Daniel R. Clymer
Mem. ASME
Department of Mechanical Engineering,
Carnegie Mellon University,
5000 Forbes Avenue,
Pittsburgh, PA 15213
e-mail: dclymer@andrew.cmu.edu

Jason Long
Department of Musculoskeletal Radiology,
Allegheny General Hospital,
320 E. North Avenue,
Pittsburgh, PA 15212
e-mail: jason.long@ahn.org

Carmen Latona
Department of Musculoskeletal Radiology,
Allegheny General Hospital,
320 E. North Avenue,
Pittsburgh, PA 15212
e-mail: carmen.latona@ahn.org

Sam Akhavan
Department of Orthopedic Surgery,
Allegheny General Hospital,
1308 Federal Street,
Pittsburgh, PA 15212
e-mail: sam.akhavan@ahn.org

Philip LeDuc[1]
Mem. ASME
Department of Mechanical Engineering,
Carnegie Mellon University,
5000 Forbes Avenue,
Pittsburgh, PA 15213
e-mail: prl@andrew.cmu.edu

Jonathan Cagan[1]
Mem. ASME
Department of Mechanical Engineering,
Carnegie Mellon University,
5000 Forbes Avenue,
Pittsburgh, PA 15213
e-mail: cagan@cmu.edu

Applying Machine Learning Methods Toward Classification Based on Small Datasets: Application to Shoulder Labral Tears

Machine learning is a powerful tool that can be applied to pattern search and mathematical optimization for making predictions on new data with unknown labels. In the field of medical imaging, one challenge with applying machine learning techniques is the limited size and relative expense of obtaining labeled data. For example, in glenoid labral tears, current imaging diagnosis is best achieved by imaging through magnetic resonance (MR) arthrography, a method of injecting contrast-enhancing material into the joint that can potentially cause discomfort to the patient, and adds expense compared to a standard magnetic resonance image (MRI). This work proposes limiting the use of MR arthrography through a medical diagnostic approach, based on convolutional neural networks (CNNs) and transfer learning from a separate medical imaging dataset to improve the efficiency and effectiveness. The results indicate an effective method applied to a small dataset of unenhanced shoulder MRI in order to diagnose labral tear severity while potentially significantly reducing cost and reducing unnecessary invasive imaging techniques. The proposed method ultimately can reduce physician workload while ensuring that the least number of patients as possible need to be subjected to an additional invasive contrast-enhanced imaging procedure. [DOI: 10.1115/1.4044645]

Introduction

Machine learning is a rapidly developing field that uses computational models to learn and make predictions on data, usually by means of pattern search and mathematical optimization. By learning and then applying effective patterns from a set of labeled training data, machine learning algorithms learn the underlying rules or structure of the data [1], and can make predictions on new data with unknown labels. In the field of image recognition, deep learning has made tremendous progress at achieving state-of-the-art performance on complex visual recognition tasks [2–4]. One of the most powerful machine learning methods used in visual-based tasks is the convolutional neural network (CNN), which takes an image as input, transforms it through several layers of convolutions (or filters) and dimensionality-reducing pooling operations, and outputs a prediction for the image, often a classification of what is contained in the image. CNNs have the advantage of requiring very little preprocessing or manual feature extraction before training, which is especially important in medical imaging, where most manual image annotation requires a trained expert and can be prohibitively expensive. CNNs applied to medical imaging have exhibited state-of-the-art performance compared to other types of pattern learning algorithms. For example, in the research by Lo et al. [5], a CNN was used to assist radiologists in the identification of lung nodules, effectively improving the diagnosis success rate for early detection of cancer. In the research by Alkabawi et al. [6], deep learning was used for early detection of dementia through CNN analysis of brain imagery. Other examples include computer-aided detection of lymph nodes in computed tomography (CT) images [7], computer-aided anatomy detection in CT scans [8], and brain tumor segmentation in MRI scans [9].

However, there are challenges to utilizing machine learning effectively for medical image recognition tasks. For one, the repercussions for an incorrect classification in medical imaging can be significant, requiring algorithms that exhibit low error rates

[1]Corresponding authors.

Contributed by the Applied Mechanics Division Technical Committee on Dynamics & Control of Structures & Systems (AMD-DCSS) of ASME for publication in the Journal of Engineering and Science in Medical Diagnostics and Therapy. Manuscript received June 18, 2019; final manuscript received August 13, 2019; published online October 21, 2019. Assoc. Editor: Davide Piovesan.

on a wide range of input data. Another common challenge is that in medical imaging, it is often prohibitively expensive to obtain a large labeled training set; therefore, many of the most widespread machine learning applications in medical imaging have come where larger datasets have been made available, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Open Access Series of Imaging Studies (OASIS).

A promising method for applying deep learning to small imaging datasets is called transfer learning, which is a method of training a CNN on a large dataset from a separate application, and then using that pretrained model as a feature extractor or a baseline for learning on the task domain [10–12]. Because deep neural networks learn hierarchical feature representations, with the earlier layers learning more generalized features such as edges [13], the earlier feature sets extracted from the original (or prior) dataset are often useful for reuse across domains and can significantly improve performance on small datasets. Work from Tajbakhsh et al. found that applying transfer learning to several modalities of medical imaging problems resulted in more accurate and robust results for deep learning algorithms, and that the improvements grew more pronounced as the target application's dataset grew smaller [14].

One particular area of interest is in medical imaging of the glenoid labrum. The glenoid labrum is a fibrocartilaginous rim that attaches around the glenoid cavity in the shoulder, providing stability to the joint by extending the "socket" in the ball and socket joint. In athletes involved in overhead sports, such as swimming and baseball, the labrum can be torn, often requiring season-ending surgery. Tears of the labrum can occur anywhere along its insertion on the glenoid but commonly occur either superiorly or along the anterior-inferior glenoid. The most common type of labral tear in overhead athletes is called a superior labrum, anterior to posterior (SLAP) tear. Sometimes, identifying these tears can be difficult with normal medical imaging, primarily because the labrum is subject to a number of normal variants that can mimic a tear. Misidentifying a healthy labral variant as a SLAP tear can subject the patient to unnecessary surgical procedures. For this reason, the "gold standard" of imaging for suspected labral tears is MR arthrography, where a diluted contrast agent is directly injected into the joint. MR arthrography enhances the capabilities of conventional MR imaging through adequate joint distension for improved tissue contrast. Drawbacks of this technique are its invasive character, which adds time and cost to the imaging procedure, as well as risk for infection, bleeding, allergic reaction, and patient discomfort or pain during the procedure [15]. However, compared to the traditional MRI, MR arthrography has been more effective at helping radiologists distinguish labral tears, with accuracies typically ranging between 83% and 94% for the detection of SLAP tears [16,17], compared to typical accuracies ranging from 70% to 83% [18,19] using the traditional MRI.

Because an arthrogram can add time, cost, and patient discomfort compared to an unenhanced MRI, a method that improves diagnosis success rate for physicians using only unenhanced shoulder labral MRIs has the potential to improve hospital workflow while providing a better patient experience throughout the imaging process at a reduced cost, and with reduced risk to the patient. This work utilizes a convolutional neural network algorithm to diagnose shoulder MRI tears, comparing between models trained from random initialization and those initialized with weights learned from a separate medical imaging application, and achieving results on a small dataset comparable to the level that radiologists achieve in practice. This method, shown at a high level in Fig. 1, ultimately can reduce physician workload while ensuring that the least number of patients as possible need to be subjected to an additional invasive MR arthrogram.

