Computers in Biology and Medicine 105 (2019) 46–53

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Deep embeddings for novelty detection in myopathy

Philippe Burlina[a,b,*], Neil Joshi[a], Seth Billings[a], I-Jeng Wang[a], Jemima Albayda[c]

[a] Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA
[b] Malone Center for Engineering in Healthcare, Baltimore, MD, USA
[c] Division of Rheumatology Johns Hopkins University School of Medicine, Baltimore, MD, USA

ARTICLE INFO

Keywords:
Novelty detection
Deep embeddings
Deep learning
Muscular diseases

ABSTRACT

We address the challenge of finding anomalies in ultrasound images via deep learning, specifically applying this to screening for myopathies and finding rare presentations of myopathic disease. Among myopathic diseases, this study focuses on the use case of myositis given the spectrum of muscle involvement seen in these inflammatory muscle diseases, as well as the potential for treatment. For this study, we have developed a fully annotated dataset (called "Myositis3K") which includes 3586 images of eighty-nine individuals (35 control and 54 with myositis) acquired with informed consent. We approach this challenge as one of performing unsupervised novelty detection (ND), and use tools leveraging deep embeddings combined with several novelty scoring methods. We evaluated these various ND algorithms and compared their performance against human clinician performance, against other methods including supervised binary classification approaches, and against unsupervised novelty detection approaches using generative methods. Our best performing approach resulted in a (ROC) AUC (and 95% CI error margin) of 0.7192 (0.0164), which is a promising baseline for developing future clinical tools for unsupervised prescreening of myopathies.

1. Introduction

Muscle diseases are an important medical problem. While they form a group of relatively rare conditions, they have a disproportionately large impact on affected individuals, with significant morbidity and mortality. In turn, they lead to outsized costs to society due to loss of productivity and high medical care expenditure. It is estimated for example, that population-wide national costs for three muscle diseases (amyotropic lateral sclerosis, Duchenne muscular dystrophy, and myotonic dystrophy) are in excess of two billion dollars per year [1]. Among these conditions, the inflammatory myopathies (Myositis), represent a heterogeneous group manifesting with a variety of presentations, for which treatment is available. They represent an ideal group for study given the spectrum of muscle involvement seen in these diseases, spanning both acute muscle edema, to chronic fatty replacement. As better treatments are sought, enhanced tools for evaluation are needed, particularly for screening and earlier detection of affected individuals.

Imaging plays an important role in myopathy evaluation by providing structural confirmation of muscle involvement that allows for confirmation of disease, evolution, and response to treatment. In this regard, muscle ultrasound has the added advantage of being point-of-care, inexpensive, and easy to use in multiple settings. In the area of genetic myopathies, ultrasound has found valuable use both for diagnosis and follow-up, particularly when using the parameter of increased muscle echointensity which is reflective of muscle damage [2–4]. However, issues of operator dependence and bias with interpretation exist, particularly when compared with the gold standard imaging modality of MRI.

Given the challenges of subjectivity with myopathy diagnostics on ultrasound, machine learning (ML) and computer aided diagnostic methods can play an important role. Originally, automated diagnostic algorithms applied to medical imaging mostly relied on classical ML methods. In the past several years however, deep learning (DL) techniques have had a transformational impact on capabilities of machines to perform medical image classification tasks. This in turn has led to a substantial rise in performance for machine learning methods applied to computer aided medical imaging diagnostics.

These DL techniques, often using deep convolutional neural networks (DCNNs), are, at a high level, quite simple in nature. They implement a cascade of operations which are mostly linear, but also combine with non-linear steps, and other more specialized steps (see more detailed explanations in later sections). These operations essentially implement a "mapping" (a transformation) between the input data (e.g., a medical image) and output values (i.e., a label that describes the disease observed in this image in case of disease

* Corresponding author. Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA.
E-mail address: pburlin2@jhu.edu (P. Burlina).

https://doi.org/10.1016/j.compbiomed.2018.12.006
Received 21 October 2018; Received in revised form 28 November 2018; Accepted 4 December 2018

classification). DCNNs however are different from traditional ML approaches in an important way: traditional ML algorithms employ human-designed image features, whereas the image features computed by DCNNs are machine-learned directly from training data.

Recently, DL approaches have been successfully applied to a wide array of medical image analysis, diagnostic, and prognostic tasks [5–10], and these methods have now mostly supplanted traditional methods [11–13] for these tasks, and reached performance often on par with humans for a range of medical diagnostic problems. However, most studies have worked using supervised assumptions, where all classes of pathologies are assumed to be known a-priori, and where an abundance of clinician-labeled training data is available for each class to train deep learning models.

Some DL techniques can address a moderate lack of training data, or few exemplars (e.g., zero-, one- or adaptive-shot learning [14]). However, a combination of challenging factors for our use case exist. This includes ambiguity and variations in presentation of disease forms on ultrasound, as well as the rarity of myopathies with possible lack of training exemplars. There can also be a total lack of side knowledge, which would prevent use of zero-shot methods, such as [14], for some rare or novel classes of the disease. This problem motivates the need for investigating the use of unsupervised novelty detection (ND) techniques, and principally the latest techniques centered on deep learning approaches. In ND, the assumption is made that some training data is available only for normal control samples, and no labeled training data exists for other classes. The ability of this ND approach to pick out cases outside of normal would find use as a prescreener in clinical situations for referral to clinical experts. Additionally, since this method would not make strong prior assumptions on what form and presentation the disease should take, we hypothesize that novelty detection methods would also be useful from a scientific perspective to allow a system to flag and recognize new or ambiguous presentations of muscle diseases on ultrasound.

