SEVERITY CLASSIFICATION IN CASES OF COLLAGEN VI-RELATED MYOPATHY WITH CONVOLUTIONAL NEURAL NETWORKS AND HANDCRAFTED TEXTURE FEATURES

Rafael Rodrigues[1]   *Susana Quijano-Roy*[2]   *Robert-Yves Carlier*[2]   *Antonio M. G. Pinheiro*[1]

[1] Instituto de Telecomunicações and Universidade da Beira Interior, Covilhã, Portugal
[2] APHP - Raymond Poincaré University Hospital, Garches, France

ABSTRACT

Magnetic Resonance Imaging (MRI) is a non-invasive tool for the clinical assessment of low-prevalence neuromuscular disorders. Automated diagnosis methods might reduce the need for biopsies and provide valuable information on disease follow-up. In this paper, three methods are proposed to classify target muscles in Collagen VI-related myopathy cases, based on their degree of involvement, notably a Convolutional Neural Network, a Fully Connected Network to classify texture features, and a hybrid method combining the two feature sets. The proposed methods were evaluated on axial T1-weighted Turbo Spin-Echo MRI from 26 subjects, including Ullrich Congenital Muscular Dystrophy and Bethlem Myopathy patients at different evolution stages. The hybrid model achieved the best cross-validation results, with a global accuracy of 93.8%, and F-scores of 0.99, 0.82, and 0.95, for healthy, mild and moderate/severe cases, respectively.

Index Terms— Collagen VI-related myopathy, MRI, Computer-aided diagnosis, Texture analysis, Convolutional Neural Networks

1. INTRODUCTION

Neuromuscular diseases (NMD) comprise a wide range of individually rare disorders, with several different causes and phenotypes [1, 2]. Ullrich Congenital Muscular Dystrophy (UCMD) is an inherited early-onset disorder, caused by mutations in the Collagen VI (COL6) genes, and is recognized as the most severe form of COL6-related myopathy. Bethlem Myopathy corresponds to a milder form, which typically has a later onset [3]. The observed symptoms include generalized muscle weakness and hypotonia, joint contractures, distal joint hyperlaxity and scoliosis. Motor development is often delayed, and walking ability might be lost or never acquired, in the most severe forms. Also, in more severe cases, patients may end up developing spinal deformities and life-threatening respiratory insufficiency [4, 5].
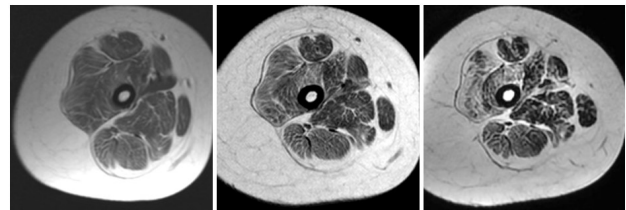
Fig. 1: Progression of COL6 myopathy in T1w MRI exams of an ambulatory young female affected with UCMD, at ages 5, 12 and 19 (from left to right).

The muscle involvement pattern is similar across the COL6 myopathy spectrum, albeit with varying degrees of severity [1, 3]. The appearance of a striped pattern, with alternating bands of hypointensity (i.e., preserved muscle) and hyperintensity (i.e., fat/connective tissue), is common in T1-weighted (T1w) Magnetic Resonance Imaging (MRI) [4]. In the thigh, muscle involvement is typically diffuse, with some common patterns in T1w MRI, such as a hyperintensity rim appearing in the *Vastus lateralis* and central hyperintensity ("target") in the *Rectus femoris* [3].

MRI provides a reliable and non-invasive clinical outcome measure for the diagnosis and monitoring of NMD [2, 3]. In specialized centers, it may be possible to assess the course of the myopathy using muscle MRI (Fig. 1), helped by standardized scoring forms of signal and volume changes [2]. Texture analysis of MRI scans has shown great potential in the development of computer-aided diagnosis (CAD) methods for NMD that might ultimately help reduce the need for biopsies, which are currently widely used in diagnosis and clinical follow-up of patients. However, developing CAD methods for these diseases is still a very challenging field, particularly due to their individual rareness [2, 6]. In this paper, we propose the use of Convolutional Neural Networks (CNN) to classify different stages of muscle involvement, on T1w Turbo Spin-Echo (TSE) MRI scans of UCMD and Bethlem Myopathy patients. Moreover, we also test this classification with a set of handcrafted texture features, classified using a Fully Connected Network (FCN), and a hybrid model combining both approaches.