Results

Evaluation on Clinical Data. A model was developed and trained to classify between healthy and unhealthy samples on the obtained shoulder MRI dataset. This model was tested by means of a leave-one-out cross-validation procedure (LOOCV) [20], which is a special case of k-fold cross-validation, and is used to estimate the predictive performance of a model. In k-fold cross-validation, the entire dataset is split into k subsets. Of these subsets, a single subset is held out as validation data in order to test the model, which is then trained on the other subsets combined. This is repeated so that each subset is held out one time. For LOOCV, k is equal to the number of total observations in the dataset, such that each observation is held out in the validation set one time. Stone [20] discusses the benefits of increasing the k value in a cross-validation assessment, when it is computationally feasible, such as in the case of a small dataset. For the clinical dataset of 34 patients, one LOOCV procedure trains 34 different models. The size of the training set for each of these trained models, which includes 27 synthetic augmentations for each training sample, was 891 images. The results of each fold of the LOOCV can be averaged to provide a better estimate for the overall expected performance of the model.

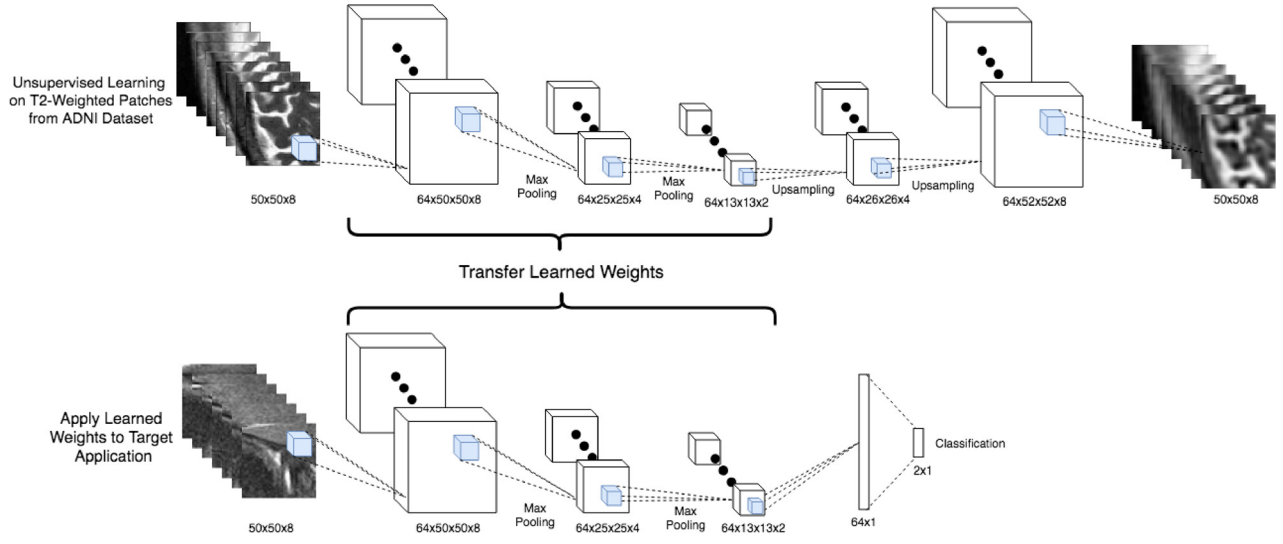


Fig. 1 High-level overview of transfer learning approach used in this work. A convolutional autoencoder was trained on 3D patches from the ADNI medical imaging database [21,22], and the learned weights from the encoding portion of this network were transferred to the target application of shoulder labral tears.

Table 1 Overall top performing algorithm classification results and comparison to radiologists' performance in the literature

Study	Imaging technique	Patients (*N*)	Sensitivity	Specificity	Accuracy
Dinauer et al. [19]	MRI	104	66–85%	75–83%	70–83%
Herold et al. [18]	MRI	35	73%	85%	77%
Bencardino et al. [17]	Contrast-enhanced MR	52	89%	91%	90%
Dinauer et al. [19]	Contrast-enhanced MR	104	84–91%	58–71%	78–87%
Herold et al. [18]	Contrast-enhanced MR	35	91%	85%	89%
Jee et al. [26]	Contrast-enhanced MR	80	84–92%	69–84%	74–86%
Waldt et al. [27]	Contrast-enhanced MR	265	82%	98%	94%
Ours—validation over full dataset—using transfer learning	MRI	34	75.3%, ±0.8%	77.3%, ±0.7%	76.6%, ±0.5%
Ours—validation over top 60% most confident responses—using transfer learning	MRI	34	82.1%, ±0.8%	85.3%, ±0.9%	84.3 ± 0.7%

To explore the effect of hyperparameter selections and model architectures on performance, a grid search, which is an exhaustive search through a manually specified subset of the hyperparameter space of an algorithm, was performed for many combinations of model parameter selections. This search was performed in order to tune the set of model hyperparameters for the given problem which has not been studied with machine learning before, as well as to gain understanding about the performance of the classifier in relation to the complexity and depth of the model. The results of the entire grid search are shown in detail in the Methods section. For each hyperparameter combination, a full leave-one-out cross-validation was performed ten times in order to obtain statistics about the robustness of the results to changes in initialized weights. The best performing model, which is described in more detail in the Methods section, achieved an average validation accuracy of 72.9 ± 0.66%, with a sensitivity (true positive rate) of 71.5 ± 0.98% and a specificity (true negative rate) of 73.8 ± 0.85%. It was also noted that when the predictions for each image were ordered from least to most confident, the accuracies on the subset of most confident predictions increased, suggesting the viability of an approach that withholds the least confident predictions for further inspection. When the least confident 40% predictions were withheld from the best trained model, the overall accuracy on the remaining 60% increased to 77.6 ± 1.0%.

Furthermore, the results from the MRIs that had been less certainly diagnosed by the physician (i.e., labeled by the physician as only suspicious for labral tear instead of definitive labral tear) were analyzed. Across the LOOCV for the best trained model, the average classification accuracy for images labeled "suspicious for labral tear" was 41.3%. These images also were within the 40% least confident predictions 79.0% of the time in the tests that were run. Thus, many of the least confident algorithmic predictions were for the same images that radiologists tended to also be less confident about clinically. This was indicated through the output of the model itself, not through a preclassification into corresponding folds.

Transfer Learning Application. Next, a transfer learning model was built using a separate, larger set of brain MRIs that was obtained from the ADNI dataset[2] [21,22], in order to explore the feasibility of transfer learning between two different three-dimensional (3D) medical imaging datasets, with the goal of learning and extracting relevant features from the larger dataset before applying them to the target application of a smaller dataset. This dataset was selected for transfer learning because it is one of the largest open-access medical imaging datasets available. Since the target dataset in this work is of the T2-weighted modality, a subset of the ADNI database that consisted of 763 T2-weighted MR images was used for the transfer learning process.

[2]www.adni-info.org

Because the MR images in the ADNI database are large and complex, much of the work in the literature with this dataset involves problem-specific preprocesses that extract local measurements or anatomical region labels for classification [23–25]. However, because the goal of this work was to learn generalized features of T2-weighted MR images that could be transferred to the shoulder labral tear application, an unsupervised learning method was employed using a convolutional auto-encoder model, which was trained to reconstruct patches taken randomly from each MRI. Fifty patches were taken from each MRI, resulting in a dataset of 38,150 3D image patches, each of size $50 \times 50 \times 8$ voxels.

