Deep Transfer Learning for Characterizing Chondrocyte Patterns in Phase Contrast X-Ray Computed Tomography Images of the Human Patellar Cartilage

Botao Deng[a,†], Anas Z. Abidin[b,*,†], Adora M. DSouza[a], Mahesh B. Nagarajan[f], Paola Coan[d,e], Axel Wismueller[a,b,c,d]

[a]University of Rochester, Department of Electrical Engineering, Rochester, U.S., 14620
[b]University of Rochester, Department of Biomedical Engineering, Rochester, U.S., 14620
[c]University of Rochester, Department of Imaging Science, Rochester, U.S., 14620
[d]Faculty of Medicine & Institute of Clinical Radiology, Ludwig Maximilians University, Munich, Germany
[e]European Synchrotron Radiation Facility, Grenoble, France
[f]Department of Radiological Sciences, University of California Los Angeles, Los Angeles, CA 90095

Abstract.

Phase contrast X-ray computed tomography (PCI-CT) has been demonstrated to be effective for visualization of the human cartilage matrix at micrometer resolution, thereby capturing osteoarthritis induced changes to chondrocyte organization. This study aims to systematically assess the efficacy of deep transfer learning methods for classifying between healthy and diseased tissue patterns. We extracted features from two different concolutional neural network architectures, CaffeNet and Inception-v3 for characterizing such patterns. These features were quantitatively evaluated in a classification task measured by the area (AUC) under the Receiver Operating Characteristic (ROC) curve as well as qualitative visualization through a dimension reduction approach t-Distributed Stochastic Neighbor Embedding (t-SNE). The best classification performance, for CaffeNet, was observed when using features from the last convolutional layer and the last fully connected layer (AUCs > 0.95), after fine-tuning. Meanwhile, off-the-shelf features from Inception-v3 produced similar classification performance (AUC > 0.95). Visualization of features from these layers further confirmed adequate characterization of chondrocyte patterns for reliably distinguishing between healthy and osteoarthritic tissue classes. Such techniques, can be potentially used for detecting the presence of osteoarthritis related changes in the human patellar cartilage.

Keywords: phase contrast imaging, patellar cartilage, deep transfer learning, convolutional neural network.

* Contact Author, anas.abidin@Rochester.edu

† Contribute equally to this paper.

1 Introduction

Phase-contrast X-ray computed tomography (PCI-CT) is an imaging technique capable of visualizing the internal architecture of tissues at a micrometer resolution.[1] This acquisition methodology exploits the fact that X-rays are not just absorbed when passing through matter but also refracted,[2,3] producing a more pronounced contrast when compared to conventional absorption based X-ray imaging modalities.[4] This allows PCI to be effective in imaging tissue types where the conventional absorption contrast is either unable to resolve the differences between different soft tissue types, i.e., breast,[5,6] or poor/absent, i.e., cartilage.[1,7,8] Although different imaging setups can be used, PCI with computed tomography, using the analyzer-based imaging (ABI) scheme[3,9,10] has been applied in different ex vivo breast,[5,6] brain,[11] and cartilage studies.[1,7] There is a huge potential for such a technique to be used for early detection of degenerative cartilage structural changes associated with osteoarthritis (OA),[12] widely recognized as one of the leading causes of disability worldwide. The advanced techniques proposed for evaluation of OA, such as delayed gadolinium-enhanced MR imaging of cartilage (dGEMRIC),[13] ^{23}Na MRI,[14] T1ρ,[15] GAG chemical exchange saturation transfer (gagCEST)[16] etc. do not possess the capability to visualize cartilage matrix structure at a cellular level. In this context the adept visualization of the cartilage matrix, using PCI-CT, can contribute towards better diagnosis and management of OA.[1] The systematic zonal achitecture maintained in the cartilage matrix as known anatomic studies[17] was clearly visualized by the imaging of samples obtained from healthy individuals. Specifically in the radial zone, the chondrocytes demonstrate an ordered arrangement (Fig. 1, top). Such a zone-specific organization progressively lost during OA progression and hence was not observed in osteoarthritic samples and was instead replaced by a more generalized clustering of cells throughout the matrix (Fig. 1, bottom).

We have previously[18,19] shown PCI-CT images can be characterized effectively with 2D or 3D texture features, in a computer aided diagnostics framework. In this study, we explore the use of deep learning for characterizing chondrocyte organization of the cartilage matrix visualized in these images. With the advent of deep-learning techniques, focus has shifted from the use of traditional "hand-crafted" features to the use of networks which learn representations best suited for a task. There has been a tremendous growth in methods to go along with vast improvements in performance of such systems in various computer vision tasks.[20] An increasing number of medical imaging studies[20] are using deep-learning methods for recognition, classification, and segmentation tasks. As these neural-networks require large amounts of annotated data, which may often not be available for medical imaging, the use of *transfer learning* has been suggested. Here, networks pre-trained on large image databases can be adapted for the specific task at hand.

One approach is to treat Convolutional Neural Network (CNN) like a feature extractor *i.e.* using "off-the-shelf" feature representations from intermediate layers of CNN and use them in a classification task. Such features have shown remarkable performance at various visual recognition tasks[21] including image classification, attribute detection and image retrieval. In medical imaging, these approaches have been applied to detection of a wide range of chest-related diseases[22] and pulmonary nodule in computed tomography scans.[23] A particular question that is not always answered in these studies is what level of representation is suited when using a convolutional neural network as a feature extractor. Although this issue has been addressed briefly[21,24] in computer-vision literature, the conclusions may not be directly applicable to medical images, as their network was both trained and tested on natural image datasets. To address such a problem with our dataset, we aim to study the characterizations obtained using different internal layers of the network for distinguishing healthy and diseased tissue classes.