2. RELATED WORK

Texture features such as intensity histogram statistics, the Gray-Level Co-occurrence Matrix (GLCM), the Run Length Matrix (RLM), Local Binary Patterns, and Wavelet-based features have been used in a few studies [7, 8, 9, 10, 11] towards the development of CAD methods for muscle diseases using MRI, particularly the Duchenne Muscle Dystrophy (DMD) and the Golden Retriever Muscle Dystrophy (GRMD), which is considered to be highly similar to DMD [6].

More recently, Cai *et al.* [12] used different CNN architectures (ResNet and VGG) to automatically assess the progression of DMD and Congenital Muscular Dystrophy on Chemical Shift-Based Water-Fat Separation MRI. The authors also introduced an improved class activation mapping, which increased the mean accuracy of the best-performing CNN (ResNet-18) up to 91.7%. Yang *et al.* [13] adopted a ResNet-50 model to diagnose dystrophinopathies on thigh muscle MRI, with an accuracy of 91%. The authors also tested other CNN architectures, such as VGG-19, Inception-V3 and two DenseNet models. The lowest reported accuracy was 87% (VGG-19).

Very little research has been published so far, regarding CAD methods for COL6-related myopathy. Bazaga *et al.* [5] proposed the first CAD method, which relies on a CNN to classify image patches, extracted from confocal microscopy images of fibroblast cultures. A majority voting algorithm was applied to obtain a global decision on a given input image. The proposed method achieved an accuracy of 95%. Recently, the authors of this paper assessed the effectiveness of several texture features in describing the level of muscle involvement on T1*w* TSE MRI [14]. Feature selection was performed using SVM Recursive Feature Elimination (SVM-RFE) [15]. The proposed method yielded accuracy values above 90% for 3 of the 4 studied muscles.

3. METHODS

3.1. MRI data and patch extraction

A group of 26 subjects, with ages ranging from 5 to 36 years old were examined using a 1.5T MRI scanner (Philips Medical Systems, Eindhoven, The Netherlands), at the Medical Imaging Department of the Raymond-Poincaré University Hospital (Garches, France). UCMD or Bethlem Myopathy was diagnosed on 17 subjects, while the remaining 9 showed no perceivable signs of myopathy in the MR scans.

In this research, we selected slices from whole body muscle MRI scans performed in the same center, with the same magnet system and technical protocol [4]. We used axial T1*w* TSE MR images of both thighs, with TR/TE = 631/16ms and slice thickness = 6mm. A different amount of slices were selected from each subject (6 to 12), to maximize the cross-sectional area of target muscles.
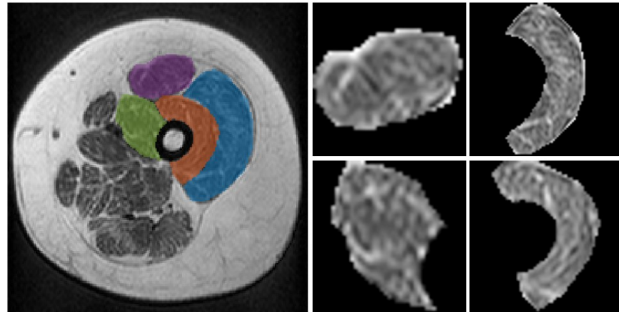

Fig. 2: Left: Segmentation of the *Quadriceps femoris* muscles in an MR image of the thigh. Right: Image patches containing the segmented muscles.