These learned weights were then used as the initialized weights for the target labral tear classification problem. A grid search of model hyperparameters was performed for this transfer learning process, the results of which are detailed in the Methods section. The best performing transfer learning process achieved an average LOO cross-validation accuracy of 76.6 ± 0.5%, with a sensitivity (true positive rate) of 75.3 ± 0.8% and a specificity (true negative rate) of 77.3 ± 0.7%. Furthermore, when the least confident 40% of predictions were withheld from the best trained model, the overall accuracy on the remaining 60% increased to 84.3 ± 0.7%, with a sensitivity of 82.1 ± 0.8% and a specificity of 85.3 ± 0.9%. This result, shown in Table 1, is comparable to that of radiologists using contrast-enhanced arthrography [17–19,26,27].

Receiver Operating Characteristic Analysis. A receiver operating characteristic (ROC) curve, which is a statistical method to evaluate the performance of a classification system, was generated for each cross-validation test. In a binary classification, the cutoff between the two expected distributions can be made at any arbitrary point. The ROC curve is made by sweeping through every possible cutoff value (in our case from 0 to 1) and plotting the corresponding true positive rate versus the false positive rate, or the rate that the algorithm correctly classifies torn labrums versus the rate that it misclassifies healthy labrums for every possible cutoff value. The area under the curve (AUC) can be interpreted as the probability that a classifier will rank a randomly chosen positive sample higher than a randomly chosen negative sample [28]. An ROC curve with an area of 1 is considered a perfect classifier, while an area of 0.5 would correspond to a completely random classifier. Often in medical image classification, an area above 0.7 is considered moderate, above 0.8 is considered good, and above 0.9 is considered excellent [29,30].

Figure 2 shows the mean ROC curves from the cross-validation tests with the clinical dataset for both the best performing model trained from random initialization and the best performing model initialized with transferred weights. The AUC for the best randomly initialized model was 0.813 ± 0.004, and the AUC for the best transfer-learned model was 0.848 ± 0.005. Figure 3 shows the mean ROC curves for the best trained models using the entire dataset as well as for a classification on only the 60% most confident model outputs. When only the 60% most confident

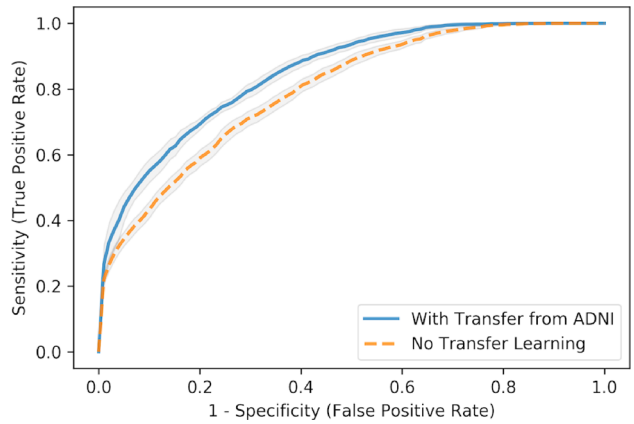
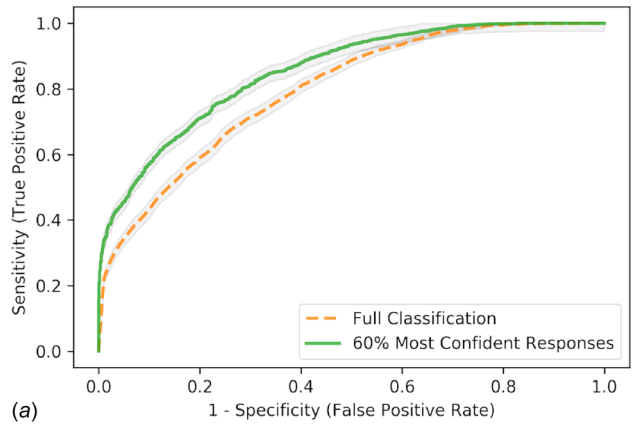
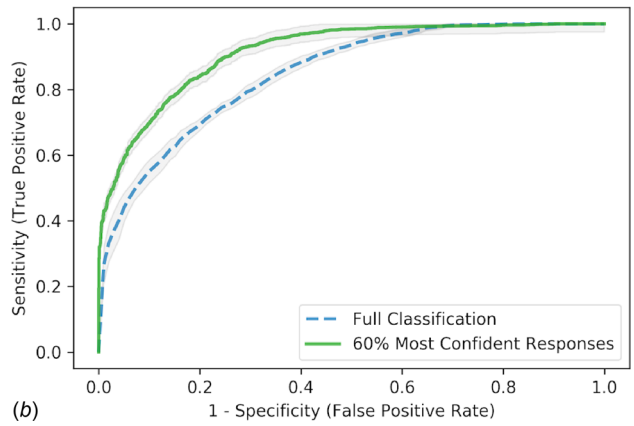


Fig. 2 Mean ROC curves for labral tear detection: comparison between the best performing models trained with and without transfer learning with features learned from the ADNI database



(*a*)



(*b*)

Fig. 3 Mean ROC curves for labral tear detection: comparison between (*a*) the full classification and 60% most confident responses of the best performing network initialized from random initialization and (*b*) comparison between the full classification and 60% most confident responses of the best performing model initialized with transfer-learned weights

classifications were considered, the AUC of the best model trained from random initialization increased to 0.854 ± 0.012, and the AUC of the best transfer-learned model increased to 0.914 ± 0.011. These ROC charts can inform the user on their selection of a classification threshold, to emphasize the true positive rate, emphasize the true negative rate, or remain balanced. For example, in many medical applications a false negative classification is costlier than a false positive, and thus the ideal classifier would choose a threshold weighted to achieve a higher true positive rate rather than a more balanced classifier.

Discussion

Machine learning has been rapidly growing in popularity in recent years, but it has been slower to gain traction in medicine, partly because large annotated datasets can be difficult and expensive to obtain. Furthermore, the cost of misclassification in these fields is often much greater. This paper explores the use of transfer learning from a larger, 3D medical imaging dataset, to address the difficulty of convergence on a small dataset. A method of sorting the algorithm predictions by model certainty is also discussed, with the purpose of effectively pairing physicians with the cases that need the most attention. It is noted that the majority of classification error corresponded to the least certain model predictions, and this error can be identified with no prior knowledge apart from the model's output of classification certainty. This suggests that higher classification accuracies with small datasets can potentially be achieved by classifying only the cases with the most confident model outputs, potentially significantly reducing physician workload and providing more time for the physician to focus on the most challenging cases coming through the system.

From the cross-validation test results, the best model architecture trained from random initialization achieved an overall classification accuracy of 72.9% on the set of unenhanced MRIs. From the hyperparameter grid search, models with fewer trainable parameters tended to achieve higher average accuracies, as expected for a problem with a very small dataset. For a given number of layers and filters, models almost always performed better with fewer nodes in the fully connected layer, a parameter which significantly affects the number of trainable parameters.

Furthermore, the transfer learning models were able to achieve generally higher accuracies, as well as obtain better results, with models that had more layers and more convolution filters per layer. The best performing transfer learning model achieved an accuracy of $76.6 \pm 0.71\%$, which is comparable to accuracies using conventional MR imaging in the literature, between 70% and 83% [18,19]. However, when the transfer learning model only made a decision on its 60% most confident results, the accuracy on that portion improved to 84.3%, a value comparable to even that obtained by radiologists using MR arthrograms, between 83% and 94% [16,17]. This work demonstrates that through effective feature transfer learning, regularization, and the use of viable augmentation methods, convolutional neural networks can be used to effectively classify small, 3D medical imaging datasets.

