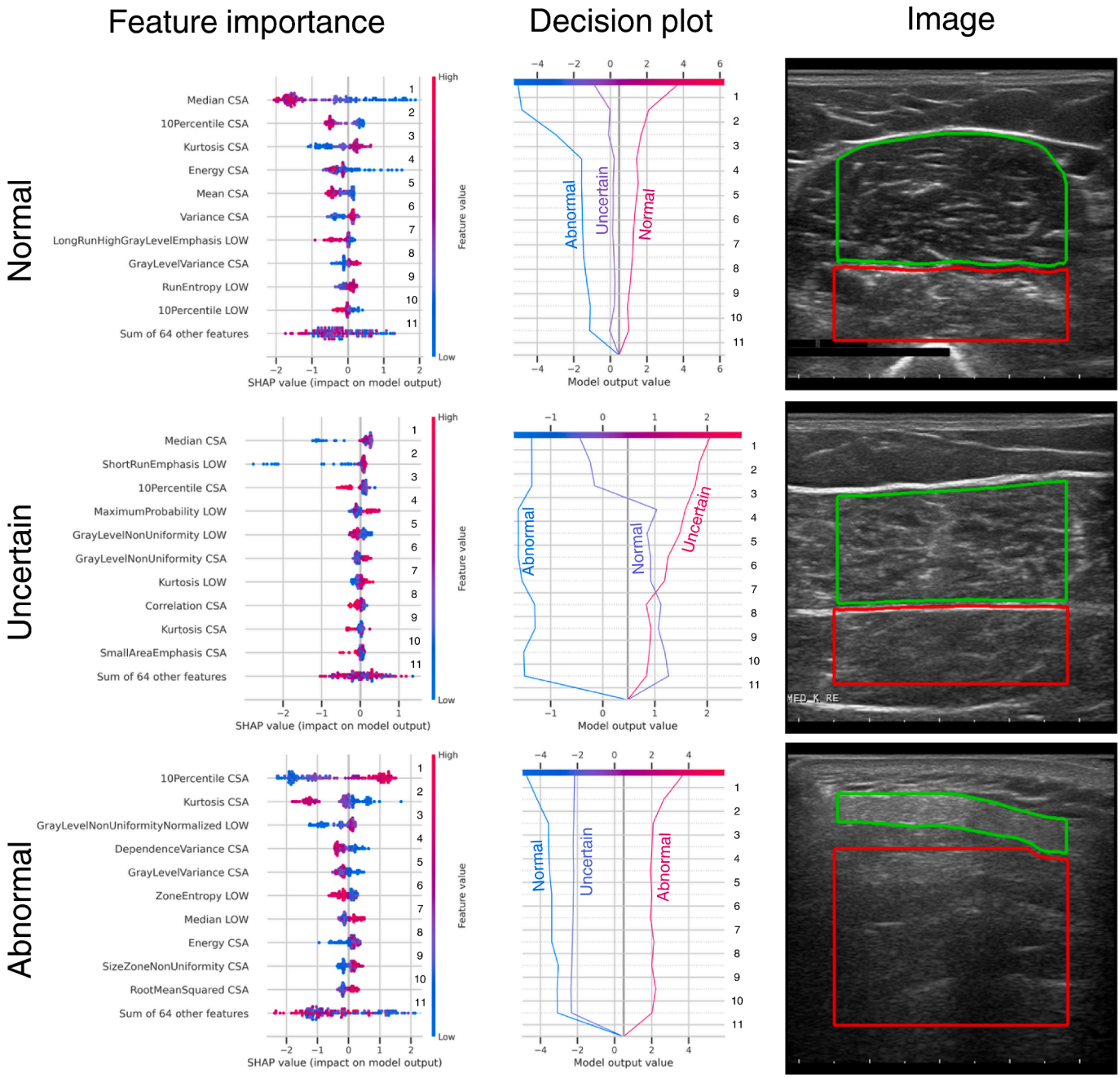

Fig. 6. Explainability analysis of the automatic Heckmatt grading model for three correctly classified images. In the first column, there are the SHAP beeswarm plots showing overall feature importance on the specific class, each point represents a single prediction in the overall dataset. In the central column, there are decision plots for the examples on the right, showing how each feature contributes to the final decision. Features are ordered by importance and are labeled as "CSA" or "LOW" depending on the extraction site (Cross-Sectional Area or the area deeper compared to the muscle). Features impact on the final model decision is displayed in terms of log odds. In the images on the right, in green is the cross-sectional area, and in red is the area below the muscle.

muscle show significant improvement over the multi-label segmentation model. However, the segmentation accuracy for facial muscles is not yet on par with bigger ones. One possible solution would be to process the entire videos of the insonation procedure exploiting multiple frames to have a more robust estimation of muscle area.