Previously reported methods for novelty detection include work such as [15–20]. While most ND methods have centered on classical machine learning, DL approaches have gained attention for ND tasks and these use neural networks to represent images in ways that are more amenable for analysis. This representation is obtained via deep embeddings by processing the image through deep networks. These approaches use either discriminative networks (DCNNs, deep belief networks, and recurrent networks, e.g. Ref. [18]), or generative approaches (via generative adversarial networks [GANs] or variational autoencoders [VAEs] e.g. Ref. [21]). GANs offer several means for embedding where novelty scores can be computed. Prior work in using deep learning for diagnostics for rare diseases such as myopathies has also been done in Ref. [5], but that work addressed the simpler use case of fully supervised binary classification. This does not consider the more difficult use case for which prior observations of affected muscle presentations are rare or non-existent, for which novelty detectors must be used.

Our study has several salient features: (a) we created (collected, curated, and annotated) the first reference dataset including control and myopathy patients for future benchmarking of unsupervised, novelty detection, zero-shot methods (Myositis3K)[1]; (b) we describe DL methods that combine deep embeddings with several novelty scoring methods that are for the first time applied to myopathy cases. We perform comparative performance evaluation with human clinicians and include supervised and generative novelty detection.

[1] This dataset can be made available by request to the authors, after approval by JHU IRB.

Table 1
Distribution of train and test datasets.

	IP - $P_{clinical}$	PP - $P_{clinical}$	IP - P_{image}	PP - P_{image}
total num images:	3586	2142	3586	2064
num train:	698	652	680	658
num train/in:	698	652	680	658
num train/out:	0	0	0	0
num test:	2888	1490	2906	1406
num test/in:	699	745	681	703
num test/out:	2189	745	2225	703

2. Methods

2.1. Data creation and annotation

We use here a dataset that is aimed at studying the presentation of myositis and the development of automatic ML classification algorithms. This dataset and study is part of an overarching project exploring the possibilities of using different deep learning strategies to detect human neuromuscular disorders and the presence of biomarkers that can help characterize those disorders.

Since such a dataset did not previously exist, we constructed this reference dataset from the ground up (and refer to it as “Myositis3K”). It is composed of ultrasound (US) images acquired of seven muscle groups imaged bilaterally per subject, including: deltoids, biceps, flexor carpi radialis, flexor digitorum profundus, rectus femoris, tibialis anterior, and gastrocnemius. All ultrasound images were acquired using one machine, a GE Logiq E system (GE, Fairfield, CT, USA) using a 12 MHz linear array transducer. The US acquisition was performed in a standardized fashion, in cross section, with ultrasound system settings held constant for all study procedures. More details on acquisition procedures are reported in Ref. [5]. This research was performed with Institutional Review Board approval from the Johns Hopkins University and under informed consent from participants.

The complete dataset consisted of images of size 476 × 476 from a total of 89 subjects, including 35 normal/control and 54 with myositis (19 with inclusion body myositis, 15 with polymyositis, and 20 with dermatomyositis). This acquisition resulted in a dataset of 3586 images. See more details on image breakdown based on specific partitioning in Table 1.

A sample of myositis images used in this study are shown in Fig. 1 which displays normal (leftmost image) and abnormal rectus muscle cases (right 3 images). Images in that figure are helpful in demonstrating that the problem of novelty detection can be especially challenging, as there are only very subtle changes between control and affected images, which leads to ambiguity.

2.2. Reference standards, challenge problem formulations, and data partitioning

For this work, two gold standards were considered and created in the Myositis3K dataset (henceforth referred to as “reference standard” (RS) to conform to JAMA nomenclature). One was from an annotation of healthy (0) versus diseased (1) muscles, created by a clinical expert (co-author JA) who took into consideration all known clinical factors including US presentation, muscle strength, muscle enzymes, histopathology, and serology (testing for the presence of antibodies), to ascertain disease diagnosis (see Ref. [5]). This reference standard is referred to as $RS_{clinical}$. The first challenge problem we considered was that of finding anomalies (non-control, class-1 individuals) based on this reference standard, and is henceforth referred to as problem $P_{clinical}$.

A second annotation was also obtained via scoring performed using conditions comparable to the machine, i.e. directly and solely from the

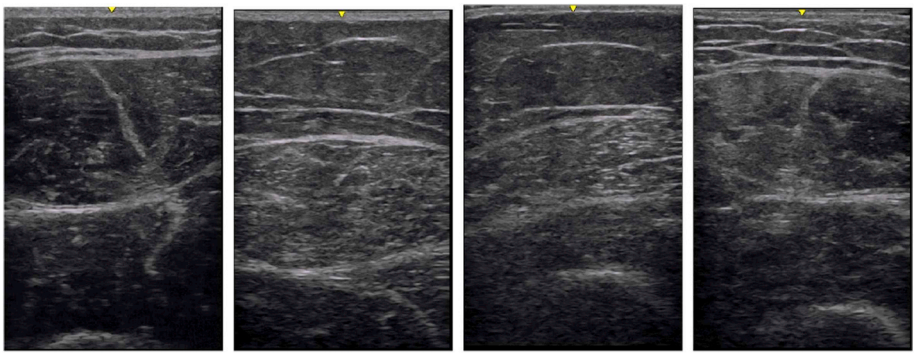

Fig. 1. Examples of myositis images used in this study: normal rectus muscles (leftmost image) versus abnormal rectus muscles (three rightmost images), demonstrating that the presence of only slight pathological changes which can easily lead to ambiguity and constitutes a challenge for novelty detection.

