# Texture Analysis for Muscular Dystrophy Classification in MRI with Improved Class Activation Mapping*

Jinzheng Cai, Fuyong Xing, Abhinandan Batra, Fujun Liu,
Glenn A. Walter, Krista Vandenborne, Lin Yang,

University of Florida
caijinzhengcn@gmail.com

## Abstract

The muscular dystrophies are made up of a diverse group of rare genetic diseases characterized by progressive loss of muscle strength and muscle damage. Since there is no cure for muscular dystrophy and clinical outcome measures are limited, it is critical to assess the progression of MD objectively. Imaging muscle replacement by fibrofatty tissue has been shown to be a robust biomarker to monitor disease progression in DMD. In magnetic resonance imaging (MRI) data, specific texture patterns are found to correlate to certain MD subtypes and thus present a potential way for automatic assessment. In this paper, we first apply state-of-the-art convolutional neural networks (CNNs) to perform accurate MD image classification and then propose an effective visualization method to highlight the important image textures. With a dystrophic MRI dataset, we found that the best CNN model delivers an 91.7% classification accuracy, which significantly outperforms non-deep learning methods, *e.g.*, $> 40\%$ improvement has been found over the traditional mean fat fraction (MFF) criterion for DMD and CMD classification. After investigating every single neuron at the top layer of CNN model, we found the superior classification ability of CNN can be explained by its 91 and 118 neurons were performing better than the MFF criterion under the measurements of Euclidean and Chi-square distance, respectively. In order to further interpret CNNs' predictions, we tested an improved class activation mapping (ICAM) method to visualize the important regions in the MRI images. With this ICAM, CNNs are able to locate the most discriminative texture patterns of DMD in soleus, lateral gastrocnemius, and medial gastrocnemius; for CMD, the critical texture patterns are highlighted in soleus, tibialis posterior, and peroneus.

## 1 Introduction

Muscular dystrophy (MD) represents a diverse group of genetic diseases often caused by either the absence or mutation in key structural proteins. Among the MD diseases, Duchenne muscular dystrophy (DMD) [4] and congenital muscular dystrophy (CMD) [3] make up a large proportion of the pediatric subtypes. Since DMD and CMD result from different genetic mutations, it is anticipated in order to treat these patients, different therapies will be required. However, DMD and CMD both present with progressive loss of skeletal muscle and function with similar symptom of muscle weakness but appear different on muscle MRI [28]. To automatically or objectively differentiate CMD from DMD, it is crucial to locate image regions that contain specific texture patterns. More importantly, the accurate texture pattern localization could become the basis for monitoring subtle changes over time in both disorders.

Considerable effort has been devoted to cross-sectional imaging, *i.e.*, the magnetic resonance imaging (MRI), to improve the clinical characterization of MD. MRI provides accurate guidance to target the specific muscle parts for tissue biopsy and histological or gene expression analysis [23]. In addition, since MRI is sensitive to abnormal fatty infiltration, it has been widely applied to assess consistent changes in both DMD and CMD [11, 27, 28, 30, 31]. Specifically, [27] demonstrated sub-clinical muscle involvement

*Preprint, to appear in Pattern Recognition (https://doi.org/10.1016/j.patcog.2018.08.012)

by measuring the fat composition of muscle tissue in MRI scans. In [31], a three-point Dixon MRI technique [8] is adopted to measure the amount of lipid-infiltration in the thigh muscles, showing that a quantitative measure of muscle adiposity correlates better with disease severity than the traditional strength measurement. To quantify the fat infiltration in pelvic and thigh muscles, [11] proposed a biomarker, *i.e.*, mean fat fraction (MFF), to assess the disease severity of DMD patients. To make the MFF a voxel level measurement of disease severity, [28] measured MFF of chemical shift-based MRI scans basing on the three-point dixon technique. However, compressing a large 3 dimensional MRI volume into a single MFF global value is a significant data loss as it may overlook some important local information.

Intuitively, implementing texture analysis on medical images [20, 33] can make full use of the available information and it has the potential to provide a more percise assessment of MD progression. To this end, we have explored advanced texture analysis methods for MRI image classification and found deep learning model, *i.e.*, convolutional neural networks (CNNs) [17], delivers the best performance in comparison to its non-deeplearning counterparts.

## 2 Related Work

Using MRI for MD progression assessment and disease subtype classification has received tremendous attention in the past few decades. Mercuri *et al.* [22, 23] proposed a grading system, in which image textures that relate to early and late disease stages are described as "moth-eaten" and "washed-out" patterns, respectively. This grading system relies on radiologists for image texture interpretation. Duda *et al.* [9] presented a semi-automated method, which exploits statistical texture analysis in the manually annotated regions of interest (ROIs) in images. They verified their method only on golden retriever dogs with muscular dystrophy. Kammoun *et al.* [16] applied a similar statistical texture analysis method on human subjusts. They also used manual ROIs to locate texture patterns before the statistical analysis [9, 16]. However, the ROIs given by human observers often suffers from inter-observer variations.

CNNs can be applied to image classification without ROI extraction in advance. To interpret the insight of CNN models, Zeiler and Fergus [32] provided a deconvolutional method and found CNNs make predictions based on the content in certain image subregions. Thus, CNNs would implicitly and automatically detect ROIs when it processes input images. To verify whether the automatic ROI is aligned with manual ROI, several methods have been proposed for visual saliency detection [13]. For instance, Simonyan *et al.* [25] used saliency maps to highlight the important subregions, which are detected by the CNN model. Zhou *et al.* [34] proposed a classification activation map (CAM) to visualize the regions that contain distinctive object parts, but CAM delivers only rough localizations and more detailed ROIs are needed for MD classification.

For CNN-based MD classification, another issue might be that training CNNs with full supervision often result in overfitting due to limited training data. However, pre-trained CNNs have been shown to contain general-purpose feature extractors [5, 12, 24], which are transferable to many other domains. Anthimopoulos *et al.* [2] used a natural image pre-trained CNN to detect texture patterns of interstitial lung diseases in chest CT scans. The CNN with transfer learning reports better classification performance than methods with hand-crafted features. To alleviate the effects of overfitting, Wang *et al.* [29] fine-tuned a pre-trained CNN for breast mass classification.