Furthermore, the cases that were clinically diagnosed with less certainty (i.e., labeled as "suspicious for labral tear") tended to be the ones that were classified with the least certainty. For the best model trained from random initialization, these images fell within the model's 40% least confident classifications 79.0% of the time. Although more data are clearly needed to validate this observation, this result provides evidence that the algorithm framework is achieving the highest confidence with the same images that radiologists are confident with clinically, and suggests the model's ability to create a discriminant boundary, where the classifications with the least prediction confidence are also the cases where the patient has the least definitive clinical condition.

Currently in many hospitals, every young patient suspected of having a labral tear is given an arthrogram to help decide whether shoulder surgery needs to be pursued. With the proposed method in this work, different potential scenarios are possible for this machine learning approach to interface and be helpful to the physician. For example, patients could instead receive a less expensive, less invasive, unenhanced MRI which would be quickly analyzed by the developed algorithm. This analysis would then generate a prediction for the most confident model classifications and suggest further physician review for the least confident classifications. The portion of patients receiving a definitive prediction

could progress with an informed treatment plan without requiring an arthrogram, which can potentially reduce the amount of resources the hospital needs to expend to diagnose and treat patients with suspected labral tears, and potentially reduce risk to those patients. In another scenario, the proposed algorithm could be used as a "second opinion" for radiologists. After a radiologist examines a case and makes a judgment, they could run the case through the algorithm and receive either a confirmation of their diagnosis or a flag for which cases might warrant another look, which would ultimately improve the confidence for a successful radiological diagnosis.

It is noted that the dataset used in this study consisted of patients with various shoulder problems, not necessarily labral tears; thus, the dataset may not be completely representative of a typical population of labral tear patients. However, because a radiologist will observe and comment on each anatomical region of the radiological images being taken, each of these cases contained feedback about the condition of the labrum. Additionally, many of the radiological reports used in this research did not go on to have their findings confirmed surgically, presenting a possible source of labeling error for those cases. This has been partially offset by the most uncertain cases being radiologically labeled as "needing further review," a method used to limit the number of false positives throughout the process. Additionally, because of the constraints caused by a relatively small dataset, the current algorithm only considers input images from the coronal viewing plane, which highlights certain types of labral tears, such as SLAP tears, better than other tears. A future model incorporating images from multiple viewing planes has the potential to further broaden the model's scope of detection capability. However, and importantly, the method introduced in this paper is highly effective with very little original data.

Methods

Dataset. A training dataset consisting of 34 unenhanced shoulder MRI studies was provided by Allegheny General Hospital in Pittsburgh, PA. All methods involving the anonymized acquisition and analysis of this data were carried out in accordance with relevant guidelines and regulations. The Allegheny Health Network Institutional Review Board has reviewed this information and finds it qualifies for exempt status according to the following category in the Code of Regulations: 45 CFR 46.101 (b) Category (Exempt Category 4) and with a "Waiver of HIPAA Authorization." All data were anonymized before being provided to the authors for use in this study. Each MRI from this set is from a patient between 18 and 35 years of age, and all are from patients that received an unenhanced shoulder MRI, and not an arthrogram. Each MRI has a corresponding radiological report on the condition of the labrum, which was reviewed by a radiologist with either 13 or 5 years of experience, and used as the label for training.

From each obtained MRI study, a T2-weighted, fat-saturated image in the coronal view was used in the analysis, because this is an image with a higher specificity for the detection of labral tears that is focused on by radiologists when searching for evidence of a SLAP tear [31]. Examples of some of the shoulder MRIs in this dataset are shown in Fig. 4. These images are slices of 3D models;

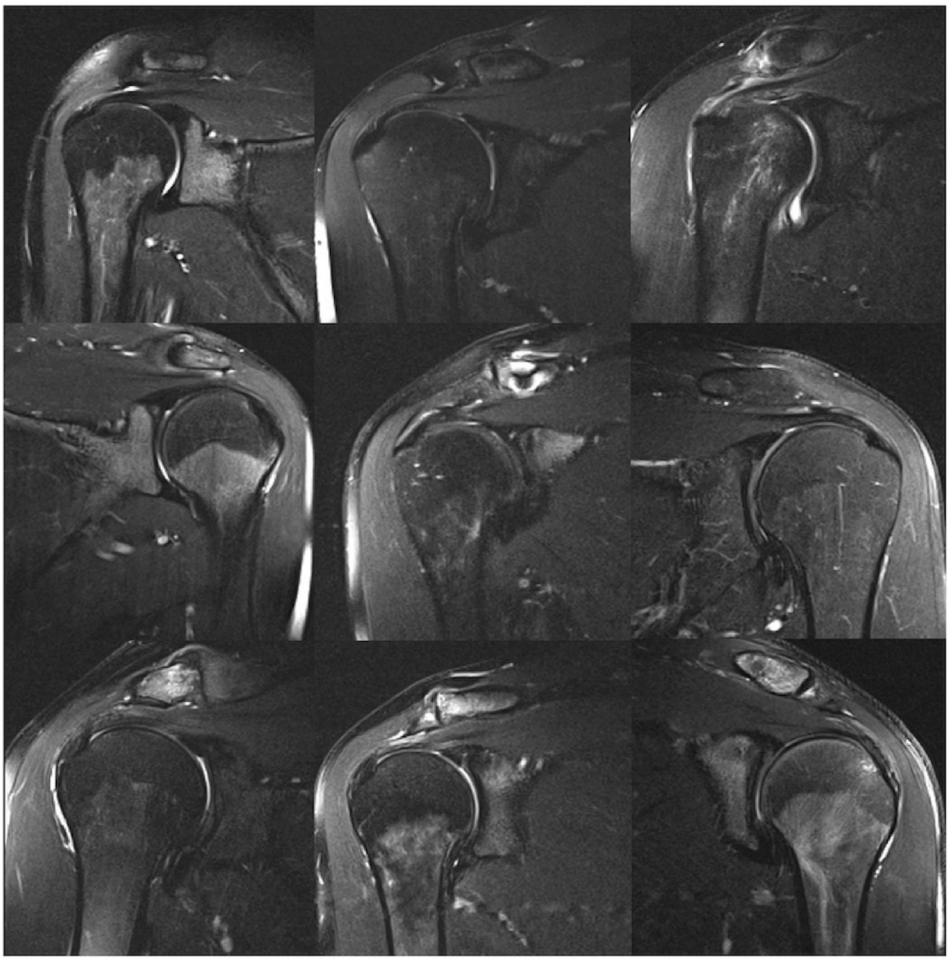


Fig. 4 Examples of one slice from several different patients' 3D shoulder MRIs, used as input into the neural network algorithm

typically around 23 slices are generated per image in the coronal view. Of the 34 MRIs, 13 are labeled by the radiologist as either having a labral tear or having suspicion of a labral tear (nine definitive, four suspicious). Of these 13, four cases went on to receive surgery and all four had their radiological diagnosis confirmed by surgical intervention. There are also three MRIs that are marked as being "motion-degraded," meaning that the patient moved slightly during the MRI and caused some distortion in the image; all three of these studies have been diagnosed by the radiologist as having no evidence of a labral tear. Five patients were found to possess one of the common labral variants (two Buford complexes, two sublabral recesses, and one sublabral foramen).

