Texture analysis and machine learning to predict water T2 and fat fraction from non-quantitative MRI of thigh muscles in Facioscapulohumeral muscular dystrophy

Paolo Florent Felisaz a,b,1, Giulia Colelli a,c,1, Elena Ballante c,d, Francesca Solazzo a, Matteo Paoletti a, Giancarlo Germani a, Francesco Santini e,f, Xeni Deligianni e,f, Niels Bergsland g,h, Mauro Monforte i, Giorgio Tasca , Enzo Ricci , Stefano Bastianello a,j, Silvia Figini k,2, Anna Pichiecchio a,j,2

a Department of Neuroradiology, IRCCS Mondino Foundation, Pavia, Italy
b Department of Radiology, Desio Hospital, ASST Monza, Desio, Italy
c Department of Mathematics, University of Pavia, Pavia, Italy
d BioData Science Center, IRCCS Mondino Foundation, Pavia, Italy
e Department of Radiology, Division of Radiological Physics, University Hospital Basel, Basel, Switzerland
f Department of Biomedical Engineering, University of Basel, Basel, Switzerland
g Buffalo Neuroimaging Analysis Center, Department of Neurology, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, State University of New York, Buffalo, NY, USA
h IRCCS, Fondazione Don Carlo Gnocchi, Milan, Italy
i Unità Operativa Complessa di Neurologia, Fondazione Polyclinico Universitario A. Gemelli IRCCS, Rome, Italy
j Department of Brain and Behavioral Sciences, University of Pavia, Pavia, PV, Italy
k Department of Political and Social Sciences, University of Pavia, Pavia, PV, Italy

ARTICLE INFO

Keywords:
Texture
Machine learning
Muscle
MRI
Quantitative
FSHD

ABSTRACT

Purpose: Quantitative MRI (qMRI) plays a crucial role for assessing disease progression and treatment response in neuromuscular disorders, but the required MRI sequences are not routinely available in every center. The aim of this study was to predict qMRI values of water T2 (wT2) and fat fraction (FF) from conventional MRI, using texture analysis and machine learning.

Method: Fourteen patients affected by Facioscapulohumeral muscular dystrophy were imaged at both thighs using conventional and quantitative MR sequences. Muscle FF and wT2 were calculated for each muscle of the thighs. Forty-seven texture features were extracted for each muscle on the images obtained with conventional MRI. Multiple machine learning regressors were trained to predict qMRI values from the texture analysis dataset.

Results: Eight machine learning methods (linear, ridge and lasso regression, tree, random forest (RF), generalized additive model (GAM), k-nearest-neighbor (kNN) and support vector machine (SVM) provided mean absolute errors ranging from 0.110 to 0.133 for FF and 0.068 to 0.115 for wT2. The most accurate methods were RF, SVM and kNN to predict FF, and tree, RF and kNN to predict wT2.

Conclusion: This study demonstrates that it is possible to estimate with good accuracy qMRI parameters starting from texture analysis of conventional MRI.

1. Introduction

Quantitative Magnetic Resonance Imaging (qMRI) is a useful non-invasive diagnostic tool in the field of neuromuscular diseases, providing clinically relevant parameters, such as Fat Fraction (FF) and muscle water T2 (wT2) [1,2]. These parameters provide sensitive...
measures of muscle damage and are aimed toward having a prognostic role. In fact, they could be used to track disease progression or response to treatments [3]. However, implementing qMRI protocols is challenging in terms of technical requirements and financial resources, thus they are not accessible in every neuromuscular center. Instead, conventional (non-quantitative) MRI is much more widespread. Conventional sequences in diagnostic protocols for muscle MRI usually include T1-weighted images and water sensitive sequences, such as STIR or T2-weighted with or without fat suppression [4]. These conventional sequences are devoted to the macroscopic evaluation of fat replacement and muscle edema. Semi-quantitative rating scales are used clinically based on visual inspection, the most commonly used being the Fischer and Mercuri scales [5,6].