## 3 Methods

In this section, we first present the collected MRI dataset, which exhibits different image textures for distinct subtypes of diseases and thus potentially enables preliminary diagnosis based on image textures. We then train CNN models with this dataset for automatic image classification, *i.e.*, DMD, CMD, and normal. Finally, we introduce the proposed improved class activation mapping (ICAM) to locate important image sub-regions and extract the learned texture patterns.

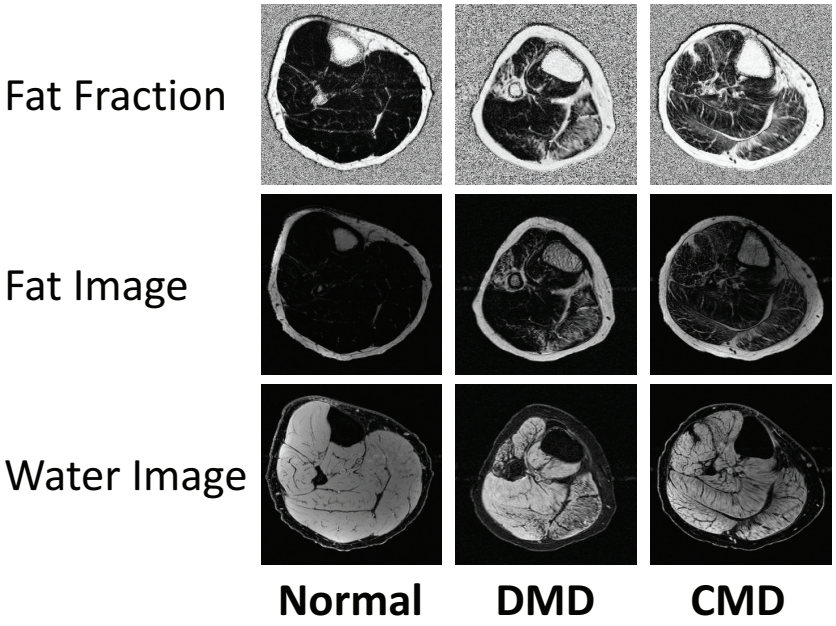


**Figure 1.** Muscular dystrophy MRI images: cross section of low leg acquired with the chemical shift-based water-fat separation MRI [28]. The normal, DMD, and CMD cases are displayed in columns from left to right, respectively. Images listed from top to bottom are the fat-fraction, fat base, and water base signals, respectively.

### 3.1 Dystrophic MRI Dataset

MRI scans from 42 subjects (16 DMD, 13 CMD, and 13 Normal) aging from 5 to 55 years old, which were collected as part of NIH funded natural history studies of DMD and CMD. From each subject, chemical shift-based water-fat separation MR imaging [8] scans of the right lower leg have been acquired on a 3T Phillips scanner. An 16 channel transmit/receive coil is used to acquire unipolar gradient echo images (TE=8.06; 9.21; 10.36 ms; the number of slices: 16-35; slice thickness: 4mm; flip angle: 20 degrees) [28]. In total, 68 MRI scans have been collected containing 20 DMD, 35 CMD, and 13 normal scans. For stability analysis, we randomly split MRI scans at subject level and run experiments under 4-fold cross-validation protocol.

Unlike the standard MRI sequences, where the signal intensity for every voxel (the measurement unit) is jointly determined based on the fat and water signal intensities within that voxel, the three-point Dixon imaging technique [8] allows separation of MRI signal intensity values for individual contributions of fat and water in each voxel of tissue. In this scenario, high-resolution water and fat maps are acquired and enable quantifying the fat fraction of individual muscles [10]. As shown in Figure 1, fat-fraction images are obtained by taking division between the water and fat signals in corresponding spatial position. We observe that DMD and CMD exhibit fat replacement of muscle tissue in different degrees comparing to the normal images. Additionally, DMD and CMD contain texture patterns that visually different from each other.

### 3.2 Learning CNN for Image Classification

A CNN is a feed-forward deep learning architecture of stacked layers with different functions, typically convolutional, activation, pooling, and dense (or fully-connected) layers. In order to resemble the receptive field in the human visual system, each neuron in the convolutional layer is connected to only a small local area of the input. Outputs of convolutional layers are then fed into an activation layer to introduce model nonlinearity. A pooling layer is used to subsample the output from its previous layer so that the size

of the receptive field would increases gradually from bottom to top layers. Finally, several dense layers are used to produce classification results. Thus, neurons from the top layers of a network could have very large receptive fields (*e.g.*, fc7 in VGG-16 [26] has a $404 \times 404$ receptive field), which is sufficient to capture useful contextual (texture) information in the MRI inputs. Here we fine-tune pre-trained CNNs of various model architectures for MRI image classification. Once a promising classification performance is achieved, *e.g.*, greater than 90% accuracy, the corresponding structure will then enable us to have a direct analysis of neurons in the top layers with respect to different types of muscular textures.

**Data Augmentation:** Typically, in order to generalize well for unseen images, CNN model training would require a large amount of annotated images. Thus, we implement data augmentation for the collected MRI images. Specifically, we first augment the images by resizing them with a size-factor, which is randomly generated from the range of $[1.0, 1.2]$. We then randomly rotate the images with an angle in the range of $[-5, 5]$ degree. To augment image contrast, we scale image intensities in pixel-wise by a scale-factor, which is randomly selected in the range of $[0.8, 1.2]$. Finally, a $224 \times 224$ sub-region is cropped for model training. In addition, to take advantage of context information of the 3D MRI scans, 3-consecutive axial slices from an MRI volume are stacked as a unified input image.

### 3.3 Key Region Localization

It is critical to know what and where in the input images enable the CNN to make correct predictions. Therefore, we propose to locate the key regions in MRI images.Our method is mainly inspired by the classification activation map (CAM) approach [34], which is designed to identify discriminative regions inside the input images. The CAM for an image category highlights discriminative image regions for identifying the corresponding category. For example, the CAM of a cat would highlight the image region that contains a cat.

The region localization in the original CAM is coarse, because the heatmap is first obtained at a low resolution (*i.e.*, $7 \times 7$) and then upsampled to have the same size as the input image. In our case, the input is $224 \times 224$, and thus significant spatial details are lost in CAMs. The first column of Figure 4 shows some exemplar heatmaps where too large areas have been covered to locate the specific texture patterns in individual muscles. In order to obtain heatmaps with a higher spatial resolution, [34] removed top layers from the original CNN architecture, but this in turn degrades the CNN's classification performance.