While being affected by subjectivity, visual Heckmatt scoring was chosen as ground truth for its wide use in clinical practice and simple interpretation. Other clinical parameters were excluded from the analysis because we wanted to focus on visual assessment without introducing other sources of bias. Subject characteristics will be included when performing a full subject assessment, including muscles based on specific evaluation protocols. ML-based automatic Heckmatt scoring demonstrated high accuracy, suggesting it can quantitatively interpret the Heckmatt scale and link ultrasound findings with more clinical endpoints. Opting for an ML model using PyRadiomics features over a DL approach allowed for the identification of key features. The ML model is trained to replicate human visual scoring. Analysis using SHAP values reveals a predominant reliance on histogram-based features over higher-order texture features. This pattern aligns with known human perceptual tendencies, where simpler object-like stimuli are processed more readily than complex texture-like stimuli. For Uncertain and Abnormal classes, the importance of features from the lower part of the image increases. This is again correlated to clinical interpretation where the presence of bone echo is used to distinguish between Heckmatt

grades. This correspondence between the visual interpretation of the images and their mathematical description is key for developing an AI system that could support clinical decisions. A further development would be integrating a visual representation of the SHAP analysis by combining the feature maps of the features that skew the classification of the image towards a specific class. This could improve the explainability by highlighting the area of the muscle that influenced the most the model prediction.

While having a lower impact on the automatic visual score, higher-order features hold the potential to develop quantitative imaging biomarkers (QIBs) since texture in the image is directly linked to tissue morphology (Mirón-Mombiela et al., 2023). Directly linking features to functional evaluations of muscle condition can merge the benefits of quantitative and visual scores expanding the usability of MUS in the clinical evaluation of muscle condition. Moreover, the manifestation of conditions like muscle fibrosis is yet to be detected using MUS, higher-order features might detect changes that are currently masked by first-order features opening new possibilities for muscle ultrasound. Higher order features also have the potential to enhance the clinical adoption of QMUS. By applying the segmentation and feature extraction method presented in this work to process a multi-device dataset and leveraging our standardized image labeling, we can learn an ultrasound image feature representation that remains robust against device and setting changes. From the SHAP analysis, it is clear how the visual Heckmatt score is mostly explained by lower order features. Since the visual preferences of the sonographer primarily link to lower order features, setting and device selection will mostly affect these lower order features, while higher order features have the potential to more robustly differentiate the various subgroups.

One limitation of the current model arises from the variability in ultrasound machines, which can affect the generalizability of our model trained on a monocentric and standardized dataset. Publishing the trained model weights could facilitate transfer learning, reducing the data requirements for developing new models on different datasets. Additionally, including other neuromuscular disorders beyond FSHD could expand the ML system's applicability, potentially allowing for a visual scoring system that classifies both the type and degree of muscle involvement.

5. Conclusion

This study demonstrates the feasibility of an automatic and reproducible pipeline for muscle classification, segmentation, and quantitative assessment via machine learning in muscle ultrasound for healthy persons and people living with facioscapulohumeral muscular dystrophy. The proposed pipeline offers a way for objective and repeatable muscle tissue evaluation. This approach aligns with physician evaluations and provides a framework for standardizing muscle ultrasound analysis. The promising results, especially in terms of segmentation accuracy and Heckmatt grading performance, highlight the potential of this pipeline to support clinical decision-making and contribute to the advancement of diagnostic procedures in neuromuscular diseases.

6. Data sharing statement

Data generated by the authors or analyzed during the study are available at: Marzola, Francesco (2024), "Machine Learning-driven Heckmatt Grading in facioscapulohumeral muscular dystrophy : A Novel Pathway for Musculoskeletal Ultrasound Analysis", Mendeley Data, V1, https://doi.org/10.17632/yzg86vb895.1

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Nens van Alfen reports a relationship with Sonoskills that includes: speaking and lecture fees. Nens van Alfen reports a relationship with Wiley Publishing that includes: non-financial support. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clinph.2025.01.016.