images, and blinded to additional clinical information. This was done by the same clinician (JA), but several months after the initial reference standard $RS_{clinical}$ annotation was done. The clinician was asked to label the images as either normal or diseased, and this was used to flag 'anomalous' images (where anomalous meant diseased). This annotation was done for two purposes: (a) for comparative performance analysis of human and machine ability to carry out novelty detection in problem $P_{clinical}$ and (b) as an alternate reference standard (henceforth referred to as RS_{image}) to be utilized in an additional challenge problem, used in teaching the machine to perform novelty detection based on the reference standard RS_{image}. We refer to the problem aimed at solving ND on RS_{image} as P_{image}. It is important to note, with regard to item (a), that the human clinician had the distinct advantage of prior knowledge (as a result of past experience and expertise) of what 'anomalous' images would look like. In addition to formulating these two problems, we also explored two types of image data partitioning for our experiments. One is image partitioning where the images are assumed to be independent entities. In this partitioning scheme, later referred to as *IP*, images are split into non-intersecting training and testing datasets. Additionally, in *IP*, since control images are less represented than affected ones, we distributed nearly equal numbers of control images (inliers) between the training and testing datasets. A second partitioning scheme was also explored, patient partitioning, whereby no images from the same patient finds its way in both training and testing. The second partitioning scheme is referred to as *PP*. In this scheme, as an alternative, we use a near equal number of inliers and outliers in the testing dataset. Because of this, PP led to a smaller number of images in testing, and also is a more difficult scenario for allowing the machine to perform learning. The resulting numbers of images used in training and testing for these two partitioning schemes and for the two challenge problems considered are summarized in Table 1. We now turn to the description of the algorithmic methods.

2.3. Novelty detection approach

Overall Pipeline. We consider two types of ND algorithms, some based on discriminative DL, and some based on generative methods. All discriminative-based ND algorithms used in this study (i.e., all other than GANomaly) rely on the following pipeline: a) we compute a new representation of the ultrasound images of the muscle via deep feature embeddings. Next we process this feature vector via b) PCA dimensionality reduction, and c) t-SNE (t-Distributed Stochastic Neighbor Embedding). Then we do novelty detection by computing d) a novelty score. Since steps b) and c) are rather conventional (see descriptions in Refs. [22,23]), all other steps (a, d) are described here in detail in the next sections. Step b) PCA is well described and has been traditionally used for dimensionality reduction (for a description see Ref. [24]). Step c), t-SNE, described in Ref. [22] allows for further dimensionality reduction to 2D in a way that maintains patterns and structure of the data. Since t-SNE is a transductive algorithm and applied to both training and testing images together (as is done also here for PCA), this makes the entire algorithmic pipeline transductive. For a discussion on transductive machine learning methods see Ref. [25]. Now we turn our attention to deep embeddings.

Deep Embedding. Deep embedding, step a) in the overall pipeline, creates a new, useful, and compact representation of the muscle ultrasound images, and was produced in this work via processing the image through a DCNN. We then used the output embedded vector as input to the subsequent scoring of the novelty of a sample input compared to 'normal' input images following dimensionality reduction steps b) and c).

We computed deep embeddings by using a discriminative network, here using a pre-trained VGG-16 DCNN [26,27]. The structure of VGG-16 includes a sequence of convolutional blocks:

$$C^1_{(2,64)} \rightarrow C^2_{(2,128)} \rightarrow C^3_{(3,256)} \rightarrow C^4_{(3,512)} \rightarrow C^5_{(3,512)} \rightarrow F$$

followed by fully connected (FC) layers:

$$FC^1_{4096} \rightarrow FC^2_{4096} \rightarrow FC^3_{1000}$$

where $C^i_{(n,k)}$ denotes the ith convolutional block, comprising n successive convolutional layers with filters of spatial size 3×3 and depth k, each of which is followed by a rectified linear unit (ReLU) activation layer, and where each such block is terminated by a pooling layer. This is then followed by a flattening step, F, producing a 4096-length vector which is fed to three successive FC layers, noted above, all on vectors of size 4096. FC embedding uses the output of the last FC^2_{4096} layer and is used as feature vector embedding for the subsequent novelty detection score computation step, described next. These embeddings are deemed to include high-level semantic information.

Novelty Detection Scoring. We describe next the four types of novelty detection scoring methods we used on the embedded feature vectors obtained via the aforementioned embedding method. In what follows, when we refer to "embedded feature vectors" we mean the output of step a) followed by the dimensionality reduction in steps b) and c).

IF:Isolation forest (later termed "IF") was used here as one of the four different ND scoring methods [28]. In this approach, the novelty score is measured using a forest of classification trees (CARTs). The method consists of building a set of random trees (a forest) from the training data consisting of the embedded feature vector inputs for normal muscle images (referred henceforth also as "inliers"). Random trees are built as follows: random features are selected from the embedded feature vector, and splits are computed on these features. The method then exploits the fact that outliers (i.e., feature vectors for muscle images presenting with anomalies), when classified by this tree, typically will stand isolated in a branch close to the root of the tree, with a short path from the node to the element indicating how "novel" the element is compared to the rest of the data. The path from the root to an element, averaged over a set of random trees, is then computed and used as the

Table 2
For problem $P_{clinical}$, performance of IF, EE, LOF, OCSVM (for two different tunings denoted OCSVM1 and OCSVM2 [a manually tuned version of the OCSVM]). 95% error margin for confidence intervals, are reported in parenthesis.

	IP - Average Precision	IP - AUC	PP - Average Precision	PP - AUC
IF	0.8494 (0.0130)	0.6383 (0.0175)	0.6036 (0.0248)	0.6058 (0.0248)
EE	0.8239 (0.0139)	0.5940 (0.0179)	0.5087 (0.0254)	0.5458 (0.0253)
LOF	0.7917 (0.0148)	0.5450 (0.0182)	0.5559 (0.0252)	0.5542 (0.0252)
OCSVM1	0.8711 (0.0122)	0.6833 (0.0170)	0.6666 (0.0239)	0.6489 (0.0242)
OCSVM2	0.8871 (0.0115)	0.7192 (0.0164)	0.6766 (0.0238)	0.6681 (0.0239)
GANomaly	0.8394 (0.0134)	0.6718 (0.0171)	0.5042 (0.0254)	0.4577 (0.0253)

novelty score, since it indicates the degree of isolation and therefore of novelty.