We can simply remove the top layers of the network, *i.e.*, ResNet18 [15], to improve the CAM resolution from $7 \times 7$ to $28 \times 28$. However, this clipping will lead to a significant performance drop in the image classification accuracy. Given the CNN as a hierarchically layered architecture, filters in its lower layers operate in small image regions to detect local image features such as edges and boundaries. With the image information forward propagated, filters in top layers operate on combinations of the local image features to detect semantic object parts, *e.g.*, muscle texture and the conjunction area between muscles. In other words, the activation maps from lower layers highlight fine-grained image features that benefit accurate texture localization. The activation maps from the top layers highlight semantic information that is critical to image classification. Instead of removing top layers, it is intuitive to combine activation maps from both lower and top CNN layers to generate fine-grained CAMs and preserve the classification accuracy. To this end, we propose an improved CAM (denoted by ICAM), which is able to deliver detailed heatmaps without a noticeable sacrifice in classification performance. We graphically illustrate the network architecture of ICAM in Figure 2.

To formally describe the proposed ICAM, we denote $f_k^n(x, y)$ the activation of $k$-th unit in the $n$-th convolutional layer at spatial location $(x, y)$. Note that the spatial resolutions could be different among convolutional layers. Similarly, the result of performing a global average pooling (GAP) is defined as $F_k^n = \sum_{x,y} f_k^n(x, y)$. Assume that there have $M$ and $N$ output channels in layer $m$ and $n$ respectively. For a given image category $c$, the input to soft-max layer $S^c$ is,

$$S^c = \sum_{k=1}^{M} w_k^c F_k^m + \sum_{k=1}^{N} w_{(k+M)}^c F_k^n, \tag{1}$$

where $w_k^c$ is an entry from the $k$-th row and the $c$-th column of a parameter matrix $\mathbf{W} \in R^{(M+N)\times C}$ in

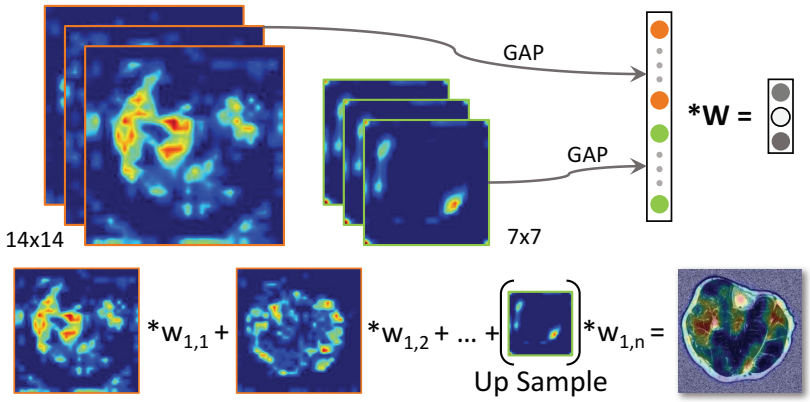


**Figure 2.** The architecture of the proposed improved class activation mapping (ICAM). In the **network** part, feature maps with high (*i.e.* $14 \times 14$) and low (*i.e.* $7 \times 7$) resolutions are fed into the global average pooling (GAP) and classification layers with parameter matrix $\mathbf{W}$. In the **generator** part, low-resolution feature maps are first upsampled with bilinear interpolation to have the same size as the high resolution feature maps, and then combined with the weights learned in the classification layer to generate activation maps.

the softmax layer. Thus, ICAM for class $c$ is defined as,

$$M^c(x,y) = \sum_{k=1}^{M} w_k^c f_k^m + \sum_{k=1}^{N} w_{(k+M)}^c U(f_k^n, 2), \tag{2}$$

where $U(\cdot, 2)$ is the function to upsamples featuremap $f_k^n$ by a factor of 2 to have the same resolution as $f_k^m$ via bilinear interpolation.

### 3.4 Network Architecture and Training Details

We use the VGG16 network [26] with 16 layers, which has demonstrated its effectiveness on the natural image classification task, as the baseline of our automatic MRI images classification. The VGG16 network is pre-trained on the ImageNet dataset [7]. Then, the original dense layers are replaced with a GAP [19] layer followed by a fully-connected layer where the output dimension is set to 3. We fine-tune the modified network on our MRI dataset using Tensorflow [1]. During model training, VGG16 is updated via stochastic gradient descent (SGD) and the mini-batch size is 8. The learning rate is set as $10^{-4}$ for the first $4,000$ iterations and then decreases to $10^{-5}$ for the later $1,000$ iterations.

Meanwhile, residual networks [15], which is a recent deep learning model that produces the state-of-the-art performance in many challenging applications, are also used for MRI image classification. Since the depth of residual networks is flexible, the optimal model size needs to be adjusted for the specific task. Intuitively, an easy task would require a small-sized network and a complicated one would need a larger model. In our experiments, we test residual networks with 18, 34, and 50 layers [15], namely ResNet18, ResNet34, and ResNet50, aiming to seek the best architecture for our application. Each of the residual network variations is pre-trained on ImageNet and fine-tuned for 3-category classification. The same training hyperparameters as those of VGG16 are applied.

### 3.5 Implementation of Non-deep Learning Counterparts

To compare with non-deep learning methods, we apply the K-Means-based texture analysis [6], the statistical texture analysis [16], and the mean fat factor (MFF) [11] for the MRI image classification.