Fascioscapulohumeral muscular dystrophy (FSHD) is one of the most common muscular dystrophies and is characterized by a progressive asymmetric loss of strength and atrophy of skeletal muscles [7,8]. Weakness usually affects first the facial and shoulder girdle muscles and then abdominal, upper and lower limbs. Conventional MRI is used for distinguishing FSHD from other myopathies and measuring disease severity with the common grading scales [9], as well as for semi-quantitatively assessing disease activity [10]. QMRI instead can potentially aid in the follow-up of patients with FSHD [11].

Texture analysis can extract quantitative features from qualitative images. These textural features are descriptors of the pixel intensity variation and distribution within an image and are related to human discriminable visual patterns, such as contrast and granularity [12]. They are commonly processed with machine learning algorithms to predict clinically useful outcomes after training from a dataset. Basically, they can fall within two broad categories: classifiers, aimed to predict discrete outcomes, for instance benign versus malignant or a particular grade of pathology, and regressors, aimed to predict continuous variables such as survival time or a quantitative biomarker.

The aim of this paper was to use machine-learning regressors to predict wT2 and FF starting from texture analysis of non-quantitative MRI sequences. We used a dataset of MRI scans from patients affected by FSHD with different degrees of intramuscular fatty replacement and edema. A cross validation framework was set to test and compare multiple parametric and nonparametric machine learning models, including linear regression, ridge regression, lasso regression, regression tree, random forest (RF), generalized additive model (GAM), support vector machine (SVM) and k nearest neighbor (kNN).

2. Methods

2.1. Study design and participants

The study was approved by the local ethics committee and all the participants provided written informed consent. Fourteen patients affected by FSHD (11 men, 3 women, mean age 45.6 years, range 32–60) were recruited in a longitudinal study and imaged at the thigh level every six months. Seven patients were imaged 3 times, four patients were imaged twice and three were imaged once, for a total of 32 examinations. Patients had different grades of muscular involvement as classified by the clinical severity scale (CSS median 3.5, 95% CI for the median 3.25–3.75). The protocol included one averaged value of FF and wT2. The images obtained by the 8th echo of the MESE sequence (T2 weighted images with TE = 88 ms) were extracted and considered as the conventional (non-quantitative) images of the study. We used the software LifeX [16] to compute the texture analysis on the T2 weighted images, using the same ROIs previously drawn on the MESE images. Voxel values were normalized and then quantized to 64 Gy levels. All possible features provided by the software were extracted, including first order statistics including features derived from the Gray Level Zone Length Matrix (GLZLM), Gray Level Run Length Matrix (GLRLM), Neighborhood-Gray-Level Different Matrix (NGLDM) and Gray Level Co-occurrence Matrix (GLCM), for a total of 47 features. Similarly to the process for FF and wT2, ROIs from all 5 slices and both sides of the same muscle were merged in order to obtain 12 observations, each made of 47 texture features, per examination. Then texture features, wT2 and FF of each observation were integrated in a unique database. FF values were already normalized whereas wT2 values were scaled ranging from 0 to 1 in order to apply the algorithms described in the following section.

2.2. MR protocols

All examinations were performed on a 3 T scanner (Skyra, Siemens AG Erlangen, Germany) using integrated spine and body surface coils. Acquisition volumes were centered on the thigh muscles at a standardized position (the last slice of the box was located at 10 cm proximally from the upper base of the patella). The protocol included a 3D 6-point multi-echo gradient echo sequence with shifted echo times (MEGE, matrix size = 432 × 396; 52 slices, TR = 35 ms; TE = 1.7–9.2 ms; resolution = 1.0 × 1.0 × 5mm³, total scan time = 6 min) and a multi-echo spin echo sequence (MESE, TE = 10.9 ms, TR = 4100.0 ms; 17 echo times; resolution = 1.2 × 1.2 × 10.0 mm; gap 30 mm, 5 slices). The Fatty Riot algorithm was used offline for the calculation of fat/water fraction maps from the MEGE sequence [13]. For the MESE sequence, extended phase graph signal simulation including slice profile was implemented offline in Python (Python Software Foundation). Python Language Reference, version 3.8) [14,15].