The other widely used technique for transfer learning, is to *fine-tune* a pre-trained network on a target medical imaging dataset.[25] The method is used to adapt the weights of the intermediate layers of the network based on the new training data. It is based on the accepted idea that in a deep learning network, the initial layers capture features that are generic (edges, orientations, simple text patterns etc.) and as we go deeper the features tend to get more abstract and specific for the task they are being trained on. To explore this effect in detail, we fine-tune the weights of a pre-trained network and test their performance in a classification task.

A common criticism of deep-learning methods is the low interpretability of the representation(s) learned by the networks. We postulate that using the representation from different layers, and comparing them before and after fine-tuning can be an initial point in deepening our understanding of these networks. Furthermore, we investigate the descriptive power of such features using a dimension reduction technique t-distributed Stochastic Neighborhood Embedding (t-SNE), which can provide further insights into the high-dimensional representations.

This is one of the first studies exploring the application of deep transfer learning on phase contrast imaging data. We aim to investigate the effectiveness of CNNs in characterizing degenerative changes occurring due to osteoarthritis in PCI-CT images of the cartilage. In order to study this, we try to address the following, technically relevant, questions in this work:

- How do off-the-shelf features from a CNN perform when classifying PCI-CT images acquired from healthy and osteoarthritic cartilage samples.
- In the commonly used architecture, CaffeNet (based on AlexNet); what is the adequate representation that can be extracted from the internal layers?
- Can fine-tuning with a small dataset help in improving the performance of tranfer learning methods?
- Futhermore, we explore the application of Inception-v3, an advanced CNN architecture which is not commonly used in medical imaging compare its performance to CaffeNet.
- Can visual exploratory analyses shed further light on the descriptors obtained from the CNN?

2 Data

We specifically focus on data acquired from cartilage in the retropatellar joint which is understood to possess significant potential for enabling early detection of treatable osteoarthritic changes. Exvivo imaging of the patellar cartilage, using PCI-CT, revealed specific differences in their internal organization chondrocytes in osteoarthritic samples. We have previously proposed quantitative metrics, based on texture and topology, that could characterize such differences and evaluate their ability to serve as diagnostic biomarkers for osteoarthritic-induced changes to the cartilage matrix.[18] In this study we aim to achieve such characterization with a deep convolutional neural network.

2.1 Samples

The selection of the patellae was based on age of the donor, macroscopic visual inspection, and probing of the cartilage surface at autopsy. Donors older than 40 years were excluded for harvest of normal samples; no constraint in age was imposed on potential donors for osteoarthritic samples. A smooth, white, and shiny surface present across the entire patellar cartilaginous surface and prompt resilience to manually performed focal indentation probing were criteria that defined macroscopically normal cartilage. Lack of these criteria in addition to visually perceived defects in the joint surface were used to select osteoarthritic samples. IRB was waived by the institutional review board of the Ludwig Maximilians University, Munich, Germany. Based on these inclusion criteria, two healthy and three osteoarthritic cylinder-shaped osteochondral samples (diameter: $7mm$) were extracted within $48h$ postmortem from the lateral facet of the four human patellae using a shell auger. Cylinders were trimmed to a total height of $12mm$ including the complete cartilage tissue and the subchondral bone. The samples were continuously rinsed by 0.9% saline during extraction, trimming, and removal of soiling from sawing. During image acquisition, samples were dipped into a 10% formalin solution. The grade of the osteoarthritic samples was assessed to be 3 (mild OA) for one subject and 4 (advanced OA) for the others, based on a histological standard.[26]