The *Quadriceps femoris* muscles, i.e., *Vastus lateralis*, *Vastus medialis*, *Vastus intermedius* and *Rectus femoris*, were manually segmented in all selected images (Fig. 2, left). The data was annotated at the muscle level by experts, according to a 4-level scale proposed in [3, 16] (healthy / no symptoms, mild, moderate, and severe). In this work, the scale was modified into a 3-level scale, by joining moderate and severe cases.

The selected T1*w* TSE MRI dataset consisted of 196 images, which included both thighs. Thus, each of the four target muscles appeared in the dataset 392 times, yielding a total of 1568 regions of interest (ROI). Image patches containing the segmented muscles were obtained by cropping the MR image within the square segmentation bounding boxes (Fig. 2, right). All pixels outside the segmented muscles were set to 0. The size of the extracted patches ranged from 11x11 to 105x105. However, the vast majority were above 30x30. All patches were resized to 96x96, which was chosen as input size of the *ConvNet* model described in Section 3.2. This input size was chosen to promote a balance between avoiding texture distortions from resizing, as much as possible, and allowing a reasonable network depth for feature extraction.

3.2. CNN model architecture

For the classification of the disease severity in each segmented muscle, we implemented a *ConvNet* encoder (Fig. 3), consisting of sequential 2D convolutional layers, with an increasing number of filters (64, 128, 256, and 512), and progressively smaller spatial kernels (7x7, 5x5, 3x3, and 3x3). The output of the first three convolutional layers is downsampled using 2x2 max pooling. To further reduce the dimensionality of the encoded features, a 1x1 convolutional layer was added at the end. All convolutional layers used 'same' *padding*, so that the height and width of their output matched those of the input. The classification part consisted of a FCN, with two layers of 512 nodes, both with 20% *dropout* to prevent *overfitting*, and a 3-node output layer. The ReLU activation function was used in every layer, with the exception of the output layer, which used a *softmax* activation.
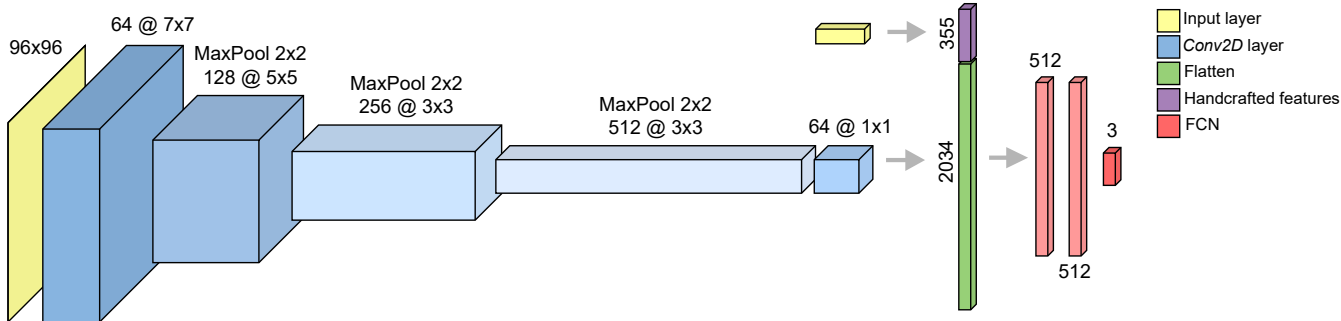
Fig. 3: Architecture of the hybrid classification model. In the *ConvNet*-only model, the architecture is similar, except for the concatenation of the handcrafted texture features (top-right).

Table 1: Performance results of the tested models. Precision, Recall, and the F-score are reported with respect to each class, whereas Accuracy refers to the overall classification accuracy.

Model	Healthy			Mild			Moderate/Severe			Accuracy
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	
ConvNet	0.96	0.95	0.96	0.73	0.80	0.76	0.95	0.93	0.94	91.3%
Texture features	0.98	0.98	0.98	0.73	0.75	0.74	0.93	0.92	0.92	90.8%
Hybrid model	0.98	0.99	0.99	0.82	0.82	0.82	0.95	0.94	0.95	93.8%

3.3. Handcrafted texture features

As a follow-up to the study presented in [14], we tested the classification of 355 ROI-based handcrafted texture features, obtained from the original images (i.e., without resizing), using an FCN classifier. These include statistics from the intensity and gradient histograms [17], GLCM and RLM-based features [11], features from the autoregressive (AR) model [17], Wavelet-based energy [10], and statistics from the Gabor response magnitude images [18].