References

Amadasun, M., King, R., 1989. Textural features corresponding to textural properties. IEEE Trans Syst, Man, Cybern 19, 1264–1274. https://doi.org/10.1109/21.44046.

Arts, I.M.P., Overeem, S., Pillen, S., Kleine, B.U., Boekestein, W.A., Zwarts, M.J., et al., 2012. Muscle ultrasonography: A diagnostic tool for amyotrophic lateral sclerosis. Clin. Neurophysiol. 123, 1662–1667. https://doi.org/10.1016/j.clinph.2011.11.262.

Caresio, C., Salvi, M., Molinari, F., Meiburger, K.M., Minetto, M.A., 2017. Fully Automated Muscle Ultrasound Analysis (MUSA): Robust and Accurate Muscle Thickness Measurement. Ultrasound Med. Biol. 43, 195–205. https://doi.org/10.1016/j.ultrasmedbio.2016.08.032.

Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM; 2016, p. 785–94. https://doi.org/10.1145/2939672.2939785.

Galloway, M.M., 1975. Texture analysis using gray level run lengths. Comput. Graphics Image Process. 4, 172–179. https://doi.org/10.1016/S0146-664X(75)80008-6.

Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Trans Syst, Man, Cybern 1973;SMC-3:610–21. https://doi.org/10.1109/TSMC.1973.4309314.

Heckmatt, J.Z., Leeman, S., Dubowitz, V., 1982. Ultrasound imaging in the diagnosis of muscle disease. J. Pediatr. 101, 656–660. https://doi.org/10.1016/S0022-3476(82)80286-2.

Lagarde, M.L.J., Van Den Engel-Hoek, L., Geurts, A.C.H., Van Alfen, N., 2023. Validity and reliability of visual assessment of orofacial muscle ultrasound images using a modified Heckmatt scale. Muscle Nerve 68, 176–183. https://doi.org/10.1002/mus.27854.

Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions 2017. https://doi.org/10.48550/ARXIV.1705.07874.

Mah, J.K., Van Alfen, N., 2018. Neuromuscular Ultrasound: Clinical Applications and Diagnostic Values. Can J Neurol Sci 45, 605–619. https://doi.org/10.1017/cjn.2018.314.

Martínez-Payá, J.J., Ríos-Díaz, J., Medina-Mirapeix, F., Vázquez-Costa, J.F., Del Baño-Aledo, M.E., 2018. Monitoring Progression of Amyotrophic Lateral Sclerosis Using Ultrasound Morpho-Textural Muscle Biomarkers: A Pilot Study. Ultrasound Med. Biol. 44, 102–109. https://doi.org/10.1016/j.ultrasmedbio.2017.09.013.

Marzola, F., Van Alfen, N., Doorduin, J., Meiburger, K.M., 2021. Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment. Comput. Biol. Med. 135, 104623. https://doi.org/10.1016/j.compbiomed.2021.104623.

Mirón-Mombiela, R., Ruiz-España, S., Moratal, D., Borrás, C., 2023. Assessment and risk prediction of frailty using texture-based muscle ultrasound image analysis and machine learning techniques. Mech. Ageing Dev. 215, 111860. https://doi.org/10.1016/j.mad.2023.111860.

Molinari, F., Caresio, C., Acharya, U.R., Mookiah, M.R.K., Minetto, M.A., 2015. Advances in Quantitative Muscle Ultrasonography Using Texture Analysis of Ultrasound Images. Ultrasound Med. Biol. 41, 2520–2532. https://doi.org/10.1016/j.ultrasmedbio.2015.04.021.

Nijboer-Oosterveld, J., Van Alfen, N., Pillen, S., 2011. New normal values for quantitative muscle ultrasound: Obesity increases muscle echo intensity. Muscle Nerve 43, 142–143. https://doi.org/10.1002/mus.21866.