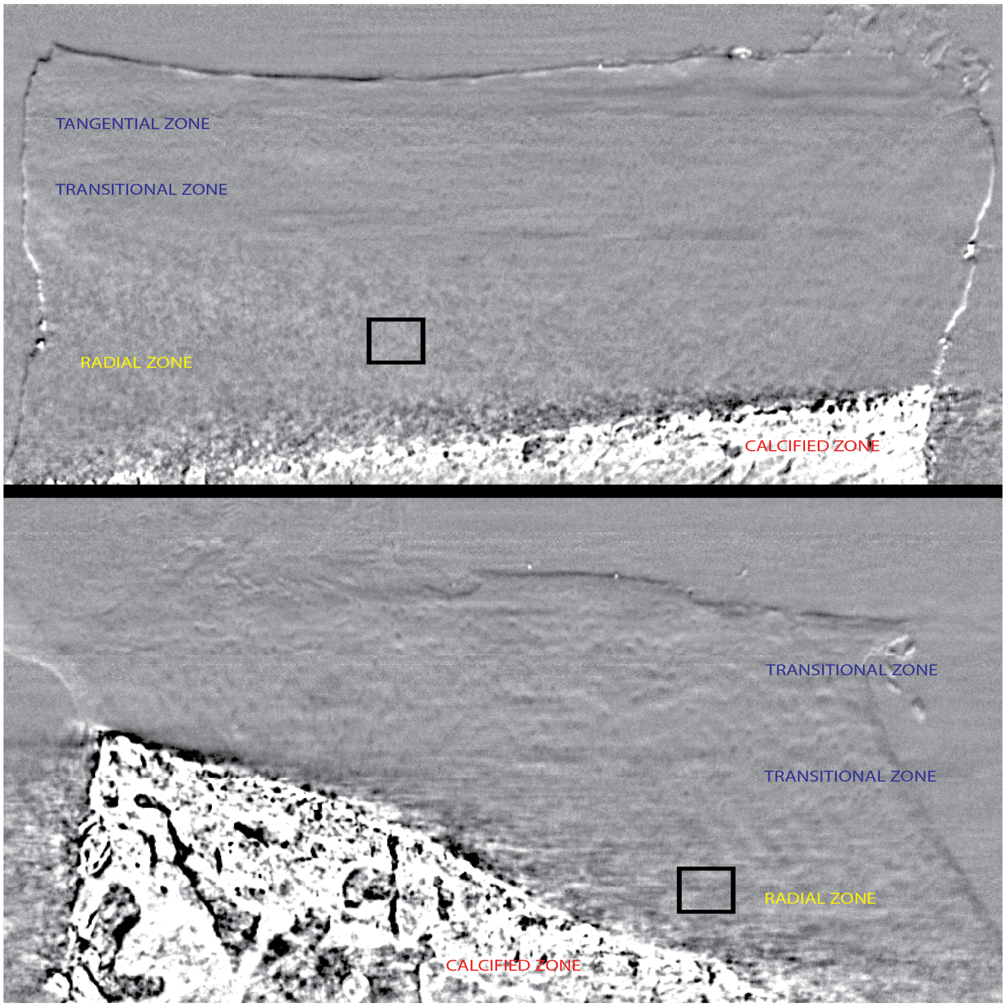

Fig 1: **Cross-sectional view of single slice of articular cartilage as obtained using PCI-CT** The presence of an orderly zonal architecture (tangential, transitional and radial) can be clearly visualized via PCI-CT imaging, particularly in healthy samples **(Top)**. This zone-specific organization of chondrocytes gradually degrades during osteoarthritis and is instead replaced by a more generalized clustering of cells throughout the matrix accompanied by a loss of clear zonal separation. **(Bottom)**. The black boxes indicate an example ROI definition in the radial zone extracted from both the groups.

2.2 Imaging Setup and Reconstruction

Details of the imaging setup have been described in detail in previous studies.[1] We briefly summarize the process here. The ABI setup consisted of a parallel monochromatic X-ray beam, used to irradiate the sample, and of a perfect crystal, the analyzer, placed between the sample, and the detector.[27] The analyzer acts as an angular filter of the radiation transmitted through the object and before being detected, the beam is modulated by the angle-dependent reflectivity of the crystal; its rocking curve (RC) has a full width at half maximum typically of the order of a few microradians. All images were acquired at the half maximum position on one slope of the RC (50% position), which was chosen to achieve the best sensitivity.

Experiments were performed at the Biomedical Beamline (ID17) of the European Synchrotron

Radiation Facility (ESRF, France). A highly collimated X-ray beam was produced by a 21-pole wiggler after monochromatization by means of a double Si (111) crystal system and an additional single Si (333) crystal.[28] The emerging, refracted, and scattered radiation from the sample was analyzed with a Si (333) analyzer crystal. Quasimonochromatic X-rays of 26 keV were used. Given the laminar shape of the stationary synchrotron beam, images were obtained by rotating and vertically scanning the sample through the X-ray beam.[29] The imaging detector used was the Fast Readout Low Noise (FReLoN) CCD camera developed at the ESRF.[30] The X-rays were converted to visible light by a 60 μm thick Gadox fluorescent screen; this scintillation light was then guided onto a 2048×2048 pixel $14 \times 14 \mu m^2$ CCD (Atmel Corp, U.S.) by a lens-based system. The effective pixel size at the object plane was $8 \times 8 \mu m^2$.

To acquire the CT images with synchrotron radiation, the sample was rotated about an axis perpendicular to the incident laminar beam; it was vertically displaced at the end of each rotation to image a different region. Both the beam and the detector were kept stationary; and an angular projection datast is acquired over $360°$, in increments of $1°$ is acquired. A flat field normalization was performed for each angular projection image to reduce the effects of the spatial and temporal X-ray beam inhomogeneities. Tomographic images were reconstructed using a direct Hamming filter backprojection algorithm.[31] An image volume of dimensions $1120 \times 1124 \times 805$ was acquired for each specimen and subsequently trimmed to eliminate background regions.

2.3 Annotation

A total of 842 square ROIs (439 were osteoarthritic and 403 were healthy), capturing chondrocyte patterns in the radial zone of the cartilage matrix, were annotated on the acquired PCI-CT images of all five specimens. In each specimen, special care was taken to ensure that the same cluster of chondrocytes was not captured by different ROIs by ensuring that annotated slices were at least $32 \mu m$ apart. We ensured that no healthy ROIs were extracted from the OA samples in this study. The annotations were made using a square of 101×101 pixels. The ground truth was extracted using analysis performed by two independent observers as described in.[1]

3 METHODS

In studies such as ours, where a large dataset is not available for training deep learning network from initialization, researachers have often used what is known as "transfer-learning" from pre-trained networks. Commonly, such networks are made publicly available after training on a large image database such as ImageNet.[32] The efficacy of repurposing activations extracted from a pre-trained deep convolutional neural network to medical image classification tasks has been illustrated in.[21,33] For characterizing the patterns in the cartilage matrix, in this study we have used and compared the performance of two different pre-trained convolutional neural networks (CNNs): CaffeNet[34] and Inception-v3.[35]

3.1 Features from CaffeNet

3.1.1 Off-the-shelf

CaffeNet, an adaptation of the winning architecture AlexNet of ILSVRC 2012,[32] is one of the most widely used networks used in transfer learning. Its simple architecture, consisting of five convolutional layers (*conv*) and 3 fully connected layers (*fc*), facilitates easier benchmarking of classification performance for different CNNs. The trained network, trained on about 1 million images of the ImageNet database, has been made available by Jia et al.[34] The last fully connected layer corresponds to the 1000 outputs required for classiying the ImageNet data. Each ROI from our study was scaled with bilinear interpolation to match the size of input layer (227 × 227) and a training set mean was subtracted in line with standard practice in deep-learning studies.[34]

It has been proposed that features in the intermediate layers capture adequate representations for transfer-learning studies[21] as these layers are neither too specific to the dataset the network was originally trained on, nor too general to not contain any representative information from images. To test whether such a rule would apply to our PCI-CT images, we extracted features from all layers of the CaffeNet. Henceforth, features from the 5 convolutional and 2 fully connected layers of the CaffeNet (excluding the normalization layer or pooling layers), are denoted as $conv_n$ $(1 \leq n \leq 5)$ or fc_n $(6 \leq n \leq 7)$, respectively.