Preprocessing the Data. To prepare the imaging data for training, a 3D volume is constructed from each MRI's set of two-dimensional slices. In order to improve the contrast of the relevant structures in the image, and because a small number of voxels in each image tend to be far brighter than the rest, the image intensity window level is set to cover 99.5% of the total voxel intensity values. Each image dimension is centered about the training dataset's mean brightness value, so that the relative contrast differences between images, as opposed to absolute brightness values, can be analyzed by the algorithm. The brightness values are normalized by the range of the training dataset. Finally, each image is associated with a label, based on the obtained information from the radiological report. In this study, a binary classification is used, indicating either a healthy or an unhealthy labrum.

Data Augmentation. Because the amount of training data has a significant impact on the success of a machine learning algorithm, it is beneficial to augment the dataset with synthetic data when possible. This can be done by several methods, including Monte Carlo methods [32], variational autoencoders [33], and more recently, generative adversarial networks [34]. However, many of these methods require a large amount of starting data to effectively create synthetic samples. When the starting dataset is very small and the distribution is unknown, a synthetic perturbation method can be effective, which augments the dataset by applying small transformations to the existing datapoints and treating each as a new, independent datapoint. This method has been used effectively to augment handwritten digit image datasets [35], as well as in the Kaggle Galaxy Challenge, to significantly augment a dataset of galaxy images [36].

The perturbation method of synthetic data augmentation has been implemented in this research, and the assumption is made that small rotations and translations made to each image will result in an image with the same classification as the original orientation. Each image is subjected to 26 random operations that are combinations of rotations and translations to the original image, multiplying the size of the training set by 27, and effectively enabling a large enough training set for the CNN to learn and classify over. This data augmentation approach can be effective in medical imaging and other areas where data are expensive or difficult to obtain [37].

Building the Convolutional Neural Network Model. A convolutional neural network model was developed to analyze these training data, which uses a 3D image as input and outputs a binary classification indicating if a healthy or unhealthy labrum was expected. The model was built and trained using the Keras library in Python, with backend calculations being computed with the TensorFlow library. To explore the effect of hyperparameter selections and model architectures on performance, a hyperparameter grid search was performed for many combinations of model parameter selections. The model parameters varied were: number of convolutional layers (1, 2, or 3), number of filters in each convolution layer (16, 32, or 64), and number of nodes in the fully connected layer at the end of the network (16, 32, or 64). The results of this grid search can be found in Table 2. For each hyperparameter combination, a full leave-one-out cross-validation was performed ten times in order to obtain statistics about the robustness of the results to changes in initialized weights.

Table 2 Results of hyperparameter grid search for convolutional neural networks trained from random initialization to perform shoulder labral tear classification on 3D MR images

Number of convolutional layers	Number of filters per layer	Number of nodes in fully connected layer	Average LOOCV accuracy	Area under ROC curve (AUC)
1	16	16	67.7 ± 0.6%	0.753 ± 0.005
		32	68.2 ± 0.8%	0.744 ± 0.006
		64	68.1 ± 0.7%	0.737 ± 0.006
	32	16	68.4 ± 0.5%	0.747 ± 0.008
		32	67.9 ± 0.9%	0.742 ± 0.007
		64	67.1 ± 0.7%	0.730 ± 0.009
	64	16	67.6 ± 0.6%	0.734 ± 0.005
		32	66.9 ± 0.8%	0.729 ± 0.007
		64	66.5 ± 0.8%	0.725 ± 0.010
2	16	16	72.1 ± 0.9%	0.801 ± 0.006
		32	71.6 ± 1.0%	0.786 ± 0.010
		64	69.8 ± 0.9%	0.765 ± 0.007
	32	16	72.9 ± 0.7%	0.813 ± 0.004
		32	71.9 ± 1.2%	0.794 ± 0.009
		64	71.8 ± 1.0%	0.780 ± 0.008
	64	16	72.4 ± 0.6%	0.797 ± 0.005
		32	71.5 ± 0.7%	0.773 ± 0.005
		64	70.3 ± 0.9%	0.769 ± 0.007
3	16	16	71.1 ± 1.0%	0.783 ± 0.009
		32	70.3 ± 0.7%	0.774 ± 0.005
		64	70.9 ± 0.9%	0.762 ± 0.009
	32	16	71.9 ± 0.7%	0.782 ± 0.004
		32	71.4 ± 0.7%	0.764 ± 0.005
		64	71.9 ± 0.6%	0.781 ± 0.08
	64	16	71.4 ± 0.4%	0.791 ± 0.004
		32	71.4 ± 1.1%	0.787 ± 0.010
		64	71.0 ± 0.9%	0.775 ± 0.009

Table 3 Architecture of best performing model trained from random initialization

Layer	Type	Input	Kernel	Stride	Pad	Output
data	Input	$50 \times 50 \times 8$	N/A	N/A	N/A	$50 \times 50 \times 8$
conv1	Convolution	$50 \times 50 \times 8$	$3 \times 3 \times 3$	1	1	$32 \times 50 \times 50 \times 8$
pool1	Max pooling	$32 \times 50 \times 50 \times 8$	$2 \times 2 \times 2$	2	0	$32 \times 4 \times 25 \times 25$
conv2	Convolution	$32 \times 4 \times 25 \times 25$	$3 \times 3 \times 3$	1	1	$32 \times 4 \times 25 \times 25$
pool2	Max pooling	$32 \times 4 \times 25 \times 25$	$2 \times 2 \times 2$	2	0	$32 \times 3 \times 12 \times 12$
fc1	Fully connected	$32 \times 3 \times 12 \times 12$	1×1	1	0	16×1
fc2	Fully connected	16×1	1×1	1	0	2×1

Table 4 Results of hyperparameter grid search for shoulder labral tear classification initialized with feature sets transferred from convolutional autoencoder learned on the ADNI database

Number of convolutional layers	Number of filters per layer	Number of nodes in fully connected layer	Fine-tuned or fixed feature extractor	Average LOOCV accuracy	Area under ROC curve (AUC)
2	16	16	FT	73.2 ± 0.4%	0.808 ± 0.004
		16	FFE	72.8 ± 0.5%	0.802 ± 0.005
		32	FT	74.6 ± 0.7%	0.824 ± 0.08
		32	FFE	74.0 ± 0.7%	0.818 ± 0.006
		64	FT	74.5 ± 0.4%	0.825 ± 0.004
		64	FFE	73.8 ± 0.4%	0.816 ± 0.005
	32	16	FT	75.3 ± 0.4%	0.841 ± 0.006
		16	FFE	74.5 ± 0.4%	0.822 ± 0.004
		32	FT	75.6 ± 0.6%	0.836 ± 0.007
		32	FFE	74.5 ± 0.6%	0.831 ± 0.008
		64	FT	75.4 ± 0.3%	0.837 ± 0.004
		64	FFE	74.7 ± 0.4%	0.825 ± 0.005
	64	16	FT	75.8 ± 0.5%	0.842 ± 0.005
		16	FFE	74.5 ± 0.5%	0.827 ± 0.006
		32	FT	75.3 ± 0.7%	0.837 ± 0.005
		32	FFE	74.9 ± 0.4%	0.828 ± 0.004
		64	FT	75.7 ± 0.6%	0.843 ± 0.007
		64	FFE	74.9 ± 0.8%	0.826 ± 0.007
3	16	16	FT	75.0 ± 0.3%	0.825 ± 0.004
		16	FFE	74.2 ± 0.5%	0.818 ± 0.006
		32	FT	75.4 ± 0.4%	0.834 ± 0.004
		32	FFE	74.8 ± 0.6%	0.831 ± 0.006
		64	FT	76.0 ± 0.3%	0.841 ± 0.003
		64	FFE	75.2 ± 0.5%	0.833 ± 0.004
	32	16	FT	75.9 ± 0.5%	0.849 ± 0.005
		16	FFE	74.9 ± 0.4%	0.828 ± 0.004
		32	FT	75.9 ± 0.3%	0.837 ± 0.005
		32	FFE	75.2 ± 0.5%	0.835 ± 0.006
		64	FT	75.8 ± 0.5%	0.831 ± 0.005
		64	FFE	75.3 ± 0.5%	0.826 ± 0.006
	64	16	FT	76.1 ± 0.4%	0.845 ± 0.005
		16	FFE	75.4 ± 0.4%	0.839 ± 0.004
		32	FT	76.3 ± 0.4%	0.843 ± 0.006
		32	FFE	75.4 ± 0.6%	0.833 ± 0.005
		64	FT	76.6 ± 0.5%	0.848 ± 0.005
		64	FFE	75.5 ± 0.4%	0.834 ± 0.006