2.3. Image processing and texture analysis

One operator drew regions of interest (ROIs) for each muscle of the thigh using the images of the first echo obtained from the MESE sequence. 12 muscle ROIs per side were drawn on each slice, for a total of 120 ROIs per examination (See Fig. 1). The ROIs were co-registered to the MEGE images with creation of new corresponding MEGE ROIs, and then manually adjusted by the same operator. FF and wT2 mean values were extracted from each ROI from the MESE and MEGE sequences, using the above-described methods. For instance the observation “Sartorius” included respectively one averaged value of FF and wT2. The images obtained by the 8th echo of the MESE sequence (T2 weighted images with TE = 88 ms) were extracted and considered as the conventional (non-quantitative) images of the study. We used the software LifeX [16] to compute the texture analysis on the T2 weighted images, using the same ROIs previously drawn on the MESE images. Voxel values were normalized and then quantized to 64 Gy levels. All possible features provided by the software were extracted, including first order features statistics based on histogram and shape, and second orders statistics including features derived from the Gray Level Zone Length Matrix (GLZLM), Gray Level Run Length Matrix (GLRLM), Neighborhood-Gray-Level Different Matrix (NGLDM) and Gray Level Co-occurrence Matrix (GLCM), for a total of 47 features. Similarly to the process for FF and wT2, ROIs from all 5 slices and both sides of the same muscle were merged in order to obtain 12 observations, each made of 47 texture features, per examination. Then texture features, wT2 and FF of each observation were integrated in a unique database. FF values were already normalized whereas wT2 values were scaled ranging from 0 to 1 in order to apply the algorithms described in the following section.

2.4. Machine learning and statistical assessment

We tested a set of parametric models including linear [17], ridge [18], lasso regression [19] and GAMs [20] and nonparametric models including kNN [21], SVM [22], tree [23] and RF [24] (Python). All observations were considered independent to maximize the number of samples for training the models. To assess the potential effect of correlation between observations from the same patients, we also implemented a mixed effect model where the different subjects were considered as random effects.

Performances of the models were estimated with the indicators mean square error (MSE) and mean absolute error (MAE). The MSE gives information on the mean quadratic discrepancy between the target values \(y_i\) and the predicted values \(\hat{y}_i\):

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]  

(1)

The MAE gives information on the mean of the absolute values of discrepancy between the target values \(y_i\) and the predicted values \(\hat{y}_i\):

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

(2)

To achieve a more realistic assessment of the performance we used cross-validation, a resampling approach in which each model is fitted multiple times on different subsets of the training data. In more detail, for a preliminary estimation of the model hyperparameters, a grid search
was used and the entire data set was divided in training and test sets. The training set included the randomly selected 80% of the data set and the test set the remaining 20%. Then, for model assessment, the k-folds cross-validation with \( k = 5 \) was used. This approach involves randomly dividing the entire data set into 5 folds of approximately equal size, using one fold as test set and the remaining four as training sets. The performance indicators are calculated on the test set five times, each time changing the test set fold. We reported means and standard deviations of the performance indicators for each model.

### 3. Results

The final dataset consisted of 384 observations, each with 2 target variables and 47 texture features (covariates) related to 32 MR examinations from 14 patients. The target variables FF values ranged from 0.027 to 0.899 (2.7 to 89.9%) and \( wT2 \) from 26.40 to 78.14 ms.

Three of the original 47 covariates (HISTO_Entropy_log10, `SHAPE_Volume_vx`, GLCM_Entropy_log10) that had strong correlation with the others, established on the basis of p-value, were not included in the analysis to avoid collinearity issues.