Features from the convolutional layers are multidimensional and have a relatively large size (e.g. $conv_2 : 27 \times 27 \times 256$, $conv_4 : 13 \times 13 \times 348$), thereby increasing the computational burden. We, therefore, applied Global Average Pooling (GAP) to pool the feature maps for converting them to linear feature vectors (e.g. $conv_1$:$55 \times 55 \times 96$ into a vector $1 \times 1 \times 96$). Although other strategies to pool features from the convolutional layers can be considered, GAP has been widely applied in the development of deeper architectures and is understood to enforce correspondance between feature maps and categories.[36]

3.1.2 Fine-tuning

To study the effect of fine-tuning, we replaced the final fully connected layer (fc_8) in the original CaffeNet with a 2 ouput layer corresponding to the healthy and osteoarthritic classification. The weights were randomly initialized based on a Gaussian distribution. During fine-tuning, we ensured that the errors are back-propagated to all layers of the network. The learning rate for the last layer is fixed at 0.001 with 0.9 momentum as this layer had to be learnt from scratch. However the learning rates for the remaining layers were set at 0.0005 as only fine adjustments from a pre-trained state would be needed. The network was trained for 1500 iterations with a batch size of 256. The multinomial logistic loss function reached a low plateau after about 1000 training iteration. Features are referred to as $fconv_n$ or ffc_n, for a fine-tuned convolutional layer, and a fine-tuned fully connected layer, respectively. We report results obtained using the two output layer of the CNN as well as those obtained when using features from the intermediate layers with a support vector machine classifier.

3.2 Features from Inception-v3

We also tested the performance of an Inception-v3 network, a contemporary architecture for characterizing our images. This network is a scaled-up version of GoogLeNet, and augments various heuristic improvisations over CNN architectures such as AlexNet, OverFeat or Decaf previously used in medical imaging studies. The pretrained Inception-v3 model was obtained from the open source deep learning framework, Lasagne[1]. Table 1 shows the architecture of Inception-v3 model, the first few layers are similar to traditional CNN, including 5 convolutional layers and 2 max pooling layers. These are followed by 11 inception modules and a Global pooling layer. A softmax layer is used to generalize a score for each category.

[1] Lasagne model zoo.

Table 1: The architecture of Inception-v3 Net. InceptionD is marked as a modified version of inceptionB module. The three dimensions of “input size” are channel dimension, width, height. In this paper, the terms filter/channel/kernel are used interchangeably.

Type of Layer	patch size / stride	input size
conv	$3 \times 3/2$	$3 \times 299 \times 299$
conv_1	$3 \times 3/1$	$32 \times 149 \times 149$
conv_2	$3 \times 3/1$	$32 \times 147 \times 147$
pool	$3 \times 3/2$	$64 \times 147 \times 147$
conv_3	$1 \times 1/1$	$64 \times 73 \times 73$
conv_4	$3 \times 3/1$	$80 \times 73 \times 73$
pool_1	$3 \times 3/2$	$192 \times 71 \times 71$
3× inceptionA	Figure 5 of[35]	$192 \times 35 \times 35$
inceptionB	Figure 10 of[35]	$288 \times 35 \times 35$
4× inceptionC	Figure 6 of[35]	$768 \times 17 \times 17$
inceptionD	Figure 10* of[35]	$768 \times 17 \times 17$
2×inceptionE	Figure 7 of[35]	$1280 \times 8 \times 8$
Global pool	—	$2048 \times 8 \times 8$
fc (Softmax)	—	$2048 \times 1 \times 1$

The core idea of the inception module is to use filters of multiple sizes with different receptive field, this allows for capturing local variablity and hence improved characterization. The original naive inception module in GoogLeNet performed convolution operations of three sizes, $1 \times 1, 3 \times 3, 5 \times 5$, and then combined them together to form a single feature matrix which is fed into the subsequent layers. Such operations when performed for larger networks quickly increase the amount of computations which can be prohibitively expensive. Improvements in the implementation of convolutions as suggested by,[36] significantly circumvents this problem. These modules also serve to enhance the representative power of networks by reducing the correlations in the activations of nearby neurons. Further details of the Inception module are discussed elsewhere.[37] Such improvements have allowed for significant enhancements in the representative power of neural networks as evidence by the performance of Inception networks in ImageNet competition.

For each image, the training set mean image was subtracted, and input image is resized to match the input layer dimension of Inception-v3. We extracted features from the last 10 inception modules, we referred to those extracted features as *inception*$_n (1 \leq n \leq 10)$.

3.3 Features from Gray-level co-occurrence matrices

RECHECK

To analyze the performance of traditional, so-called “hand-crafted” features, measures were extracted from *Gray-level co-occurrence matrices* (GLCM) constructed using the ROIs as described in.[38] On each ROI, the gray-levels were quantized to 32 gray-level values. GLCMs were then generated in the four principal directions and then summed up element-wise resulting in one non-directional GLCM. From this, the least correlated and most frequently used statistical features were computed, i.e., absolute value, entropy, contrast, energy, correlation and homogeneity.[39]

3.4 Classification

Subsequent to feature extraction we performed a supervised learning step where the ROIs were classified as healthy or osteoarthritic. Briefly, a SVM finds a linear separating hyperplane with the maximal margin in the feature space. The SVM implementation was taken from the Scikit-learn library.[40] As the performance obtained with the features extracted was optimal using the linear kernel, we have not explored additional higher dimensional kernels in this study. This was reduced the number of parameters to be optimized during cross-validation.

Due to the limitation imposed by the small dataset used in this study, we imposed the following restrictions during our analyses. This was done in the following ways: (1) We have defined *non-overlapping* ROIs in the radial zone of the cartilage matrix in the PCI-CT images for each subject. This avoided over-representation of specific patterns from each subject. (2) For each iteration of the machine-learning step we have separated out 1 healthy and 1 diseased subject randomly for testing. In contrast to random splitting of the dataset into a training/test sets this ensures that ROIs from the same subject are never used towards training as well as testing, thereby preventing overfitting of the classifier to patterns from a single subject and biasing the performance evaluation.