The K-Means-based method is proposed in [6], which presents an effective descriptor for texture patterns to differentiate the foreground from the background. We implement this method for texture

| Method | Test accuracy (%) | | |
|---|---|---|---|
| | Mean | Max. | Min. |
| **MFF (kNN)** | 48.6 | 55.4 | 39.6 |
| **STP (kNN)** | 50.1 | 57.4 | 45.0 |
| **K-Means (SVM)** | 53.2 | 58.4 | 47.3 |
| **STP (SVM)** | 54.9 | 61.8 | 45.6 |
| **VGG16-FC7 (SVM)** | 80.6 | 86.7 | 54.7 |
| **ResNet18-GAP (SVM)** | 83.7 | 93.2 | 76.5 |
| **VGG16** | 84.1 | 91.2 | 76.5 |
| **ResNet50** | 88.6 | 91.6 | 83.5 |
| **ResNet34** | 89.2 | 93.9 | 84.6 |
| **ResNet18** | **90.7** | **94.2** | **86.1** |

**Table 1.** Image classification comparison: classification results are represented in the form of testing error. With 4-fold cross validation, the mean, maximum, and minimum of testing errors are reported. For non-deep learning methods, classifiers kNN and SVM are implemented.

classification. We first crop image patches from image ROIs which are manually drawn from the soleus, lateral and medial gastrocnemius muscles. We then cluster them by K-Means clustering method. Each of the MRI images is then represented as a histogram of the learned K centroids. Finally, we apply Support vector machine with rbf kernel (RBF-SVM) on these histograms for image classification. We use the implementation of [6] and the number of centroids is set to $K = 900$ for $18 \times 18$ image patches (We also test different image patch sizes and find patches smaller than $18 \times 18$ would not contain enough context information and larger patches would significantly reduce the amount of training data).

For the second non-deep learning method, we extract the statistical texture patterns (STP) from image ROIs as the combination of intensity gray scale histogram and co-occurence matrix used in [16]. Support vector machine (SVM) [18] and k-nearest-neighbor (kNN) [14] are used to classify the statistical texture patterns. We search over different hyperparameters, *e.g.*, the panelty value in SVM and the number of neighbors $k$ in kNN, to find the best settings for classification.

Finally, we use CNN features that extracted from pre-trained CNN models for additional comparisons. We first use the dense layer of the pre-trained VGG16 to represent each MRI image as a vector of 4096 dimensions, denoted as VGG16-F7. We then train an RBF-SVM on the extracted CNN features for image classification. Similarly, ResNet18 coverts each MRI image into a 512-dimensional feature vector, denoted as ResNet18-GAP.

# 4 Results and Analysis

## 4.1 Image Classification

Table ?? shows the image classification results with different models and ResNet18 achieves the best performance. In comparison with non-deep learning methods (*e.g.*, K-Means), ResNet18 boosts the testing accuracy by greater than 30%. Meanwhile, both ResNet18 and VGG16 outperforms their pre-trained counterpars, ResNet18-GAP and VGG16-FC7, respectively, proving that model fine-tuning is important to transfer the pre-trained model towards the medical domain. However, ResNet18 outperforms the baseline deep learning model VGG16 by 6.5%. Compared with ResNet18, ResNet34 and ResNet50 are observed with slight performance degradations. The number of model parameters for ResNet34 and ResNet50 are about 22 and 25 millions respectively, which are two times as many as ResNet18. This may explains that ResNet34 and ResNet50 are more easily over-fitted to training images. We thus design ResNet18 to be the backbone architecture for the following experiments.

## 4.2 Network Analysis

To investigate the reason for ResNet18 to achieve the best performance, we proceed to analyzing the GAP layer. The GAP layer in ResNet18 locates between the last convolutional layer and the dense classification layers that transform $7 \times 7$ feature maps into the averaged pixel values [19]. A high value in the output of GAP layer indicates its corresponding network path has a strong response to textures in the input image. In ResNet18, we define each entry in the GAP layer as a neuron and in total there are 512 neurons. Each neuron, like a "weak classifier", has its own discriminative ability to differentiate input images. For instance, a well-trained neuron may have positive responses for CMD images, negative for DMD images, and nearly zero responses for the normal cases as demonstrated in the box-plot B of Figure 3. We then analyze the neuron responses from two aspects, namely a single neuron's outputs to all input images and all of the neurons reacting to the same input image.

First, we analyze each individual neuron by measuring differences of the neuron's outputs to all input images from the DMD, CMD, and normal subjects. For quantitative measurements, we separate the neuron outputs into three groups corresponding to the image categories. From each response group, we extract a normalized histogram with $n$ (*e.g.*, $n = 20$ in our implementation) bins. Euclidean $\ell^2$-distance and Chi-square $\chi^2$-distance are used to measure similarities between the histograms. More specifically, the $\ell^2$-distance between DMD ($h_d$) and CMD ($h_c$) is:

$$\ell^2(h_d, h_c) = \sqrt{\sum_{i=1}^{N} \left(h_d(i) - h_c(i)\right)^2}. \tag{3}$$

where $h_\cdot(i)$ represents the $i$-th bin value in the histogram. Similarly, the $\chi^2$-distance is:

$$\chi^2(h_d, h_c) = \frac{1}{2} \sum_{i=1}^{N} \left[ \frac{\left(h_d(i) - h_c(i)\right)^2}{h_d(i) + h_c(i)} \right]. \tag{4}$$

where $\chi^2(h_d, h_c)$ is weighted by variables and sample units. Compared with $\ell^2$-distance, $\chi^2$-distance is more sensitive to small histogram values. Thus, they are complementary to each other. Meanwhile, the mean fat-fractions (MFF) [11] is calculated inside the soleus muscle from selected keyframes as the output of a spacial classifier (human annotation), and we compare it with all the other deep learning neurons. We observe that there are 91 and 118 neurons performing better than the MFF criterion under the distance measurements $\ell^2$ and $\chi^2$, respectively.

We then study the responses across all neurons for the common input image and the overview of neuron activations is shown on the left panel of Figure 3. As we can see, more network neurons have the negative responses (*i.e.*, blue columns) to DMD inputs compared with CMD and the normal subjects. Meanwhile, for the normal cases, most of the columns are displayed in white color, which means most of the neurons give near zero responses to the unaffected cases. In fact, inside the muscle of normal cases, there have very little textured regions.

In particular, the right panel of Figure 3 indicates that neurons No.1, No.156, and No.382 serve as markers to differentiate DMD, CMD, and normal cases. For example, neuron No.156 outputs positive values for CMD, negative for DMD, and almost zero values for the normal cases. In comparison, we plot MFF values in the bottom-right panel. Each of No.1, No.156, and No.382 has better discriminative power to distinguish different subtypes of disease than the MFF. In practice, an ensemble of many neurons would deliver a strong classifier for dystrophic image classification, which is verified by its 90.7% accuracy on evaluation cases.