EE: During training, an *elliptic envelope* (later termed "EE") is obtained by fitting a unimodal Gaussian distribution to the embedded feature vectors for the inlier training exemplars. We then use as ND score the Mahalanobis distance from the test element to the Gaussian centroid.

LOF:local outlier factor (or "LOF") was first proposed in Ref. [29]. It works as follows: it considers the set of points including 1) the test element and 2) the inlier training exemplars, in embedded space. Then the LOF computes a novelty score that is based on the relative density of the space surrounding that test point, when compared to its nearest neighbors' densities. It uses the simple hypothesis that, if a point is an outlier, it should stand more isolated in embedded space when compared to its nearest neighbors. Consequently, a measure of isolation (and therefore of novelty) is computed by taking the ratio of the local density measured around that test element, divided by the average density measured around the nearest neighbors of that test point.

OCSVM: Finally, we also used a novelty score based on *one class* support *vector machines* (OCSVM). While there exist several approaches to OCSVM, here we specifically used support *vector data description* (SVDD) (see Refs. [30–32]). The geometric interpretation of this approach is that it bounds the set of training inlier exemplars inside the smallest enclosing hypersphere (with center denoted $\mathbf{a}$ and radius R). Mathematically, finding this hypersphere is done by minimizing the error:

$$F(R, \mathbf{a}) = R^2 \tag{1}$$

with constraints:

$$\|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2, \forall\, i \tag{2}$$

where $\mathbf{x_i}$ are the training exemplars. Skipping derivation steps reported in Ref. [30] and in Appendix A, a score is derived by computing the distance $\mathrm{d}(\mathbf{y})$ from the test element $\mathbf{y}$ to the center of the hypersphere [30]:

$$\mathrm{d}(\mathbf{y}) = \frac{1}{R^2}\left[K(\mathbf{y}, \mathbf{y}) - 2\sum_i \alpha_i K(\mathbf{y}, \mathbf{x}_i) + \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)\right]$$

where $K(.,.)$ denotes a kernel replacing the linear dot product.

Generative methods. As mentioned earlier, GANs offer alternative means of computing embeddings (e.g., latent space) and novelty scores (e.g., directly using the GAN discriminator) when compared to the discriminative networks described previously. We used the recent method GANomaly [21], which was demonstrated to have best performance results against other generative ND methods. This approach uses a conditional generative adversarial network that performs both learning of the generation of synthetic images as well as the generation of latent space embeddings. This is done by using an encoder-decoder-encoder network structure for the generator network, which allows learning the mapping from an input image to a latent space representation. The novelty score is then measured in this latent space.

3. Experiments

In this study, we subdivided the data for the control/unaffected (inlier) and affected (outlier) patients between testing and training for the data partitioning schemes that were either patient-based *PP* or image-based *IP* as described earlier and also detailed in Table 1. Results for problems $P_{clinical}$ and P_{image} are reported below for both partitionings *PP* and *IP*.

For performance characterization and comparison, we report Average Precision as well as the area under the ROC curve (AUC) (Tables 2 and 3). Note that for novelty detection tasks, these are preferable performance metrics compared to conventional metrics like accuracy, sensitivity, and specificity, as these latter metrics are highly dependent on the threshold used by novelty scoring to divide inliers vs. outliers. Better comparisons are also made via ROC curves, which we report in Figs. 2 and 3.

In Table 2 and Fig. 2, we report novelty detection performance for problem $P_{clinical}$ for all novelty scoring methods (IF, EE, LOF, OCSVM), performed on top of the DCNN embeddings and computed against the clinician reference standard $RS_{clinical}$, which is based on the full panel of clinical information (i.e., actual diagnostics using serology, muscle strength, and ultrasound image analysis information). We also compare performance of these ND methods to the performance of a reference algorithm solving a binary supervised classification problem (myositis vs. control) [5] as well as to the performance of a human clinician basing decisions on only image information. These two additional evaluations are shown through the two points plotted on Fig. 2, as computation of Average Precision and AUC was either not possible (in the human evaluation) or was simply not done (in the case of [5]). Finally, we also report the result of ND using the generative approach (GANomaly).

As a second set of experiments dedicated to solve problem P_{image},

Table 3
For problem P_{image}, performance of IF, EE, LOF, OCSVM (for two different tunings denoted OCSVM1 and OCSVM2 [a manually tuned version of the OCSVM]) on image-based ground truth. 95% error margin for confidence intervals are reported in parenthesis.

	IP - Average Precision	IP - AUC	PP - Average Precision	PP - AUC
IF	0.9190 (0.0099)	0.7620 (0.0155)	0.7596 (0.0223)	0.7511 (0.0226)
EE	0.8457 (0.0131)	0.6782 (0.0170)	0.5566 (0.026)	0.6389 (0.0251)
LOF	0.7710 (0.0153)	0.5026 (0.0182)	0.4770 (0.0261)	0.4962 (0.0261)
OCSVM1	0.9445 (0.0083)	0.8292 (0.0137)	0.8281 (0.0197)	0.7855 (0.0215)
OCSVM2	0.9498 (0.0079)	0.8431 (0.0132)	0.8418 (0.0191)	0.8041 (0.0207)
GANomaly	0.8084 (0.0143)	0.6032 (0.0178)	0.4300 (0.0259)	0.3919 (0.0255)