Performance of the classifier over the different iterations is evaluated using Receiver operating characteristics (ROC) curve. Here the SVM is used to produce probability scores for each class. Probabilities for a binary classification are calibrated using Platt scaling: logistic regression on the SVMs scores, fit by an additional 5-fold cross-validation on the training data.

These scores are then thresholded systematically to obtain a measure of True-positive rate (sensitivity) and False positive rate (1 - specificity), resulting in a ROC curve. The area under the ROC curve (AUC) is used as the evaluation metric for a particular feature set.

3.5 Visualization of CNN features

3.5.1 Activation of layers/neurons

An alternative to study the features obtained from CNNs is through appropriate visualization techniques. One such approach, is ti visualize the activation of neurons during a forward pass through the network. In CNNs, individual neurons act as filters applied in a 2D convolution over the two spatial dimensions of the image, which in turn produces activations (or input) for subsequent layers. Such tools have been used previously to identify neuron which respond to specific patterns in the images such as text, flowers, faces etc.[41] Qualitative visualization of the features can help gain insight into CNNs especially with regards to distinguishing patterns of osteoarthritis in the cartilage matrix.

3.5.2 Dimension Reduction

We also explored the use of an unsupervised dimension reduction technique known as t-Distributed Stochastic Neighbor Embedding (t-SNE),[42] as an alternative mechanism for comparing the features representation power for distinguishing between healthy and osteoarthritic ROIs.

Stochastic Neighbor Embedding (SNE) converts Euclidean distances between high-dimension texture feature vectors into conditional probabilities representing similarities; the closer the feature vectors, the higher the similarity.[42] Once conditional probability distributions are established for both the high-dimension feature vectors and their corresponding low-dimension representations, the goal of the algorithm is to minimize the mismatch between the two. t-SNE was developed as an improvement over SNE to further simplify cost function optimization and overcome the so-called crowding problem inherent to SNE.[42] Details pertaining to this algorithm and its cost function minimization can be found in,[42] and a review of the algorithm can be found in.[43,44] This technique has been shown to be particularly applicable for visualization of the high-dimensional data.

4 Results

In this section, we evaluate and compare the classification performance of off-the-shelf features from multiple layers of two different networks and also commonly used GLCM based features on the PCI-CT dataset.

4.1 Features from CaffeNet

4.1.1 Off-the-shelf

For features extracted from a pre-trained CaffeNet, the best classification performance was obtained using the features extracted from the last convolutional layer (*conv*$_5$, AUC=0.91) and the fully connected layer (*fc*$_7$, AUC=0.90). Interestingly, the commonly used first fully connected layer in such studies does not perform as well (AUC=0.81 ± 0.17). The initial convolutional layers perform poorly at the classification task (Fig. 2).

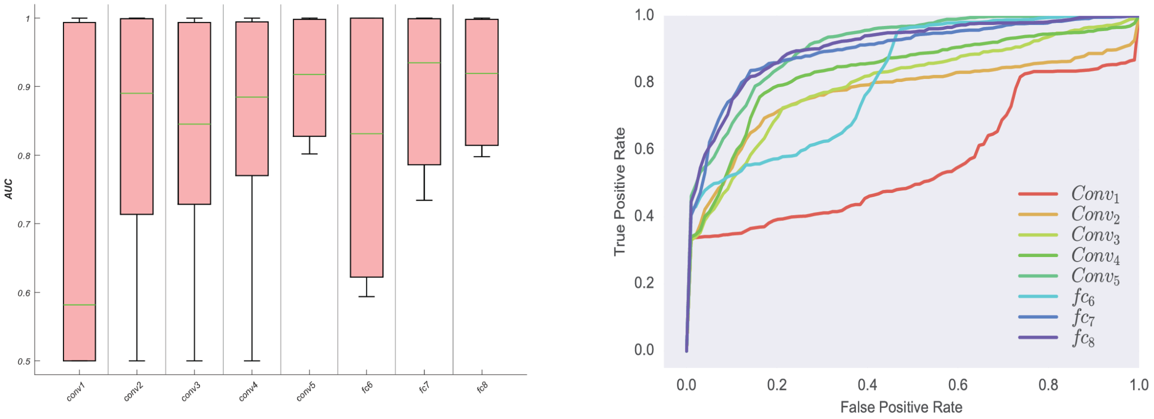

Fig 2: **Comparison of classification performance for features extracted from different layers of a pre-trained CaffeNet. (Left)** Boxplots representing the performance of features from different layers over multiple iterations. Each of the colored region indicates the 25^{th} and the 75^{th} percentile and the central line corresponds to the median AUC value across different test/train splits. The circles, when seen represent statistical outliers. Some AUC values for the first few layers were < 0.5. **(Right)** The corresponding mean ROC curves generated with features from different layers. Representations extracted the last convolutional as well as the last fully connected layer perform best at the classification task.

4.1.2 *Fine-tuning*

The feature extracted from layers of a fine-tuned CaffeNet exhibited an improvement in performance in most layers. Specifically, features from *fconv*$_3$ and *ffc*$_6$, which performed poorly prior to fine-tuning showed a significant boost in performance. Features from *fconv*$_5$ and *ffc*$_7$ show improvement over their previous classification results. Table 2 and Fig. 3 list all the results obtained using these methods.

Furthermore, when classifying using the fine-tuned CNN we obtained AUCs of 0.96 ± 0.07.