## 4.3 Texture Pattern Localization

In order to demonstrate the effectiveness of the proposed ICAM, we list the image classification results in Table 2 and show qualitative analysis in Figure 4. Comparing with the original CAM framework, ICAM produces better classification performance and more fine-grained localization. In Table 2, ResNet18-14[2]

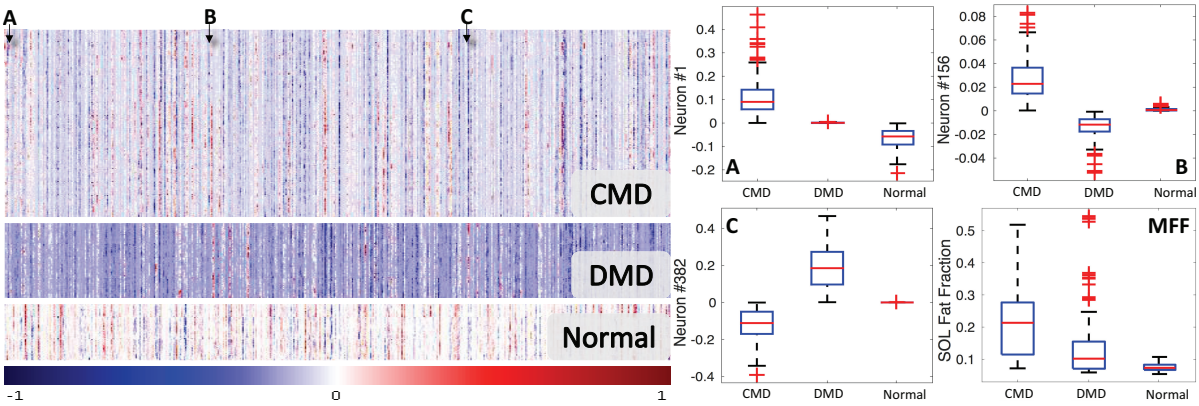


**Figure 3.** CNN output analysis. **left**, each heatmap contains 512 columns corresponding to the 512 "neurons" from the GAP layer of ResNet18. The cross-validation fold with 9% testing error is selected to report. The $1^{st}$, $2^{nd}$, and $3^{rd}$ rows are activation maps for CMD, DMD, and normal cases, respectively. The last row displays the color bar that maps the normalized neuron activations in $[-1, 1]$ to different colors. Activations of network "neurons" No.1, No.156, and No.382 are compared with the mean fat-fraction (MFF) as displayed on the **right** panel.

| MS | Method | Testing Accuracy (%) | | | | |
|---|---|---|---|---|---|---|
| | | Fold-1 | Fold-2 | Fold-3 | Fold-4 | Mean |
| 28x28 | ResNet | 88.7 | 94.5 | 87.8 | 75.6 | 86.6 |
| | ICAM | 88.4 | 90.3 | 90.3 | 77.9 | 86.7 |
| 14x14 | ResNet | 83.2 | 94.5 | 88.4 | 80.8 | 86.7 |
| | ICAM | **93.3** | **96.8** | **93.4** | **84.6** | **91.7** |

**Table 2.** Image classification comparison: to obtain higher map size (**MS**), it requires change of the original network architecture (*e.g.*, removing the top convolutional layers); however, this might degrade classification performance. Here we compare the proposed method (*i.e.* ICAM) with those simply removing top layers (*i.e.* ResNet). Details for model architecture can be found in Section 4.3.

is created by removing layers between the last MaxPooling and GAP layer (including the MaxPooing layer), which delivers heatmaps at a $14 \times 14$ resolution. Similarly, ResNet-$28^2$ is obtained by removing layers between the second last MaxPooing and GAP layer. Meanwhile, ICAM-$14^2$ has the same number of layers as ResNet18-$7^2$ but doubles heatmap resolutions. It is achieved by combining feature maps (map resolution is $14 \times 14$) before its last MaxPooing and feature maps (map resolution is $7 \times 7$) before the GAP layers, as visually depicted in Figure 2. Similarly, ICAM-$28^2$ is modified from the ResNet18-$14^2$ architecture. Both of the ICAM models perform better than their counterparts while delivering doubled heatmap map resolutions (see Table 2).

Figure 4 shows the visual comparison of region localization between ResNet18-$7^2$, ResNet18-$14^2$, ResNet18-$28^2$, and the proposed ICAM-$28^2$ on example DMD and CMD cases. Both ResNet18-$28^2$ and ICAM-$28^2$ generate more detailed heatmaps than ResNet18-$7^2$ and ResNet18-$14^2$, where fine-grained inner muscle areas are highlighted with red. Meanwhile, ICAM-$28^2$ achieves better classification performance than ResNet18-$28^2$, which is evidenced by the example that ResNet18-$28^2$ highlights the wrong muscle area in the white circle in Figure 4 and the highlighted regions by other models are consistent with each other. In addition, Figure 5 shows MRI images from CMD, DMD and normal cases and the highlighted regions are consistent across the entire MRI volume.

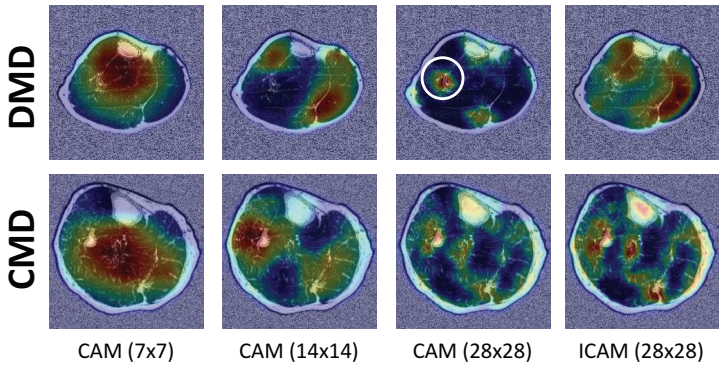


**Figure 4.** Comparison of ROI localization. Importance of image regions from high to low is represented with colors ranging from red to blue. The proposed ICAM with a $28 \times 28$ map resolution has more fine-grained region localization than baselines. The white circle indicates irrelevant regions incorrectly highlighted by the baseline with a $28 \times 28$ resolution.