These texture features were computed using the *MaZda* software [17], with the exception of Gabor-based features, which were extracted using MATLAB R2020b. The FCN classifier included two densely connected layers of 256 nodes, using a ReLU activation, and a similar 3-node output layer with a *softmax* activation.

3.4. Hybrid model

Finally, a hybrid classifier was also tested, which combines the handcrafted texture features and the flattened output of the *ConvNet* into a single tensor. In this case, the FCN architecture was the same as in the *ConvNet* experiment. The weights of the convolutional layers, which had been trained beforehand in each fold, were stored, and loaded for the corresponding test subset. The *ConvNet* features in the hybrid model were computed without further training, whereas the parameters of the FCN classifier were fully trained in this experiment. Fig. 3 shows the full architecture diagram of the hybrid classification model.

3.5. Model training and evaluation

A leave-one-out cross-validation was set at the subject level for model training and evaluation, in order to maximize the number of patches in the training subset. Given the label imbalance within the dataset, class weights were computed in each fold, using the *compute_class_weights* function of the *scikit-learn Python* library. The weight for each label (w_l) was defined as $w_l = {}^N\!/_{(L \times n_l)}$, where N and L are the total number of images and labels, respectively, and n_l is the number of images belonging to label l. The weights were passed as an argument to the fitting algorithm model, thus assigning a higher importance to patches of the least represented classes.

To address the issue of poor network generalization that might arise from a relatively small dataset, we implemented a data augmentation scheme. At each iteration the images were transformed through the random application of horizontal flipping, horizontal translation within a range of [-10%, 10%] of the image width, rotation within a range of [-0.2, 0.2] $\times 2\pi$, and contrast adjustment with a contrast factor (c_f) within a range of [0.8, 1.2]. The pixel values x are adjusted according to $c_f(x - \overline{x}) + \overline{x}$), where $\overline{x}$ represents the mean of the pixel values. The grayscale input images were also normalized to the [0, 1] range.

The proposed models were implemented using the Keras API, with a Tensorflow backend, and optimized using the Adam algorithm for 100 epochs, with a batch size of 25, using the categorical cross-entropy loss function. The learning rate was initially set at 10^{-3}, with a step-wise decay of 0.1 at every 20 epochs. We also implemented L2 weight regulariza-
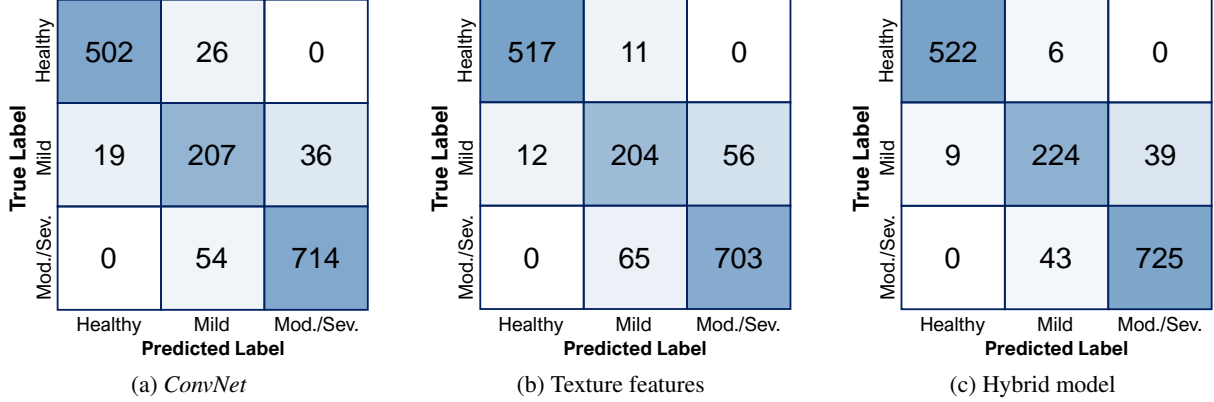
Fig. 4: Cross-validation confusion matrices.