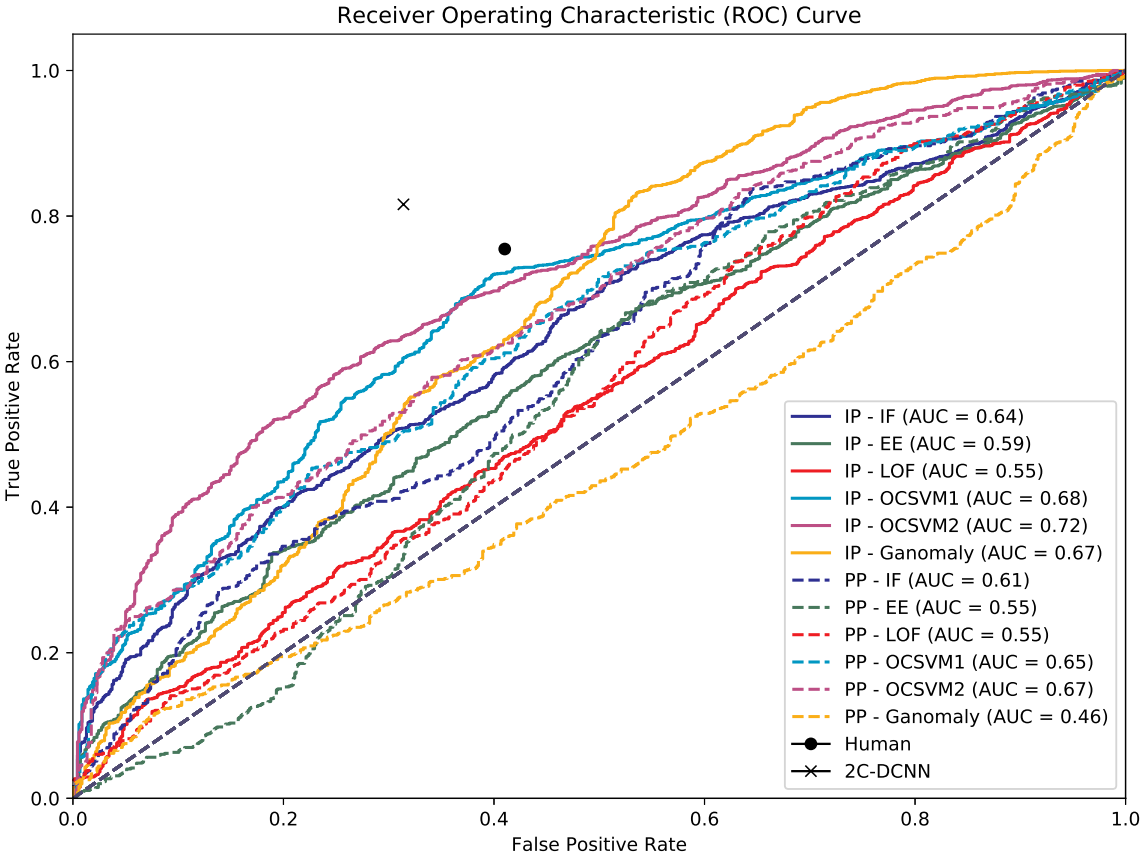

Fig. 2. For problem $P_{clinical}$, ROC curve comparison of all presented methods. Also reported is human performance modeled as a point on the curve. In addition, we report results from [5] (marked as "2C-DCNN") which also solves the simpler binary classification problem.

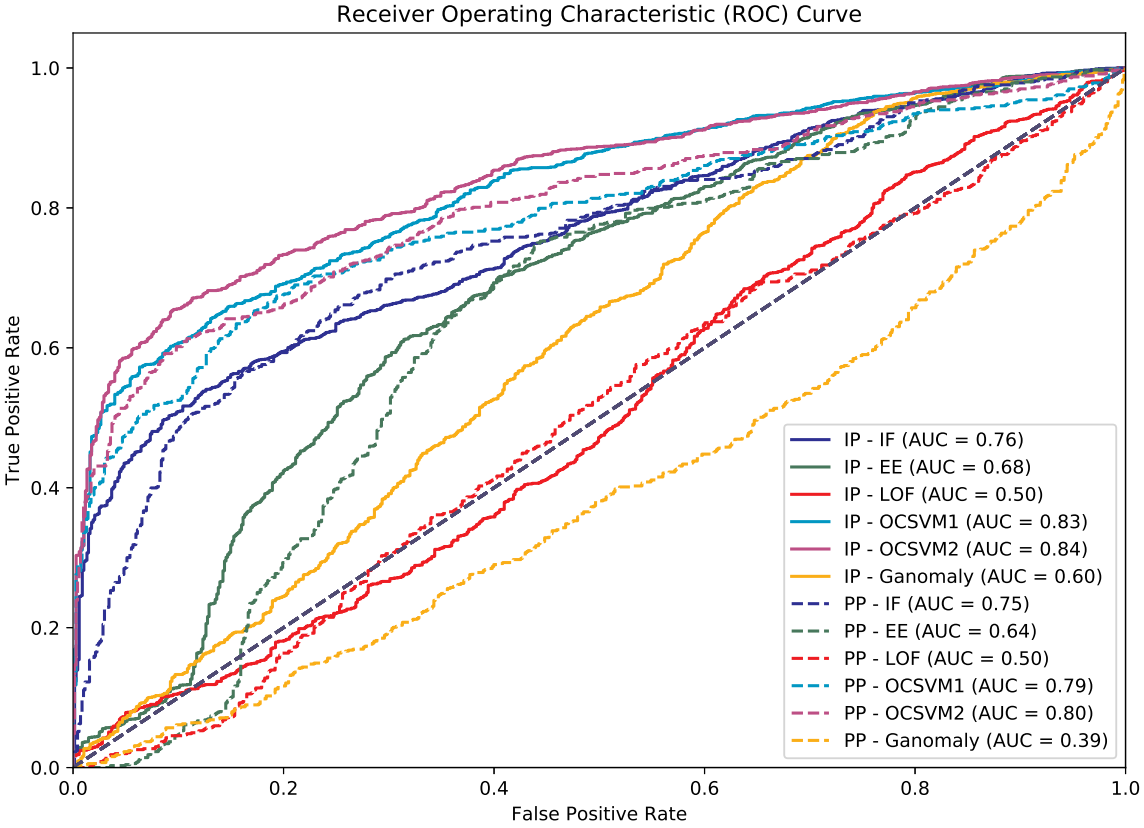

Fig. 3. For problem P_{image}, ROC curve comparison of all presented methods on image-based ground truth. The human annotations evaluated in Fig. 2 are used as ground truth here, which is why a human performance point is not shown. In addition, the 2C-DCNN model in Fig. 2 was never evaluated on this ground truth, which is why it is absent here.

shown in Table 3 and Fig. 3, we used the previously described human annotations RS_{image} as reference standard (and also trained the methods with "normal" exemplars based on this reference standard). We then evaluated each novelty detection approach against this reference standard. This additional experiment is useful to show novelty detection performance on ground truth that is based solely on image data. This problem and experiment is very informative in so far as being able to show the ability of the machine in replicating human ND behavior when that human is basing his/her adjudication on image-only information.