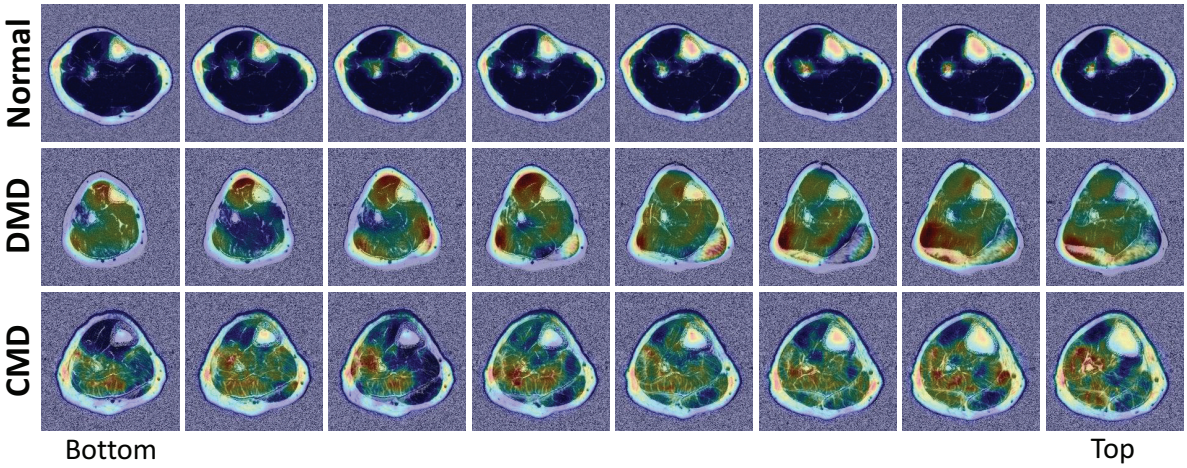


**Figure 5.** Consistency of discriminative region localization. Exemplar MRI sequences from the normal, DMD, and CMD subjects. In each row, the discriminative regions highlighted by ICAM evolve continuously across the consecutive MRI images. In the second row, muscle areas with thick fat tissues are presented as the most discriminative texture in the DMD case. In the last row, the highlighted regions exhibit rich fat textures that are very different from the regions presented in the normal and DMD cases.

### 4.4 Texture Pattern Understanding

In order to answer the following two questions, what kind of texture patterns are related to differentiate subtypes of disease and where these texture patterns appear in affected subjects, we visualize the low-dimensional embedding of detected discriminative regions in Figure 6 and quantify the *muscle texture score* in Table 3 and Figure 7. The detected texture regions are cropped as $k \times k$ image patches from MRI images and then embedded into 2-dimension with t-SNE [21] for better visualization, as shown in Figure 6. We set $k = 24$ so that image patches could be large enough to cover highlighted regions. The most representative image regions are listed at the bottom of Figure 6, 4 from DMD and 4 from CMD cases. These patches indicate what image textures are utilized the most by our ICAM-$28^2$ model to achieve accurate disease subtype classification.

We align the detected texture areas with the lower leg muscles, *e.g.*, lateral gastrocnemius (LG), medial gastrocnemius (MG), soleus (Sol), tibialis posterior (TP), peroneus (Per), extensor digitorium longus (EDL), and tibialis anterior (TA). Since the lengths of muscles are different, some axial MRI slice may contain only a sub-group of the lower leg muscle. However, for ease of illustration, Figure 7 presents

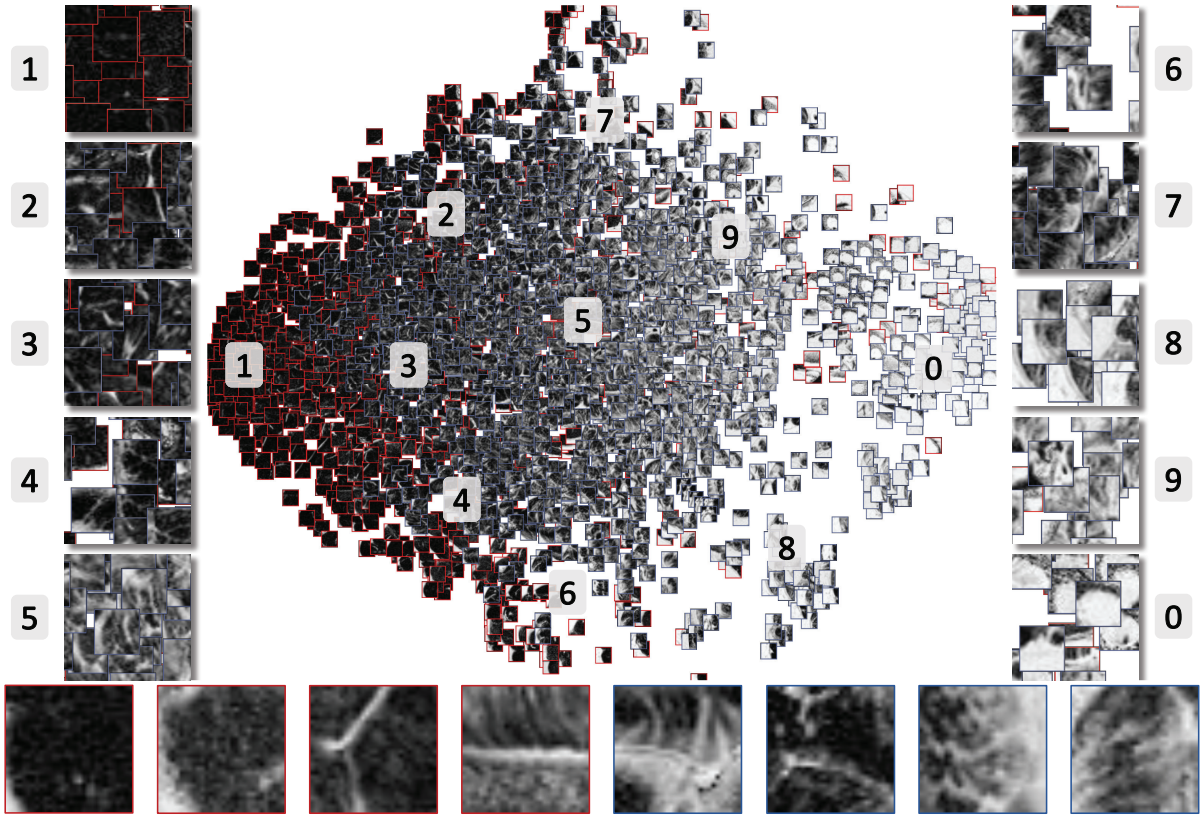


**Figure 6.** Visualization of 2-dimensional embedding of the learned discriminative regions. **top center** depicts discriminative texture regions learned from the DMD and CMD training images projected into a 2D space by t-SNE [21]. **Top sides** presents local details of the 2D embedding. The most representative image patches are shown in the **bottom row**. Cropped subregions from the DMD and CMD images are marked with red and blue color, respectively. Best viewed in color with zooming.