Table 2: Overall accuracy and average F-score (across all labels) considering each target muscle independently.

Model	Vastus lateralis		Vastus intermedius		Vastus medialis		Rectus femoris	
	Accuracy	F-score	Accuracy	F-score	Accuracy	F-score	Accuracy	F-score
ConvNet	94.6%	0.92	93.4%	0.89	83.9%	0.80	93.6%	0.93
Texture features	89.3%	0.84	96.4%	0.94	85.0%	0.82	94.1%	0.93
Hybrid model	94.4%	0.91	97.0%	0.94	87.5%	0.85	96.4%	0.96

tion with $\alpha = 0.01$, to further help reducing the probability of *overfitting*.

4. RESULTS AND DISCUSSION

Table 1 shows performance measures of the three tested models (i.e., precision, recall, and F-score) considering each severity grade, as well as the global accuracy. The hybrid model achieves better results, with a higher global classification accuracy of 93.8%, mainly due to the better classification of mild cases. A comparison between the *ConvNet* and the texture features classifier shows that the first performed better with the mild and moderate/severe cases (Fig. 4a), while the second improved the correct identification of healthy cases (Fig. 4b). All models performed well in separating healthy from affected cases, even between healthy and mild, which shows good potential for the early detection of myopathy without biopsy. From the confusion matrices in Fig. 4, we may conclude that misclassifications occurred only between adjacent classes, i.e., healthy/mild or mild/moderate-severe. This suggests the potential of using the proposed approaches in follow-up studies, e.g., in monitoring treatment outcome with reduced invasiveness.

Considering only results for each muscle, which are summarized in Table 2, the *Vastus medialis* led to the worst overall results, and particularly in mild cases with 32 misclassifications (*ConvNet*), 31 (texture features), and 26 (hybrid model), out of 80 patches. Most errors were mild samples classified as moderate/severe, but a few cases of mild classified as

healthy also occurred (7 with *ConvNet* and 1 with the other two methods. These results are in line with those obtained in [14], where the *Vastus medialis* also led to the the worst results (recall = 0.77 / precision = 0.72) considering the three classes. Fewer misclassifications of mild cases occurred in other muscles, with either model (maximum of 15 with the texture features for the *Rectus femoris*). However, the smaller number of samples, when compared with the healthy and moderate/severe classes, accentuates the weight of these errors in the performance measures.

This study differs from [14], as ROI patches from all muscles were combined in model training, with the intention of working towards a more generalized and automated solution. Nonetheless, as is common with deep learning-based solutions, the lack of training data is very likely to be a relevant constraint to the improvement of these results.

5. CONCLUSIONS AND FUTURE WORK

The proposed models have shown potential to assist radiologists in performing non-invasive diagnosis and follow-up of patients with COL6-related myopathies, as well as other NMD. As these diseases are fortunately rare, it is even more important to find solutions that provide accurate classifications with reduced amounts of data.

In future research efforts, it will be essential to test this approach with an increased number of different muscles, towards enabling fully automated whole body myopathy screenings, combined with a muscle segmentation method.

6. REFERENCES

[1] L. ten Dam, A. J. van der Kooi, C. Verhamme, M. P. Wattjes, and M. de Visser, "Muscle imaging in inherited and acquired muscle diseases," *Eur. J. Neurol.*, vol. 23, no. 4, pp. 688–703, 2016.

[2] S. Quijano-Roy and R.-Y. Carlier, "Neuroimaging in non-dystrophic myopathies," in *Clinical Neuroradiology: The ESNR Textbook*, F. Barkhof, H. R. Jäger, M. M. Thurnher, and À. Rovira, Eds., pp. 2145–2183. Springer International Publishing, 2019.

[3] E. Mercuri et al., "Muscle MRI in Ullrich congenital muscular dystrophy and Bethlem myopathy," *Neuromuscular Disorders*, vol. 15, no. 4, pp. 303–310, 2005.