4. Discussion and interpretation of the results

Ultrasound has been found to be a useful tool for screening myopathies, particularly for childhood neuromuscular diseases, with performance similar to EMG for detecting muscle alterations but with improved patient tolerability [33–35]. Although many advantages exist for the use of ultrasound for muscle disease diagnostics, the lack of standardization for muscle evaluation and the lack of refined characterization of sonographic representations of disease have hampered its more widespread use outside of specialty centers. To help overcome this problem, as well as the ambiguity with which myopathy may present on ultrasound, this study has taken the approach of combining computer aided diagnostics and ultrasound imaging with an unsupervised novelty detection formulation of the problem.

This study found promising results for the use of DL techniques for ND applied to myositis. Using the techniques we have employed, the best results were obtained for discriminative embeddings, by OCSVM, then EE, IF, and LOF. In general, the resulting AUC performance appears very promising considering the inherent complexity of the novelty detection problem applied to myopathies and myositis.

Machine results come very close to human performance (as demonstrated in the ROC in Fig. 2). However, human and machine comparisons are not one-to-one, as the machine is asked to solve a pure ND problem and as such is only trained and aware of negative/normal/control examples, whereas human performance is greatly helped by the clinician's prior experience using ultrasound for myositis evaluations and the fact that—by experience—the clinician has learned from binary (control vs. affected) examples, and is therefore implicitly solving an easier binary classification problem.

Results obtained for *PP* partitioning were inferior, as should be expected, compared to *IP*. The *IP* scheme assumes that images are individual entities, while the other assumes that patients are independent entities. Using the *PP* scheme the machine learns less variations in presentation of normal/inlier data and therefore its generalization ability suffers accordingly. The two assessments are useful to bracket the overall expected performance of the algorithms presented.

In this study we also compared performance to prior work using a machine utilizing a nearly identical dataset but solving a simpler binary classification problem using DCNNs (see Fig. 2). Note that this comparison should additionally be qualified by the fact that the partitioning of data in the prior study [5] was different from this study ([5] used muscle partitioning) so the comparison is also not strictly one-to-one. Considering this, the methods proposed herein, which solve a much harder one-class problem, show remarkably promising performance.

In addition, we compared the results to a generative method (GANomaly) [21] using GANs. GANomaly was demonstrated to yield performance that placed it as a best-of-breed GAN implementation of ND, when compared to other generative-based novelty detection methods. In our study this method also demonstrated promising performance with an AUC of 0.6718 (see Fig. 2). This GAN method was able to yield performance which is slightly inferior but comes close to the best discriminative novelty detector performance (AUC of 0.7192 from OCSVM2). Since we have relative paucity of inlier training data in this study, a trade-off must be struck between resolution and reconstruction for the GANomaly detector, with 32^2 resolutions yielding better results for AUC than the 64^2 and 128^2 resolutions (which were also tested but both led to AUCs in the low and mid fiftieth percentiles, not reported here). A head-to-head comparison between generative- and discriminative-based methods however is not possible, since GANs use much less information than the discriminative methods, which use a pre-trained network. Using that as an additional factor makes the use of GANs (and GANomaly here) very appealing. Therefore, future work using GAN-based ND applied to myositis is of definite interest, if the issue of training GANs with less training exemplars can be addressed. GANs have also had other limitations in the past, including dealing with image resolution greater than 128^2. This situation is rapidly changing however, as emerging research has recently lead to the development of GANs able to generate medium to high resolutions images (e.g., ProGANs now allows resolutions up to 1024^2 [36]). In sum, there is a promising path for continued future investigation of the application of GAN-based ND for myositis and myopathies in general, as evidenced by GANomaly results here.

We also ran a set of experiments (problem P_{image}) comparing novelty detection methods when training and evaluating the machine on image-based human annotations (see Table 3 and Fig. 3). This analysis shows performance improvements for most algorithms when compared to the problem $P_{clinical}$, which is as expected since in this experiment humans and machines are provided on par information. This demonstrates that the ND algorithms are able to approximate well the behavior of human clinicians for clinicians performing image-only ND tasks. Considering again that the human clinician had the benefit of years of prior training with both normal and diseased cases, results with AUC of 0.8431 for the best performing ND method indicate that the machine likely has performance characteristics very close to humans, while solving a pure and more difficult ND problem.

Our study has several limitations: we use, as one of our reference standards, the annotation of the clinician utilizing all clinical information and knowledge of the disease diagnosis, with the assumption that all muscles from an affected individual are affected. However, there may be cases where an affected individual has been treated, leading to a return to normalcy for some imaged muscles, or the muscle of a normal subject may have some other type of muscle involvement, which could create noise in the annotation. Blinding the clinician to clinical information and using only image information (similar to the machine), as we did in our second reference standard, would appear to be a better way of ascertaining a discriminator for normal vs. abnormal muscle that is unbiased. However, this too may not provide detection of the true anomalous cases. To be most conservative, only images which were clearly normal have been annotated as such.