a cross-section with all of the lower leg muscles. The corresponding ICAM emphasizes regions in the EDL, Per, LG muscle area. It indicates that the texture patterns presented in EDL, Per, and LG are critical for CNN models to make the correct predictions. Thus, by producing ICAMs for the MRI scan, we can measure the importance of each lower leg muscle for MD sub-disease classification. Formally, we refer to the "muscle importance" as the *muscle texture score* that is defined as,

$$s_i = n_i/m_i, \tag{5}$$

where $i$ presents the index of a muscle in the lower leg,and $m_i$ is the number of MRI slices that contain the $i$-th muscle, where $n_i$ of them are highlighted by ICAM. $s_i$ ranges from 0 to 1, where 1 means the $i$-th muscle has full textures and 0 represents no texture patterns. The muscle texture scores for selected MD cases are displayed in Table 3. The top-3 textured muscles in DMD are the soleus, the lateral gastrocnemius, and the medial gastrocnemius. For the CMD cases, the peroneus, the tibialis posterior, and the soleus are mostly highlighted. This indicates the locations of useful texture patterns could be utilized to differentiate disease subtypes. These textured muscles could lead to a dystrophic study for further investigation of physiological significance and determine longitudinal changes as a course of disease progression and following therapeutic intervention.

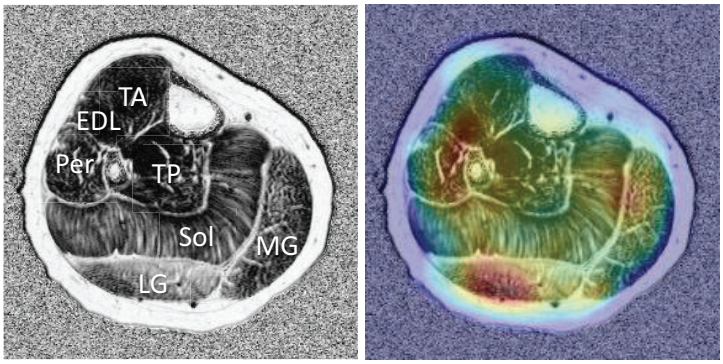


**Figure 7.** ROI context. One representative fat-fraction MRI image of the lower leg from a DMD subject is displayed on the left. This MRI image is utilized to present fat replacement in lower leg muscles. The CNN model predicts this image as DMD based on multiple highlighted regions including EDL, Per, LG, and MG, which is displayed on the right.

| **Cases** | **Muscle Regions** | | | | | | |
|---|---|---|---|---|---|---|---|
| | TA | EDL | Per | TP | Sol | LG | MG |
| | | | | **DMD** | | | |
| **Case No.1** | 0.300 | 0.050 | 0.000 | 0.100 | 0.900 | 0.688 | 0.250 |
| **Case No.2** | 0.188 | 0.625 | 0.625 | 0.375 | 0.688 | 0.750 | 0.750 |
| **Case No.3** | 0.050 | 0.100 | 0.500 | 0.182 | 0.545 | 0.526 | 0.667 |
| **Case No.4** | 0.348 | 0.087 | 0.087 | 0.043 | 0.435 | 0.947 | 0.316 |
| **DMD Mean** | 0.222 | 0.216 | 0.303 | 0.175 | **0.642** | **0.728** | **0.496** |
| | | | | **CMD** | | | |
| **Case No.1** | 0.000 | 0.050 | 0.363 | 0.055 | 0.406 | 0.154 | 0.146 |
| **Case No.2** | 0.031 | 0.062 | 0.719 | 0.375 | 0.031 | 0.063 | 0.063 |
| **Case No.3** | 0.034 | 0.074 | 0.804 | 0.239 | 0.875 | 0.278 | 0.147 |
| **Case No.4** | 0.215 | 0.421 | 0.479 | 0.596 | 0.685 | 0.238 | 0.551 |
| **CMD Mean** | 0.074 | 0.151 | **0.591** | **0.316** | **0.499** | 0.183 | 0.226 |

**Table 3.** Muscle texture scores in CMD and DMD cases. The testing set from the $1^{st}$ fold of the cross-validation is reported.

## 5 Discussion

Apart from the progress reported in this work, the localization results of ICAM still require further validations. We present examples of the ICAM's results for qualitative analysis. We observe the ICAM achieves better classification performance than ResNets because the former is able to locate more discriminative texture regions. We also embed the located image regions for visualization. However, all of these analyses are not a direct quantitative evaluation for the whole MRI dataset. In future work, we will further improve the metric of *muscle texture score* based on the ROIs annotated in the MRI images and measure the consistency between the CNN highlighted regions and human annotations.

## 6 Conclusion

MRI provides a powerful way for noninvasive observation of fat-tissue replacement in muscular dystrophy subjects. In this work, we proposed to fully automate the analysis of dystrophic MRI by using convolutional neural networks (CNNs) for image classification and texture understanding. We tested multiple state-of-

the-art CNN variations on the top of 68 MRI scans. Comparing with the conventional mean fat factor (MFF) and non-deep learning counterparts, CNN models produced superior classification results, among which the best performance is 91.7% mean accuracy over 4-fold-cross-validation. The CNN-based deep learning models are often remarked as data-driven methods, and the current performance is possibly limited by the small amount of data. It is of high potential for CNN models to achieve better results if a larger-sized training dataset is available. Thus one important aspect of our future work is to extend the current data collection with many more study cases for deep model learning.