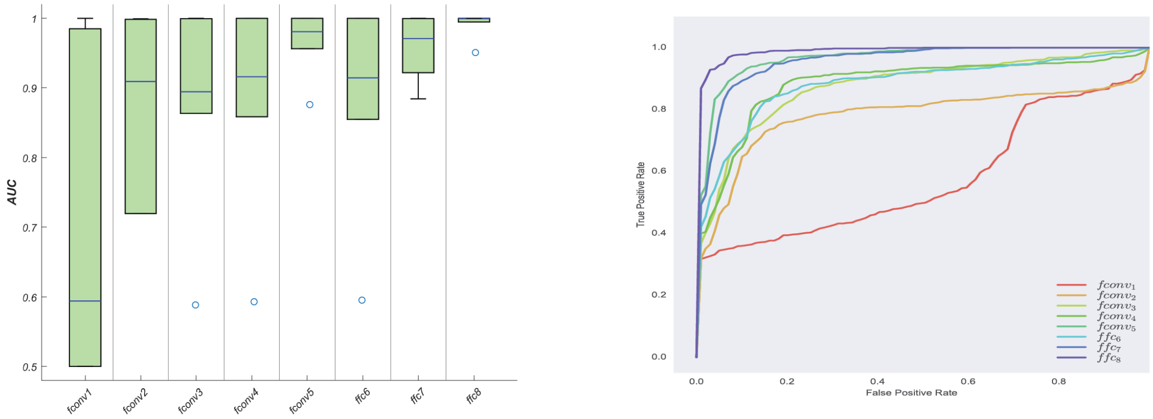

Fig 3: **Comparison of classification performance for features extracted from different layers of a fine-tuned CaffeNet. (Left)** Boxplots representing the performance of features from different layers over multiple iterations. Each of the colored region indicates the 25^{th} and the 75^{th} percentile and the central line corresponds to the median AUC value across different test/train splits. Some AUC values for the first layer were < 0.5. The circles, when seen represent statistical outliers. **(Right)** The corresponding mean ROC curves generated with features from different layers. There is a noticeable improvement in performance of features from all layers post fine-tuning, particularly in the later layers of the network.

4.2 *Features from Inception-v3*

The performance of features from the Inception-v3 network are shown in Fig 4. In general features from all inception modules can accurately distinguish between the two classes (AUC > 0.95), with no significant differences in performance seen through the different layers.

4.3 *Features from Gray-level co-occurrence matrices*

Standard texture features extracted with the ROIs were also evaluated within the same cross validation scheme. Most GLCM derived texture features perform poorly (Table 2) with the exception of *Correlation* which produced a high AUC of (0.93 ± 0.07).

4.4 *Visualization of CNN features*

4.4.1 *Activation of layers/neurons*

To perform qualitative visualization of the features produced by the datast we used the DeepVis toolbox[41] to study activations produced by the network. We noticed differential activation in layers

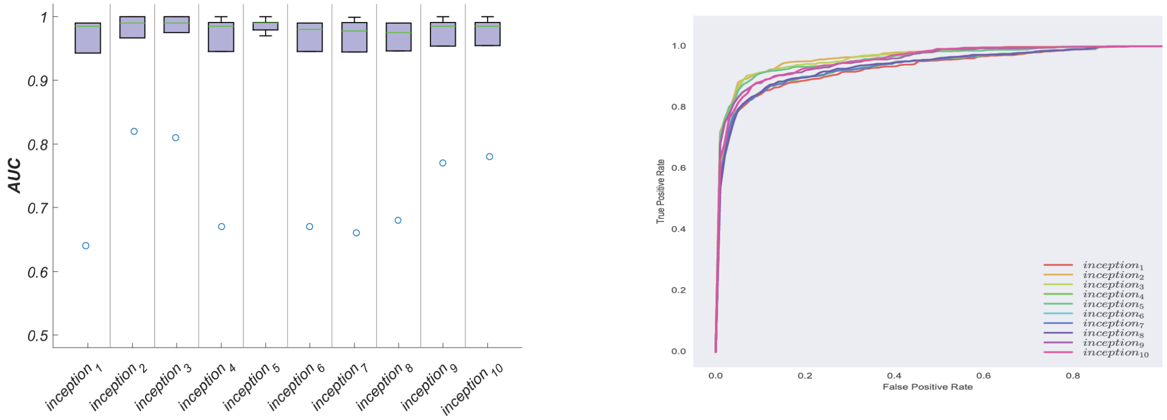

Fig 4: **Comparison of classification performance for features extracted from different inception modules of a pre-trained Inception-v3 network. (Left)** Boxplots representing the performance of features from different layers over multiple iterations. Each of the colored region indicates the 25^{th} and the 75^{th} percentile and the central line corresponds to the median AUC value across different test/train splits. The circles, where seen represent statistical outliers. **(Right)** The corresponding mean ROC curves generated with features from different modules. Features from all modules perform well at the classification task.

Table 2: Comparison of AUC (mean ± standard deviation) values obtained for the different feature sets used in this study

CaffeNet	*conv*$_1$	*conv*$_2$	*conv*$_3$	*conv*$_4$	*conv*$_5$	*fc6*$_6$	*fc*$_7$	*fc*$_8$		
AUC	0.56 ± 0.36	0.77 ± 0.31	0.80 ± 0.22	0.83 ± 0.19	0.91 ± 0.08	0.81 ± 0.17	0.90 ± 0.10	0.91 ± 0.08		
FT-CaffeNet	*fconv*$_1$	*fconv*$_2$	*fconv*$_3$	*fconv*$_4$	*fconv*$_5$	*ffc*$_6$	*ffc*$_7$	*ffc*$_8$		CNN o/p
AUC	0.57 ± 0.35	0.78 ± 0.31	0.88 ± 0.14	0.88 ± 0.14	0.96 ± 0.04	0.88 ± 0.14	0.96 ± 0.04	0.99 ± 0.02		0.96 ± 0.07

Inception-v3	*inception*$_1$	*inception*$_2$	*inception*$_3$	*inception*$_4$	*inception*$_5$	*inception*$_6$	*inception*$_7$	*inception*$_8$	*inception*$_9$	*inception*$_{10}$
AUC	0.93 ± 0.13	0.97 ± 0.06	0.96 ± 0.07	0.94 ± 0.12	0.96 ± 0.07	0.94 ± 0.12	0.94 ± 0.12	0.94 ± 0.11	0.96 ± 0.08	0.95 ± 0.07