Table 1 reports means of MSE and MAE related to FF prediction for each model implemented.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression</td>
<td>0.029 (0.008)</td>
<td>0.122 (0.015)</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>0.028 (0.008)</td>
<td>0.127 (0.018)</td>
</tr>
<tr>
<td>Lasso regression</td>
<td>0.030 (0.010)</td>
<td>0.133 (0.021)</td>
</tr>
<tr>
<td>GAM</td>
<td>0.028 (0.007)</td>
<td>0.122 (0.013)</td>
</tr>
<tr>
<td>Regression Tree</td>
<td>0.033 (0.008)</td>
<td>0.125 (0.013)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.023 (0.008)</td>
<td>0.105 (0.018)</td>
</tr>
<tr>
<td>kNN</td>
<td>0.024 (0.008)</td>
<td>0.110 (0.020)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.026 (0.008)</td>
<td>0.114 (0.016)</td>
</tr>
<tr>
<td>Mixed effect model</td>
<td>0.025 (0.005)</td>
<td>0.118 (0.009)</td>
</tr>
</tbody>
</table>
As a result, the prediction performances were good for all models in terms of MSE and MAE and the values were stable (low values of standard deviation). RF was the model with minimum predictive errors (mean value MSE = 0.023, MAE = 0.105), closely followed by KNN (MSE = 0.024) and SVM (MSE = 0.026). We can thus conclude that in the best case (RF) the algorithm predicted the expected FF value with a mean error of approximately +/-11 percentage points (pp), and in the worst case (Trees) of +/-13pp.

Table 2 reports means of MSE and MAE related to water T2 for each model implemented. The prediction performances were good for all models in terms of average MSE and MAE and in terms of stability, but in this case the kNN gave an outperformance result (MSE = 0.010, MAE 0.068,) followed by RF (MSE = 0.022) and Ridge and Trees (both with MSE = 0.023). Thus the algorithms could predict wT2 with an error ranging from +/- 7pp to +/- 12pp, equivalent to a minimum of 3.50 ms (kNN) to a maximum of 5.94 ms (Linear regression).

The performance of the mixed effect model was similar to the other models (for FF MSE = 0.025 and for wT2 MSE = 0.022)

4. Discussion

In this study, a set of machine learning models are proposed to predict wT2 and FF of thigh muscles using texture analysis of conventional MRI, starting from a dataset of MR examinations from patients affected with FSHD. To the authors’ knowledge, this is the first attempt to derive qMRI parameters from texture analysis and our promising results should be considered a proof of principle for further improvements in the future.

There is a growing need to have muscle qMRI parameters accessible and easily obtainable in most neuromuscular centers for assessing disease progression or response to new therapies in rare muscle disorders [25]. Studies showed that qMRI parameters such as wT2 and FF are strongly correlated to the clinical outcome of patients affected with FSHD [3]. There is evidence that wT2 and FF change over time as disease progresses [26,27] and that FF can be responsive to the effects of treatments [11]. However, having accurate qMRI results is technically challenging and even though commercial and open source qMRI packages are available, further modifications or tuning are often needed to avoid drawbacks. For instance, post-processing correction of stimulated echo artifacts is mandatory for robust measures of wT2 maps obtained with MESE sequences [14]. Likewise, 6-point MEGE sequences FF maps are superior to the commonly used 2- or 3-point Dixon techniques used for FF maps. However, 6-point MEGE sequences are not consistently available across different scanning platforms as of now.

Texture analysis and machine learning algorithms can predict clinically relevant outcomes starting from non-quantitative imaging. Most studies used texture analysis to classify discrete outcomes, for instance atrophic versus normal muscles ex vivo [28] and in vivo [29], or to distinguish different types of myopathies [30]. In our study we used machine learning models, more specifically regression models, to predict continuous outcomes such as muscle qMRI parameters FF and wT2.

Our results are encouraging. We observed a minimum MAE of 11pp in FF, which is enough to automatically score muscles with a clinical 5-point scale [5,6] allowing also a more precise grading of intermediate levels of FF. In fact, 5-point scales have extreme values consisting in normal (0) or completely fatty substituted (4) muscles, and three intermediate scores consisting in mild, moderate and severe fat substitution. For wT2 the results were even better, with errors ranging from 3.50 ms to 5.94 ms. Since we have no previous examples of such similar analysis in the literature, we can speculate that the better performance for wT2 might reflect a more homogeneous increase of signal overall the muscle, possibly related to textural features sensitive to signal intensity. FF, instead, corresponds to a pattern more related to the morphology, with well-defined strands of fat substituting the muscles from the epimysium. The application of deep learning methods would likely further increase the performance of FF prediction and will be the target of future studies.