In this work, we also proposed an effective CNN visualization method, *i.e.*, the improved class activation mapping (ICAM), to visualize the textural regions that are discriminative for disease subtype classification. By visualizing and clustering the highlighted sub-image regions, we verified the hypothesis that there have certain common visual patterns shared by MRI scans of the same disease subtype. These visual patterns are also found to be aligned with lower leg muscles that exhibit high *muscle texture scores*. The proposed ICAM provides an efficient way to understand CNN's predictions and also helps to convince users in terms of the correctness of the automatic classification. However, the validity of current localization results would require further quantitative evaluations. Manual ROI annotations will be acquired and applied to analyze the localization results in our future work.

## References


1. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, and G. Brain. Tensorflow: A system for large-scale machine learning tensorflow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16), pages 265–284, 2016.

2. M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Transactions on Medical Imaging (TMI), 35(5):1207–1216, 2016.

3. E. Bertini, A. D'Amico, F. Gualandi, and S. Petrini. Congenital muscular dystrophies: a brief review. In Seminars in pediatric neurology, volume 18, pages 277–288. Elsevier, 2011.

4. J. Burakiewicz, C. D. J. Sinclair, D. Fischer, G. A. Walter, H. E. Kan, and K. G. Hollingsworth. Quantifying fat replacement of muscle by quantitative mri in muscular dystrophy. Journal of Neurology, 264(10):2053–2067, Oct 2017.

5. K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In British Machine Vision Conference (BMVC), 2014.

6. A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng. Text detection and character recognition in scene images with unsupervised feature learning. In International Conference onDocument Analysis and Recognition (ICDAR), pages 440–445. IEEE, 2011.

7. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. pages 248–255. IEEE, 2009.

8. W. T. Dixon. Simple proton spectroscopic imaging. Radiology, 153(1):189–194, 1984.

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, pages 255–266. Springer, 2015.

10. E. Finanger, B. Russman, S. Forbes, W. Rooney, G. Walter, and K. Vandenborne. Use of skeletal muscle mri in diagnosis and monitoring disease progression in duchenne muscular dystrophy. Physical Medicine and Rehabilitation Clinics of North America, 23(1):1–10, 2 2012.

11. M. Gaeta, S. Messina, A. Mileto, G. L. Vita, G. Ascenti, S. Vinci, A. Bottari, G. Vita, N. Settineri, D. Bruschetta, et al. Muscle fat-fraction and mapping in duchenne muscular dystrophy: evaluation of disease distribution and correlation with clinical assessments. Skeletal radiology, 41(8):955–961, 2012.

12. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 580–587, 2014.

13. J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen. Recent advances in convolutional neural networks. Pattern Recognition, 77:354 – 377, 2018.

14. A. Hafiane, K. Palaniappan, and G. Seetharaman. Joint adaptive median binary patterns for texture classification. Pattern Recognition, 48(8):2609 – 2620, 2015.

15. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

16. M. Kammoun, S. Meme, W. Meme, M. Subramaniam, J. R. Hawse, F. Canon, and S. F. Bensamoun. Impact of tieg1 on the structural properties of fast- and slow-twitch skeletal muscle. Muscle & Nerve, 55(3):410–416, 2017.

17. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

18. S. Liao, M. W. K. Law, and A. C. S. Chung. Dominant local binary patterns for texture classification. IEEE Transactions on Image Processing, 18(5):1107–1118, May 2009.

19. M. Lin, Q. Chen, and S. Yan. Network in network. 2014.

20. X. Liu and J. Tang. Mass classification in mammograms using selected geometry and texture features, and a new svm-based feature selection method. IEEE Systems Journal, 8(3):910–920, 2014.

21. L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. Journal of Machine Learning Research (JMLR), 9(Nov):2579–2605, 2008.

22. E. Mercuri, K. Bushby, E. Ricci, D. Birchall, M. Pane, M. Kinali, J. Allsop, V. Nigro, A. Sáenz, A. Nascimbeni, et al. Muscle mri findings in patients with limb girdle muscular dystrophy with calpain 3 deficiency (lgmd2a) and early contractures. Neuromuscular Disorders, 15(2):164–171, 2005.

23. E. Mercuri, B. Talim, B. Moghadaszadeh, N. Petit, M. Brockington, S. Counsell, P. Guicheney, F. Muntoni, and L. Merlini. Clinical and imaging findings in six cases of congenital muscular dystrophy with rigid spine syndrome linked to chromosome 1p (rsmd1). Neuromuscular Disorders, 12(7):631 – 638, 2002.

24. M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1717–1724, 2014.

25. K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. International Conference on Learing Representations (ICLR), Workshop, 2013.

26. K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations (ICRL), pages 1–14, 2015.

27. S. Sookhoo, I. Mackinnon, K. Bushby, P. Chinnery, and D. Birchall. Mri for the demonstration of subclinical muscle involvement in muscular dystrophy. Clinical radiology, 62(2):160–165, 2007.

28. W. T. Triplett, C. Baligand, S. C. Forbes, R. J. Willcocks, D. J. Lott, S. DeVos, J. Pollaro, W. D. Rooney, H. L. Sweeney, C. G. Bönnemann, et al. Chemical shift-based mri to measure fat fractions in dystrophic skeletal muscle. Magnetic resonance in medicine, 72(1):8–19, 2014.

29. H. Wang, J. Feng, Z. Zhang, H. Su, L. Cui, H. He, and L. Liu. Breast mass classification via deeply integrating the contextual information from multi-view data. Pattern Recognition, 80:42 – 52, 2018.

30. R. Willcocks, W. Triplett, S. Forbes, H. Arora, C. Senesac, D. Lott, T. Nicholson, W. Rooney, G. Walter, and K. Vandenborne. Magnetic resonance imaging of the proximal upper extremity musculature in boys with duchenne muscular dystrophy. Journal of Neurology, 264(1):64–71, 2017.

31. T. A. Wren, S. Bluml, L. Tseng-Ong, and V. Gilsanz. Three-point technique of fat quantification of muscle tissue as a marker of disease progression in duchenne muscular dystrophy: preliminary study. American Journal of Roentgenology, 190(1):W8–W12, 2008.

32. M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In European Conference on Computer Vision (ECCV), pages 818–833. Springer, 2014.

33. Z. Zhang, S. Liu, X. Mei, B. Xiao, and L. Zheng. Learning completed discriminative local features for texture classification. Pattern Recognition, 67:263 – 275, 2017.

34. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2921–2929, 2016.