GLCM	*Absolute Value*	*Entropy*	*Contrast*	*Energy*	*Correlation*	*Homogeneity*
AUC	0.78 ± 0.12	0.57 ± 0.04	0.78 ± 0.13	0.59 ± 0.06	0.93 ± 0.07	0.76 ± 0.12

that produced high classification performance. For example, for *conv5* layer we noticed specific neurons that mostly produced high activation for diseased samples (Fig. 5, red box). Similarly some neurons responded preferentially to healthy samples (Fig. 5, green box). Interestingly, layer *fc6* qualitatively produced similar activations for ROIs from both groups, however *fc7* did not [[[[[[[[[[[[[[Supplementary Fig]]]]]]]]]]]]]. Although quantitative analysis for these neurons was not performed, such visualization can aid in the choice of specific layers for transfer learning. CHK CHK

4.4.2 Dimension Reduction

We have also explored the visualization of features from both networks using t-SNE (Fig.6). The visualizations produced distinct clustering of healthy and diseased ROIs, in-line with classification performance as obtained in the previous sections. We show here visualizations of features exhibiting best performance in CaffeNet (with and without fine tuning), namely $conv_5$ and fc_8. Given

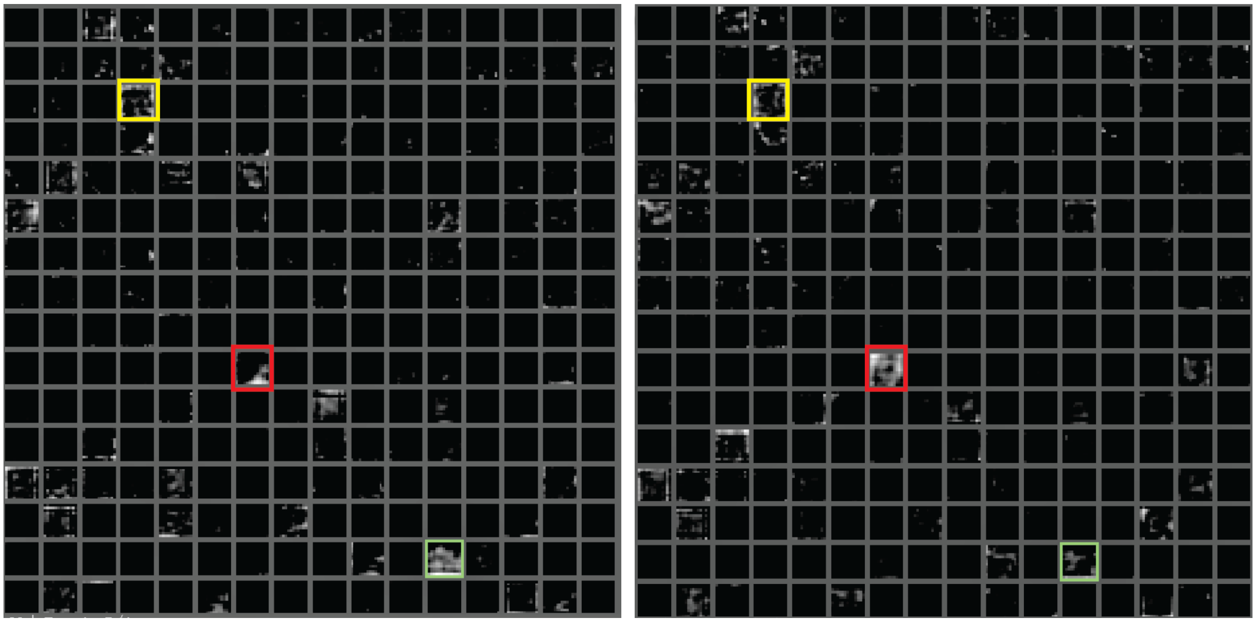

Fig 5: **Differential activation of neurons to healthy (left) and osteoarthritic (right) samples in all 256 neurons in the *conv5* layer of CaffeNet visualized using the DeepVis toolbox.** It was noticed that neuron 150 (red box) generally produced high activation in response to OA samples. In contrast neuron 221 (green box) produced an opposite response. Many neurons produced similar responses in this layer, for ex. neuron 36. Exploratory visualization techniques can motivate the choice of specific layers for extraction of CNN features. These visualization have been adapted using the DeepVis toolbox[41]

that features from Inception-v3 performs equally well, we included the last two inception modules for a comparative analysis.

5 Discussion

The application of Phase Contrast Imaging with Computed Tomography (PCI-CT) for visualizing structural details of the cartilage matrix with high spatial resolution has been demonstrated previously.[1,18] This technique enables the capture of differences in chondrocyte organization between healthy and osteoarthritic cartilage samples. In this study, we explored the application of deep transfer learning for characterizing the organizational patterns using multiple layers of convolutional neural networks, in a computer-aided diagnostics framework. Our results illustrate that features extracted from networks pre-trained on non-medical imaging datasets can accurately classify chondrocyte patterns as normal or osteoarthritic with high accuracy.

The current popularity of deep-learning techniques stems from the ability of such networks to accurately capture patterns within images and hence produce high performance at various different tasks. This is further emphasized through the application of transfer learning methods, wherein networks trained in different setting can be adapted for a new task. These methods have recently gained popularity in medical imaging applications, where, networks trained on the ImageNet dataset have been used for the classification and detection of diseases,[22,23] often with improved level of performance over traditional, "hand-crafted", features. We investigated the applicability