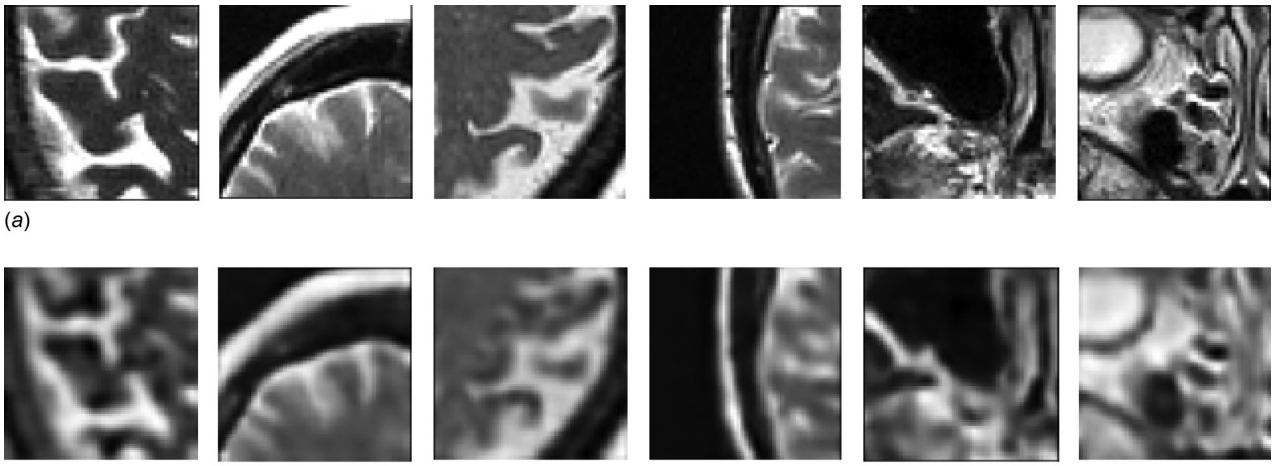
The model architecture that achieved the best results is shown in Table 3. It consists of two 3D convolutional layers and two max pooling layers, followed by a fully connected flattened layer and a binary classification output layer. All convolution layers have 32 kernels, all of size $3 \times 3 \times 3$ with stride 1. The convolution weights are initialized with the uniform distribution proposed by Glorot and Bengio [38], which has been shown to be effective at preventing overfitting in deep network architectures. ReLU activation functions are used for each intermediate layer which has been shown to help avoid vanishing gradient problems in deeper networks [39], and a sigmoid activation function applied for the output. The model is optimized by the Adam optimizer [40], with an initial learning rate of 0.001, and with a binary cross-entropy loss function. Because there is a relatively large imbalance between the two classes in the clinical training dataset, the loss function is "class-weighted" [41,42], that is, each input image is weighted according to its corresponding class' percent representation in the total training set.

Model selection was achieved by saving the network that achieved the highest validation accuracy at the end of each training step. The final layer of the model has two nodes, which output the expected probabilities, between 0 and 1, of the two classifications. When a new image is analyzed by the network, the output node with the higher value is treated as the model's classification for that image. The larger the difference between the two output values, the more confident the model is in its prediction.

The primary challenge in applying CNN models to small datasets is the risk of overfitting, meaning that the model performs well at classifying the training data, but does not generalize well to new, unseen data. To account for this problem, several methods of regularization are used to constrain the model. One method is dropout [43], in which a random selection of nodes in a neural

Table 5 Architecture of convolutional autoencoder model used to transfer learned weights to shoulder labral tear classification

Layer	Type	Input	Kernel	Stride	Pad	Output
data	Input	$50 \times 50 \times 8$	N/A	N/A	N/A	$50 \times 50 \times 8$
conv1	Convolution	$50 \times 50 \times 8$	$3 \times 3 \times 3$	1	1	$64 \times 50 \times 50 \times 8$
pool1	Max pooling	$64 \times 50 \times 50 \times 8$	$2 \times 2 \times 2$	2	0	$64 \times 25 \times 25 \times 4$
conv2	Convolution	$64 \times 25 \times 25 \times 4$	$3 \times 3 \times 3$	1	1	$64 \times 25 \times 25 \times 4$
pool2	Max pooling	$64 \times 25 \times 25 \times 4$	$2 \times 2 \times 2$	2	0	$64 \times 13 \times 13 \times 2$
conv3	Convolution	$64 \times 13 \times 13 \times 2$	$3 \times 3 \times 3$	1	1	$64 \times 13 \times 13 \times 2$
up1	Up-sampling	$64 \times 13 \times 13 \times 2$	$2 \times 2 \times 2$	2	0	$64 \times 26 \times 26 \times 4$
conv4	Convolution	$64 \times 26 \times 26 \times 4$	$3 \times 3 \times 3$	1	1	$64 \times 26 \times 26 \times 4$
up2	Up-sampling	$64 \times 26 \times 26 \times 4$	$2 \times 2 \times 2$	2	0	$64 \times 52 \times 52 \times 8$
conv5	Convolution	$64 \times 52 \times 52 \times 8$	$3 \times 3 \times 3$	1	1	$64 \times 52 \times 52 \times 8$
output	Convolution—linear activation	$64 \times 52 \times 52 \times 8$	$3 \times 3 \times 1$	1	0	$1 \times 50 \times 50 \times 8$



(*a*)

(*b*)

Fig. 5 (*a*) Example slices from the 3D images of the ADNI database used to train the convolutional autoencoder and (*b*) the output predictions from the autoencoder on those slices

Table 6 Architecture of best performing shoulder labral tear classification model using transferred weights from the convolutional autoencoder described in Table 5