In sum, the ability of the machine to learn what is normal, and to then flag what is abnormal, could be of great value in the work-up of a patient with suspected myopathy. These methods could also find applications in cases where access to more sophisticated tools (EMG, MRI, muscle biopsy) or specialists are not immediately available, and could help triage the need for further work-up. Our results in general appear competitive and encouraging for the use of DL techniques as a prescreener of affected muscles in the scenario where new or unseen presentations of the myopathic diseases are encountered. It could help in scientific investigations of yet unknown myopathic cases in an unsupervised learning fashion and help characterize myositis or myopathies in general. Future work should encompass the addition of multiple other myopathies for more comprehensive representations of muscle disease.

5. Conclusion

This study considered the problem of novelty detection applied to rare myopathies, and in particular myositis, observed from ultrasound. We developed a fully annotated reference dataset of more than three thousand images consisting of both control and affected patients. We presented generative and discriminative methods based on deep

embeddings and several novelty scores and evaluated their performance against human and best-of-breed algorithms with promising outcomes.

CBM COI statement

The authors do not have any conflict of interest.

Acknowledgments

This work was supported by the Johns Hopkins Discovery Award, and the Jerome L. Greene Foundation, USA.

Appendix A. OCSVM details

We used a novelty score based on *one class* support *vector machines* (OCSVM), here using support vector data description (SVDD) [30] which models inliers as a mixture of kernels probability density function. Geometrically, it works by enclosing a set of training inlier exemplars using the smallest enclosing hypersphere (with center denoted $\mathbf{a}$ and radius R). Finding this hypersphere is done by minimizing the error function:

$$F(R, \mathbf{a}) = R^2 \tag{3}$$

with constraints:

$$\|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2, \forall\, i \tag{4}$$

where $\mathbf{x}_i$ denote the training dataset of inlier exemplars. Slack variables are then used in the above constraint to allow for some outliers present in the training dataset to violate boundaries, with excess violations expressed via the additive term $\xi_i \geq 0$:

$$\|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0 \quad \forall\, i \tag{5}$$

The original minimization problem including penalties on ξ_i magnitude now becomes:

$$F(R, \mathbf{a}) = R^2 + C \sum_i \xi_i \tag{6}$$

with C a weighting for slack variables. Then, using Lagrange multipliers $\alpha_i \geq 0$ and $\gamma_i \geq 0$, this problem can be restated as that of minimizing L with respect to R, $\mathbf{a}$, $\mathbf{x}_i$, and maximizing L with respect to α_i and γ_i:

$$L(R, \mathbf{a}, \alpha_i, \gamma_i, \xi_i) = R^2 + C \sum_i \xi_i - \sum_i \gamma_i \xi_i$$

$$- \sum_i \alpha_i \{R^2 + \xi_i - \|x_i - \mathbf{a}\|^2\}$$

One can show that ([30]) the problem reduces to minimizing:

$$L = \sum_i \alpha_i (\mathbf{x}_i \cdot \mathbf{x}_i) - \sum_{i,j} \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{7}$$

with constraints in (5). The linear dot product is replaced by a nonlinear kernel $K(\mathbf{x}, \mathbf{y})$ to allow for nonlinear decision boundaries. Minimizing L produces a set of weights α_i for the corresponding samples $\mathbf{x}_i$, as well as the center $\mathbf{a}$ and radius R of the hypersphere. By invoking the complementary slackness constraint, training exemplars with nonzero weights become the *support vectors* of the data defining the bounding hypersphere. In the end, to perform ND inference, and determine if a test element $\mathbf{y}$ is an inlier, one can use the distance of a sample to the center of the hypersphere [30–32] as an effective metric of novelty:

$$\mathrm{d}(\mathbf{y}) = \frac{1}{R^2}\left[K(\mathbf{y}, \mathbf{y}) - 2 \sum_i \alpha_i K(\mathbf{y}, \mathbf{x}_i) + \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)\right]$$

References

[1] Jane Larkindale, Wenya Yang, Paul F. Hogan, J Simon Carol, Yiduo Zhang, Anjali Jain, Elizabeth M. Habeeb-Louks, Annie Kennedy, Valerie A. Cwik, Cost of illness for neuromuscular diseases in the United States, Muscle Nerve 49 (3) (2014) 431–438.

[2] Craig M. Zaidman, Nens van Alfen, Ultrasound in the assessment of myopathic disorders, J. Clin. Neurophysiol. 33 (2) (2016) 103–111.

[3] Sigrid Pillen, O Tak Ramon, J Zwarts Machiel, Martin MY. Lammens, Kiek N. Verrijp, MP Arts Ilse, Jeroen A. van der Laak, M Hoogerbrugge Peter, Baziel GM van Engelen, Aad Verrips, Skeletal muscle ultrasound: correlation between fibrous tissue and echo intensity, Ultrasound Med. Biol. 35 (3) (2009) 443–446.

[4] S. Pillen, A. Verrips, N. Van Alfen, IMP Arts, L.T.L. Sie, M.J. Zwarts, Quantitative skeletal muscle ultrasound: diagnostic value in childhood neuromuscular disease, Neuromuscul. Disord. 17 (7) (2007) 509–516.

[5] Philippe Burlina, Seth Billings, Neil Joshi, Jemima Albayda, Automated diagnosis of myositis from muscle ultrasound: exploring the use of machine learning and deep learning methods, PLoS One 12 (8) (2017) e0184059.

[6] Philippe M. Burlina, Neil Joshi, Michael Pekala, Katia D. Pacheco, David E. Freund, Neil M. Bressler, Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks, JAMA Ophthalmol. 135 (11) (2017) 1170–1176.

[7] P. Burlina, N. Joshi, K.D. Pacheco, D.E. Freund, J. Kong, N.M. Bressler, Utility of deep learning methods for referability classification of age-related macular degeneration, JAMA Ophthalmol. 136 (11) (2018) 1305–1307.

[8] Philippe M. Burlina, Neil Joshi, Katia D. Pacheco, David E. Freund, Jun Kong, Neil M. Bressler, Use of deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration, JAMA Ophthalmol. (2018).