We kept almost all features in the analysis, removing only three that were highly correlated with one another. To reduce a potential cause of over-fitting we opted for implementing regularization instead of an initial features selection. In fact, L1 regularization included in the lasso regression, favoring a sparse solution, implicitly implements feature selections. Since there were no relevant differences in performances between regularized (ridge and lasso) and unregularized (linear) models, we performed the other models (in particular non parametric) while keeping all 44 features as described in the results section. Nonparametric models worked better in both analyses, suggesting a complex and non linear relationship between predictors and target variables.

This study has some limitations. The first is related to the relative homogeneity of the training dataset. To expand the applicability of our results, algorithms should be trained with a mixed and larger database encompassing more patients with other muscular diseases, and including normal subjects. Also, they should include examinations from lower field MRI scanners and different vendors. Another limitation is the presence of multiple examinations from the same patients at different time points and different muscles associated to the same patient. However, the mixed effect model did not demonstrate clear differences in performance with respect to the analyses where observations were considered uncorrelated. We applied texture analysis to T2 weighted images without fat-suppression, extracted from the MESE sequence, but T1 weighted and STIR are usually preferred in diagnostic protocols. In this study, as a proof of concept, we chose the sequence where both signals from wT2 and fat were present in order to maximize the chances to prove the feasibility of our method. However, future studies should investigate other sequences with different parameters (TR, TE, slice thickness etc.) including STIR and T1w sequences. We deliberately considered “observations” from the average of both left and right. This was a technical limitation related to the necessity of having a minimum amount of voxels for the texture analysis. In facts, some ROIs were too small and not all muscles were represented in all slices. Higher resolution images would have led to twice the number of observations, at least one per muscle per side.

Future studies may be conducted using deep learning methods that would remove the necessity of texture analysis. These methods, commonly used for muscle segmentation, provide faster analysis of qMRI sequences [31] and have been used for classifying different subtypes of muscular dystrophies [32]. However, such techniques typically need much larger datasets to be properly trained. One solution could be the use of data augmentation techniques, which artificially increase the training dataset.

In conclusion, we demonstrated the feasibility of predicting the qMRI parameters FF and wT2 using texture analysis and machine learning methods, starting from conventional T2 weighted images. Our encouraging results may extend the implementation of qMRI to all centers dealing with neuromuscular diseases that use standard sequences. Training data is critical and future studies with larger and mixed cohorts are warranted to improve the performance of these methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE (0.008)</th>
<th>MAE (0.016)</th>
<th>MAE in T2 scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression</td>
<td>0.025</td>
<td>0.115</td>
<td>5.943 (0.831)</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>0.023</td>
<td>0.105</td>
<td>5.450 (0.882)</td>
</tr>
<tr>
<td>Lasso regression</td>
<td>0.024</td>
<td>0.108</td>
<td>5.580 (0.755)</td>
</tr>
<tr>
<td>GAM</td>
<td>0.025</td>
<td>0.113</td>
<td>5.865 (0.909)</td>
</tr>
<tr>
<td>Regression Tree</td>
<td>0.023</td>
<td>0.098</td>
<td>5.048 (0.0762)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.022</td>
<td>0.098</td>
<td>5.089 (0.692)</td>
</tr>
<tr>
<td>kNN</td>
<td>0.010</td>
<td>0.068</td>
<td>3.501 (0.438)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.026</td>
<td>0.099</td>
<td>5.134 (0.810)</td>
</tr>
<tr>
<td>Mixed effect model</td>
<td>0.022</td>
<td>0.109</td>
<td>5.664 (0.810)</td>
</tr>
</tbody>
</table>
Declaration of Competing Interest

The authors report no declarations of interest.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi: https://doi.org/10.1016/j.ejrad.2020.109460.

References


Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi: https://doi.org/10.1016/j.ejrad.2020.109460.

References