Layer	Type	Input	Kernel	Stride	Pad	Output
data	Input	$50 \times 50 \times 8$	N/A	N/A	N/A	$50 \times 50 \times 8$
conv1	Transferred convolution	$50 \times 50 \times 8$	$3 \times 3 \times 3$	1	1	$64 \times 50 \times 50 \times 8$
pool1	Max pooling	$64 \times 50 \times 50 \times 8$	$2 \times 2 \times 2$	2	0	$64 \times 25 \times 25 \times 4$
conv2	Transferred convolution	$64 \times 25 \times 25 \times 4$	$3 \times 3 \times 3$	1	1	$64 \times 25 \times 25 \times 4$
pool2	Max pooling	$64 \times 25 \times 25 \times 4$	$2 \times 2 \times 2$	2	0	$64 \times 13 \times 13 \times 2$
conv3	Transferred convolution	$64 \times 13 \times 13 \times 2$	$3 \times 3 \times 3$	1	1	$64 \times 13 \times 13 \times 2$
fc1	Fully connected	$64 \times 13 \times 13 \times 2$	1×1	1	0	64×1
fc2	Fully connected	64×1	1×1	1	0	2×1

network is ignored at each step during training, causing the network to become less sensitive to any specific weights during training and better able to generalize to new data. Dropout layers of 15% are used after each convolutional layer in the network, as well as a 40% dropout layer before the final output layer of the model. Another powerful regularization method is L2 regularization, which is sometimes referred to as "weight decay." This method adds a term to the optimization that penalizes larger weights, tending to only allow large weights for the most important features in the network [44,45]. L2 regularization is used on the weights of each convolution layer of the model. The influence of this term on the loss function is determined by a linear parameter, λ, which was set at 0.01 for this research.

Transfer Learning From the Alzheimer's Disease Neuroimaging Initiative Dataset. Next, a method of transfer learning was explored in which convolutional features were learned from a subset of the ADNI database and used as the initialization for the target application of shoulder labral tears. Since the target dataset was of the T2-weighted modality, a subset of the ADNI database that consisted of 763 T2-weighted MR images was used for the transfer learning process. Because the MR images in the ADNI database are large and complex, much of the work in the literature with this dataset involves problem-specific preprocesses that extract local measurements or anatomical region labels for classification [23–25]. However, because the goal of this work was to learn generalized features of T2-weighted MR images that could be transferred to the shoulder labral tear application, an unsupervised learning method was employed using a convolutional auto-encoder model, which was trained to reconstruct patches taken randomly from each MRI. Fifty patches were taken from each MRI, resulting in a dataset of 38,150 patches, each of size $50 \times 50 \times 8$ voxels.

A hyperparameter grid search was performed to explore the effect of model parameter selections on the model performance. The model parameters varied were: number of convolutional

layers (1, 2, or 3), number of filters in each convolution layer (16, 32, or 64), and number of nodes in the fully connected layer at the end of the network (16, 32, or 64). Also, each parameter combination was tested using fine-tuning (FT) or fixed training (FFE), signifying if the weights were allowed to be further trained or if they would be fixed after transfer. For each hyperparameter combination, a full leave-one-out cross-validation was performed ten times with different initialized weights each time. The results of this grid search are found in Table 4. In almost all cases, fine-tuning performed better than fixed training for this application.

The autoencoder from the best performing transfer learning model architecture is shown in Table 5. It consists of convolutions and max pooling operations in the encoder portion of the network, and convolutions and up-sampling operations in the decoder portion. Each convolution has 64 kernels and has a ReLU activation function except for the output layer, which has one kernel and a linear activation function. The autoencoder is optimized with the Adam optimizer with an initial learning rate of 0.001 and a mean squared error loss function. The autoencoder was trained until the point that the test loss stopped improving; outputs from the training of this network can be seen in Fig. 5. For the transfer learning step, the learned weights from the convolutions in the encoder portion ("conv1," "conv2," and "conv3") are used as the initialization for the convolutions in target classification model, described in Table 6. This model was trained in the same manner as the models trained from random initialization, except that the learning rate was decreased to 0.00001.

Conclusions

In this work, the viability of machine learning applications for medical image classification and diagnosis for small datasets is examined, and applied to the diagnosis of labral tears in unenhanced MRI. A transfer learning method to extract relevant features between separate 3D medical imaging datasets is explored by means of a convolutional autoencoder and shown to provide better results compared to random initialization. A clinical workflow is discussed in which the machine learning algorithm filters out a portion of incoming patient images to a high degree of accuracy, providing the physician more time to focus on the remaining cases. The results demonstrate that the proposed algorithm appears to not only be able to filter out a large portion of incoming cases, but that the remaining cases tend to be the same ones that radiologists considered harder to diagnose clinically. Achieving model convergence and accuracy on a very small dataset provides optimism that these methods can be effective for analysis of complicated medical images. Future work will apply these methods to other clinically relevant datasets. This work strives toward less invasive, faster, and cheaper MRI processes to diagnose labral tear severity and can also be extended to many other medical imaging and other applications.

Acknowledgment

Data used in preparation of this article were obtained from the ADNI database.[3] As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found online.[4]

Funding Data

- Disruptive Health Technologies Institute at Carnegie Mellon University.
- The Office of Naval Research (N00014-17-1-2566).
- The Pennsylvania Infrastructure Technology Alliance.
- Air Force Office of Scientific Research (FA9550-18-1-0262).
- The Pennsylvania Department of Health (SAP4100077084).

[3] adni.loni.usc.edu

[4] http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

References

[1] Whiting, M., Cagan, J., and LeDuc, P., 2016, "Efficient Probabilistic Grammar Induction for Design," AI EDAM, **32**(2), pp. 177–188.

[2] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., 2015, "Going Deeper With Convolutions," IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June 7–12, pp. 1–9.

[3] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., 2016, "Rethinking the Inception Architecture for Computer Vision," Conference on Computer Vision and Pattern Recognition (CVPR), June 27–30, Las Vegas, NV, pp. 2818–2826.

[4] Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., and Geman, D., 2011, "Comparing Machines and Humans on a Visual Categorization Test," Proc. Natl. Acad. Sci., **108**(43), pp. 17621–17625.

[5] Lo, S.-C. B., Lou, S.-L. A., Lin, J.-S., Freedman, M. T., Chien, M. V., and Mun, S. K., 1995, "Artificial Convolution Neural Network Techniques and Applications for Lung Nodule Detection," IEEE Trans. Med. Imaging, **14**(4), pp. 711–718.

[6] Alkabawi, E. M., Hilal, A. R., and Basir, O. A., 2017, "Computer-Aided Classification of Multi-Types of Dementia Via Convolutional Neural Networks," IEEE International Symposium on Medical Measurements and Applications (MeMeA), Rochester, MN, May 7–10, pp. 45–50.

[7] Roth, H. R., Le Lu, A. S., Cherry, K. M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., and Summers, R. M., 2014, "A New 2.5D Representation for Lymph Node Detection Using Random Sets of Deep Convolutional Neural Network Observations," International Conference on Medical Image Computing and Computer-Assisted Intervention, Boston, MA, Sept. 14–18, pp. 520–527.

[8] Zheng, Y., Liu, D., Georgescu, B., Nguyen, H., and Comaniciu, D., 2015, "3D Deep Learning for Efficient and Robust Landmark Detection in Volumetric Data BT—Medical Image Computing and Computer-Assisted Intervention," International Conference on Medical Image Computing and Computer-Assisted Intervention, Cham, Switzerland, Munich, Germany, Oct. 5–9, pp. 565–572.

[9] Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., and Larochelle, H., 2017, "Brain Tumor Segmentation With Deep Neural Networks," Med. Image Anal., **35**(Suppl. C), pp. 18–31.

[10] Azizpour, H., Razavian, A. S., Sullivan, J., Maki, A., and Carlsson, S., 2015, "From Generic to Specific Deep Representations for Visual Recognition," IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, June 7–12, pp. 36–45.

[11] Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S., 2014, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Columbus, OH, June 23–28, pp. 806–813.