[9] SW Ting Daniel, Yong Liu, Philippe Burlina, Xinxing Xu, Neil M. Bressler, Tien Y. Wong, AI for medical imaging goes deep, Nat. Med. 24 (5) (2018) 539.

[10] Philippe Burlina, Katia D. Pacheco, Neil Joshi, David E. Freund, Neil M. Bressler, Comparing humans and deep learning performance for grading AMD: a study in using universal deep features and transfer learning for automated AMD analysis, Comput. Biol. Med. 82 (2017) 80–86.

[11] Philippe Burlina, David E. Freund, Benedicte Dupas, Neil Bressler, Automatic screening of age-related macular degeneration and retinal abnormalities, Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, IEEE, 2011, pp. 3962–3966.

[12] Albert K Feeny, Mongkol Tadarati, David E. Freund, Neil M. Bressler, Philippe Burlina, Automated segmentation of geographic atrophy of the retinal epithelium via random forests in AREDS color fundus images, Comput. Biol. Med. 65 (2015) 124–136.

[13] Srihari Kankanahalli, Philippe M. Burlina, Yulia Wolfson, David E. Freund, Neil M. Bressler, Automated classification of severity of age-related macular

degeneration from fundus photographs, Invest. Ophthalmol. Vis. Sci. 54 (3) (2013) 1789–1796.
[14] Philippe M. Burlina, Aurora C. Schmidt, I-Jeng Wang, Zero shot deep learning from semantic attributes, IEEE 14th International Conference on Machine Learning and Applications, (ICMLA), 2015, pp. 871–876 2015.
[15] Varun Chandola, Arindam Banerjee, Vipin Kumar, Anomaly detection: a survey, ACM Comput. Surv. 41 (3) (2009) 15.
[16] Marco A.F. Pimentel, David A. Clifton, Clifton Lei, Lionel Tarassenko, A review of novelty detection, Signal Process. 99 (2014) 215–249.
[17] Stefania Matteoli, Marco Diani, Theiler James, An overview of background modeling for detection of targets and anomalies in hyperspectral remotely sensed imagery, IEEE J. Selected Top. Appl. Earth Observ. Remote Sens. 7 (6) (2014) 2317–2336.
[18] Sarah M. Erfani, Sutharshan Rajasegarar, Shanika Karunasekera, Christopher Leckie, High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning, Pattern Recogn. 58 (2016) 121–134.
[19] Diego Carrera, Giacomo Boracchi, Alessandro Foi, Brendt Wohlberg, Detecting anomalous structures by convolutional sparse models, Neural Networks (IJCNN), 2015 International Joint Conference on, IEEE, 2015, pp. 1–8.
[20] Donghwoon Kwon, Hyunjoo Kim, Jinoh Kim, Sang C. Suh, Ikkyun Kim, J Kim Kuinam, A survey of deep learning-based network anomaly detection, Cluster Comput. (2017) 1–13.
[21] Samet Akcay, Amir Atapour-Abarghouei, Toby P. Breckon, Ganomaly: Semi-Supervised Anomaly Detection Via Adversarial Training, (2018) 1805.06725.
[22] Laurens van der Maaten, Geoffrey Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.
[23] Rasmus Bro, Age K. Smilde, Principal component analysis, Anal. Method. 6 (9) (2014) 2812–2831.
[24] Svante Wold, Esbensen Kim, Geladi Paul, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.
[25] Chapelle Olivier, Bernhard Scholkopf, Zien Alexander, Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews], IEEE Trans. Neural Network. 20 (3) (2009) 542–542.
[26] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-scale Image Recognition, (2014) 1409.1556.
[27] Ali Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson, CNN features off-the-shelf: an astounding baseline for recognition, Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on, IEEE, 2014, pp. 512–519.
[28] Fei Tony Liu, Kai Ming Ting, Zhi-Hua Zhou, Isolation forest, Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, IEEE, 2008, pp. 413–422.
[29] Markus M. Breunig, Hans-Peter Kriegel, T Ng Raymond, Jörg Sander, LOF: identifying density-based local outliers, ACM Sigmod Record, vol. 29, ACM, 2000, pp. 93–104.
[30] David MJ. Tax, Robert PW. Duin, Support vector data description, Mach. Learn. 54 (1) (2004) 45–66.
[31] Amit Banerjee, Philippe Burlina, Chris Diehl, A support vector method for anomaly detection in hyperspectral imagery, IEEE Trans. Geosci. Rem. Sens. 44 (8) (2006) 2282–2291.
[32] David E. Freund, Neil Bressler, Philippe Burlina, Automated detection of drusen in the macula, Biomedical Imaging: from Nano to Macro, 2009. ISBI'09. IEEE International Symposium on, IEEE, 2009, pp. 61–64.
[33] A.M. Alanen, B. Falck, H. Kalimo, M.E. Komu, V.H. Sonninen, Ultrasound, computed tomography and magnetic resonance imaging in myopathies: correlations with electromyography and histopathology, Acta Neurol. Scand. 89 (5) (May 1994) 336–346.
[34] K. Brockmann, P. Becker, G. Schreiber, K. Neubert, E. Brunner, C. Bonnemann, Sensitivity and specificity of qualitative muscle ultrasound in assessment of suspected neuromuscular disease in childhood, Neuromuscul. Disord. 17 (7) (Jul 2007) 517–523.
[35] S. Pillen, A. Verrips, N. van Alfen, I.M. Arts, L.T. Sie, M.J. Zwarts, Quantitative skeletal muscle ultrasound: diagnostic value in childhood neuromuscular disease, Neuromuscul. Disord. 17 (7) (Jul 2007) 509–516.
[36] Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen, Progressive Growing of GANS for Improved Quality, Stability, and Variation, (2017) arXiv preprint arXiv:1710.10196.