[4] S. Quijano-Roy, D. Avila-Smirnow, R.-Y. Carlier, and WB-MRI Muscle Study Group, "Whole body muscle MRI protocol: pattern recognition in early onset NM disorders," *Neuromuscular Disorders*, vol. 22, pp. S68–S84, 2012.

[5] A. Bazaga, M. Roldán, C. Badosa, C. Jiménez-Mallebrera, and J. M. Porta, "A Convolutional Neural Network for the automatic diagnosis of Collagen VI-related muscular dystrophies," *Applied Soft Computing*, vol. 85, pp. 105772, 2019.

[6] J. D. De Certaines et al., "Application of texture analysis to muscle MRI: 1-What kind of information should be expected from texture analysis?," *EPJ Nonlinear Biomedical Physics*, vol. 3, no. 1, pp. 1–14, 2015.

[7] S. Herlidou, Y. Rolland, J. Y. Bansard, E. Le Rumeur, and J. D. De Certaines, "Comparison of automated and visual texture analysis in MRI: characterization of normal and diseased skeletal muscle," *Magnetic Resonance Imaging*, vol. 17, no. 9, pp. 1393–1397, 1999.

[8] Z. Fan et al., "Characteristics of magnetic resonance imaging biomarkers in a natural history study of Golden Retriever Muscular Dystrophy," *Neuromuscular Disorders*, vol. 24, no. 2, pp. 178–191, 2014.

[9] D. Duda, M. Kretowski, N. Azzabou, and D. Jacques, "MRI texture analysis for differentiation between healthy and golden retriever muscular dystrophy dogs at different phases of disease evolution," in *IFIP International Conference on Computer Information Systems and Industrial Management*. Springer, 2015, pp. 255–266.

[10] M.-H. Zhang, J.-S. Ma, Y. Shen, and Y. Chen, "Optimal classification for the diagnosis of Duchenne Muscular Dystrophy images using Support Vector Machines," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 9, pp. 1755–1763, 2016.

[11] A. Eresen, L. Alic, S. M. Birch, W. Friedeck, J. F. Griffin IV, J. N. Kornegay, and J. X. Ji, "Texture as an imaging biomarker for disease severity in Golden Retriever Muscular Dystrophy," *Muscle & Nerve*, vol. 59, no. 3, pp. 380–386, 2019.

[12] J. Cai, F. Xing, A. Batra, F. Liu, G. A. Walter, K. Vandenborne, and L. Yang, "Texture analysis for muscular dystrophy classification in MRI with improved class activation mapping," *Pattern Recognition*, vol. 86, pp. 368–375, 2019.

[13] M. Yang, Y. Zheng, Z. Xie, Z. Wang, J. Xiao, J. Zhang, and Y. Yuan, "A deep learning model for diagnosing dystrophinopathies on thigh muscle MRI images," *BMC Neurology*, vol. 21, no. 1, pp. 1–9, 2021.

[14] R. Rodrigues, M. G.-G. de la Banda, M. Tordjman, D. Gómez-Andrés, S. Quijano-Roy, R.-Y. Carlier, and A. M. G. Pinheiro, "Texture analysis of T1-weighted Turbo Spin-Echo MRI for the diagnosis and follow-up of Collagen VI-related myopathy," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1274–1278.

[15] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using Support Vector Machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[16] A. E. Lamminen, "Magnetic resonance imaging of primary skeletal muscle diseases: patterns of distribution and severity of involvement," *The British Journal of Radiology*, vol. 63, no. 756, pp. 946–950, 1990.

[17] P. M. Szczypiński, M. Strzelecki, A. Materka, and A. Klepaczko, "MaZda — a software package for image texture analysis," *Computer Methods and Programs in Biomedicine*, vol. 94, no. 1, pp. 66–76, 2009.

[18] Y. Zhang, W. Li, L. Zhang, X. Ning, L. Sun, and Y. Lu, "Adaptive learning Gabor filter for finger-vein recognition," *IEEE Access*, vol. 7, pp. 159821–159830, 2019.