[12] Penatti, O. A. B., Nogueira, K., and Santos, J. A. D., 2015, "Do Deep Features Generalize From Everyday Objects to Remote Sensing and Aerial Scenes Domains?," IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, June 7–12, pp. 44–51.

[13] Olah, C., Mordvintsev, A., and Schubert, L., 2017, "Feature Visualization," Distill, **2**(11), p. 1.

[14] Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J., 2016, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," IEEE Trans. Med. Imaging, **35**(5), pp. 1299–1312.

[15] Magee, T., 2015, "Usefulness of Unenhanced MRI and MR Arthrography of the Shoulder in Detection of Unstable Labral Tears," Am. J. Roentgenol., **205**(5), pp. 1056–1060.

[16] Chang, D., Mohana-Borges, A., Borso, M., and Chung, C. B., 2008, "SLAP Lesions: Anatomy, Clinical Presentation, MR Imaging Diagnosis and Characterization," Eur. J. Radiol., **68**(1), pp. 72–87.

[17] Bencardino, J. T., Beltran, J., Rosenberg, Z. S., Rokito, A., Schmahmann, S., Mota, J., Mellado, J. M., Zuckerman, J., Cuomo, F., and Rose, D., 2000, "Superior Labrum Anterior- Posterior Lesions: Diagnosis With MR Arthrography of the Shoulder," Radiology, **214**(1), pp. 267–271.

[18] Herold, T., Hente, R., Zorger, N., Finkenzeller, T., Feuerbach, S., Lenhart, M., and Paetzel, C., 2003, "Indirect MR-Arthrography of the Shoulder-Value in the Detection of SLAP-Lesions," Rofo, **175**(11), pp. 1508–1514.

[19] Dinauer, P. A., Flemming, D. J., Murphy, K. P., and Doukas, W. C., 2007, "Diagnosis of Superior Labral Lesions: Comparison of Noncontrast MRI With Indirect MR Arthrography in Unexercised Shoulders," Skeletal Radiol., **36**(3), pp. 195–202.

[20] Stone, M., 1974, "Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion)," J. R. Stat. Soc., Ser. B, **36**, pp. 111–147.

[21] Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L., 2005, "The Alzheimer's Disease Neuroimaging Initiative," Neuroimaging Clin. North Am., **15**(4), pp. 869–877.

[22] Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L., 2005, "Ways Toward an

Early Diagnosis in Alzheimer's Disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI)," Alzheimer's Dementia, **1**(1), pp. 55–66.
[23] Whitwell, J. L., Shiung, M. M., Przybelski, S. A., Weigand, S. D., Knopman, D. S., Boeve, B. F., Petersen, R. C., and Jack, C. R., 2008, "MRI Patterns of Atrophy Associated With Progression to AD in Amnestic Mild Cognitive Impairment," Neurology, **70**(7), pp. 512–520.
[24] C. R., McDonald, L. K., McEvoy, L., Gharapetian, C., Fennema-Notestine, D. J., Hagler, Jr., D., Holland, A., Koyama, J. B., Brewer, and A. M., Dale, and A. D. N. I., 2009, "Regional Rates of Neocortical Atrophy From Normal Aging to Early Alzheimer Disease," Neurology, **73**(6), pp. 457–465.
[25] Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., and Colliot, O., 2011, "Automatic Classification of Patients With Alzheimer's Disease From Structural MRI: A Comparison of Ten Methods Using the ADNI Database," Neuroimage, **56**(2), pp. 766–781.
[26] Jee, W.-H., McCauley, T. R., Katz, L. D., Matheny, J. M., Ruwe, P. A., and Daigneault, J. P., 2001, "Superior Labral Anterior Posterior (SLAP) Lesions of the Glenoid Labrum: Reliability and Accuracy of MR Arthrography for Diagnosis," Radiology, **218**(1), pp. 127–132.
[27] Waldt, S., Burkart, A., Lange, P., Imhoff, A. B., Rummeny, E. J., and Woertler, K., 2004, "Diagnostic Performance of MR Arthrography in the Assessment of Superior Labral Anteroposterior Lesions of the Shoulder," Am. J. Roentgenol., **182**(5), pp. 1271–1278.
[28] Fawcett, T., 2006, "An Introduction to ROC Analysis," Pattern Recognit. Lett., **27**(8), pp. 861–874.
[29] Metz, C. E., 1978, "Basic Principles of ROC Analysis," Semin. Nucl. Med., **8**(4), pp. 283–298.
[30] Obuchowski, N. A., 2003, "Receiver Operating Characteristic Curves and Their Use in Radiology," Radiology, **229**(1), pp. 3–8.
[31] De Coninck, T., Ngai, S. S., Tafur, M., and Chung, C. B., 2016, "Imaging the Glenoid Labrum and Labral Tears," RadioGraphics, **36**(6), pp. 1628–1647.
[32] Tanner, M. A., and Wong, W. H., 2010, "From EM to Data Augmentation: The Emergence of MCMC Bayesian Computation in the 1980s," Stat. Sci., **25**(4), pp. 506–516.
[33] Kingma, D. P., and Welling, M., 2014, "Auto-Encoding Variational Bayes," The International Conference on Learning Representations, Banff, Canada, Apr. 14–16.
[34] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., 2014, "Generative Adversarial Networks," Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, Dec. 8, pp. 2672–2680.
[35] Ha, T. M., and Bunke, H., 1997, "Off-Line, Handwritten Numeral Recognition by Perturbation Method," IEEE Trans. Pattern Anal. Mach. Intell., **19**(5), pp. 535–539.
[36] Dielman, S., 2015, "My Solution for the Galaxy Zoo Challenge," London, UK, accessed May 11, 2017, http://benanne.github.io/2014/04/05/galaxy-zoo.html
[37] Asperti, A., and Mastronardo, C., 2017, "The Effectiveness of Data Augmentation for Detection of Gastrointestinal Diseases From Endoscopical Images," Bioimaging., 11th International Joint Conference on Biomedical Engineering Systems and Technologies, Vol. 2. Jan. 19–21, Funchal, Madeira, Portugal. pp. 199–205.
[38] Glorot, X., and Bengio, Y., 2010, "Understanding the Difficulty of Training Deep Feedforward Neural Networks," 13th International Conference on Artificial Intelligence and Statistics, Vol. 9, Sardinia, Italy, May 13–15, pp. 249–256.
[39] Glorot, X., Bordes, A., and Bengio, Y., 2011, "Deep Sparse Rectifier Neural Networks," 14th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, Apr. 11–13, pp. 315–323.
[40] Kingma, D. P., and Ba, J., 2015, "Adam: A Method for Stochastic Optimization," International Conference on Learning Representations (ICLR), San Diego, CA, May 7–9.
[41] Krawczyk, B., 2016, "Learning From Imbalanced Data: Open Challenges and Future Directions," Prog. Artif. Intell., **5**(4), pp. 221–232.
[42] Buda, M., Maki, A., and Mazurowski, M. A., 2018, "A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks," Neural Networks, **106**, pp. 249–259.
[43] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., 2014, "Dropout: A Simple Way to Prevent Neural Networks From Overfitting," J. Mach. Learn. Res., **15**, pp. 1929–1958.
[44] Tikhonov, N., 1943, "On the Stability of Inverse Problems," Dokl. Akad. Nauk SSSR, **39**(5), pp. 195–198.
[45] Bishop, C. M., 2006, *Pattern Recognition and Machine Learning*, New York, pp. 256–269.