Paris, M.T., Mourtzakis, M., 2021. Muscle Composition Analysis of Ultrasound Images: A Narrative Review of Texture Analysis. Ultrasound Med. Biol. 47, 880–895. https://doi.org/10.1016/j.ultrasmedbio.2020.12.012.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pillen, S., Van Alfen, N., 2015. Muscle ultrasound from diagnostic tool to outcome measure—Quantification is the challenge. Muscle Nerve 52, 319–320. https://doi.org/10.1002/mus.24613.

Sahinis, C., Kellis, E., 2023. Hamstring Muscle Quality Properties Using Texture Analysis of Ultrasound Images. Ultrasound Med. Biol. 49, 431–440. https://doi.org/10.1016/j.ultrasmedbio.2022.09.011.

Salvi, M., Caresio, C., Meiburger, K.M., De Santi, B., Molinari, F., Minetto, M.A., 2019. Transverse Muscle Ultrasound Analysis (TRAMA): Robust and Accurate Segmentation of Muscle Cross-Sectional Area. Ultrasound Med. Biol. 45, 672–683. https://doi.org/10.1016/j.ultrasmedbio.2018.11.012.

Salvi, M., Michielli, N., Meiburger, K.M., Cattelino, C., Cotrufo, B., Giacosa, M., et al., 2024. cyto-Knet : An instance segmentation approach for multiple myeloma plasma cells using conditional kernels. Int J Imaging Syst Tech 34, e22984. https://doi.org/10.1002/ima.22984.

Seoni, S., Matrone, G., Meiburger, K.M., 2023. Texture analysis of ultrasound images obtained with different beamforming techniques and dynamic ranges – A robustness study. Ultrasonics 131, 106940. https://doi.org/10.1016/j.ultras.2023.106940.

Sun, C., Wee, W.G., 1982. Neighboring gray level dependence matrix for texture classification. Comput. Graphics Image Process. 20, 297. https://doi.org/10.1016/0146-664X(82)90093-4.

Thibault, G., Fertil, B., Navarro, C., Pereira, S., Cau, P., Lévy, N., et al., 2009. Texture indexes and gray level size zone matrix. Application to cell nuclei classification.

Van Alfen, N., Mah, J.K., 2018. Neuromuscular Ultrasound: A New Tool in Your Toolbox. Can J Neurol Sci 45, 504–515. https://doi.org/10.1017/cjn.2018.269.

Van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., et al., 2017. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 77, e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339.

Vincenten, S.C.C., Teeselink, S., Voermans, N.C., Van Engelen, B.G.M., Mul, K., Van Alfen, N., 2023. Establishing the role of muscle ultrasound as an imaging biomarker in facioscapulohumeral muscular dystrophy. Neuromuscul. Disord. 33, 936–944. https://doi.org/10.1016/j.nmd.2023.10.015.

Vincenten SCC, Voermans NC, Cameron D, Van Engelen BGM, Van Alfen N, Mul K. The complementary use of muscle ultrasound and MRI in FSHD: Early versus later disease stage follow-up. Clinical Neurophysiology 2024:S1388245724000646. https://doi.org/10.1016/j.clinph.2024.02.036.

Watanabe, T., Murakami, H., Fukuoka, D., Terabayashi, N., Shin, S., Yabumoto, T., et al., 2017. Quantitative Sonographic Assessment of the Quadriceps Femoris Muscle in Healthy Japanese Adults: Quantitative Sonographic Assessment of the Quadriceps Femoris Muscle. J Ultrasound Med 36, 1383–1395. https://doi.org/10.7863/ultra.16.07054.

Wijntjes, J., Saris, C., Doorduin, J., Van Alfen, N., Van Engelen, B., Mul, K., 2024. Improving Heckmatt muscle ultrasound grading scale through Rasch analysis. Neuromuscul. Disord. 42, 14–21. https://doi.org/10.1016/j.nmd.2024.07.001.

Wijntjes, J., Van Der Hoeven, J., Saris, C.G.J., Doorduin, J., Van Alfen, N., 2022. Visual versus quantitative analysis of muscle ultrasound in neuromuscular disease. Muscle Nerve 66, 253–261. https://doi.org/10.1002/mus.27669.

Zhang W, Pang J, Chen K, Loy CC. K-Net: Towards Unified Image Segmentation 2021. https://doi.org/10.48550/ARXIV.2106.14855.