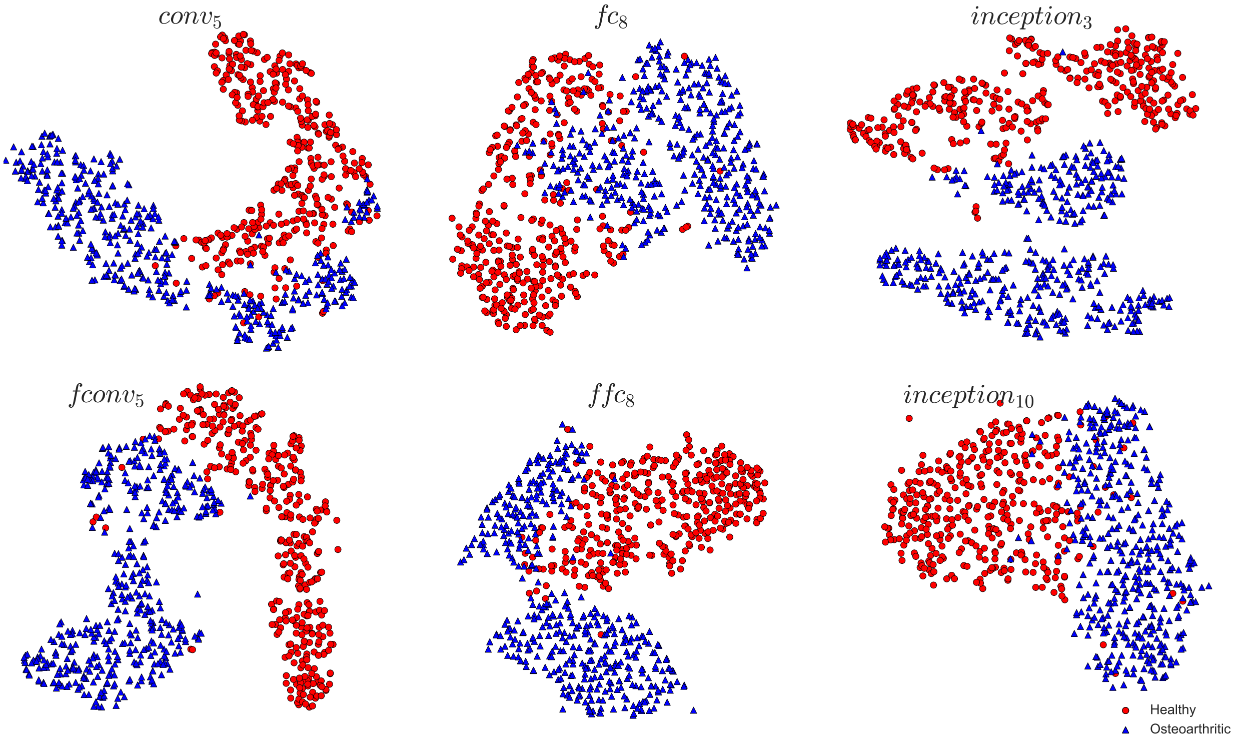

Fig 6: **Visualization of high-dimensional features obtained using t-SNE dimension reduction.** Here visualizations of the features that performed best at the classification task are compared. We notice a clear distinction between healthy (red) and diseased (blue) clusters re-iterating that such features capture adequate information for distinguishing between the two classes.

of such methods on a dataset comprising of images of healthy and osteoarthritic cartilage acquired using PCI-CT.

It is widely accepted that in a deep learning network, the initial layers learned features that are generic (edges, wavelet filters) and as we go deeper they get more specific for the task they are being trained on. Yosinski et al.[24] explored this aspect of transfer learning and noted that not all features are easily transferable, especially if, the datasets for two tasks are significantly different. Based on this it would be expected that features from the earlier layers of the CNN would be transferred more easily. However, our results suggest, that the choice of layer from which specific features are extracted affects the overall results significantly. It is noteworthy that features extracted from the first fully connected layer (fc_6) of CaffeNet, which has been widely used in transfer learning studies does not perform as well as other layers. In fact, pooled features from the last convolutional layer ($conv_5$) and the last fully connected layer (fc_7) perform the best indicating adequate characterization of chondrocyte patterns in these (Fig. 2). This firmly suggests that the choice of layer from CaffeNet for transfer learning should be made in a principled manner based on the target application, or in the case of using a smaller network such as CaffeNet should reported for all layers as previously suggested by.[20] When the CaffeNet was fine-tuned with our dataset, an improvement in performance was noted for all layers of the network, although the best performing feature were still obtained from $fconv_5$ and fc_7 (Fig. 3). This is suggestive of the fact that significant improvement in the performance of convolutional neural networks can be obtained with a small amount of fine-tuning when the parameters (learning rate, momentum, initialization scheme etc.) are properly chosen. Additionally, obtaining labels from the fine-tuned CNN using a 2 output layer also produces a high AUC (0.96 ± 0.07) similar to fine-tuned features when used with SVM. Post-fine tuning very high AUC values are obtained, indicating that some of the features learnt accurately capture distinguishing characteristics in healthy and osteoarthritic ROIs.

Interestingly, the features from the inception modules of the Inception-v3 network, all exhibit accurate classification performance (Fig. 4). Layers of this network significantly, outperformed the CaffeNet suggesting that a more state-of-the art network provides an enhanced characterization and hence may not require fine-tuning of the network parameters for classification tasks. As performance obtained was fairly high (AUCs > 0.94) for most layers, we have not performed fine-tuning for this network. Although, variants of the AlexNet are still popular and continue to be widely used, it can be advocated that modern architectures such as Inception Networks or ResNets[45] could enhance performance of transfer learning methods.

Apart from quantitative analysis of the characterization power of deep learning features through classification, we also suggest two different visualization techniques for CNNs in a medical imaging setting. It has been shown previously[41] that when working with non-medical datasets, specific neurons within layers elicit preferential response to high-level features such as text, flower or faces. As the ROIs used in this study were annotated specifically in the radial zone of the matrix, such high-level features are not present. Using the DeepVis toolbox, we were however, able to identify neurons which tend to pick up clustered/checkered patterns seems to respond more to ROIs extracted from normal subjects compared to those from OA 5. We anticipate the use of such a visualization technique could potentially contribute further in more complicated tasks such as organ segmentation, lesion detection etc as larger medical imaging datasets become available. The visualization of activations in such a manner given an intuitive understanding of the patterns picked up by CNNs and tends to improve the interpretability features. It can provide valuable insights for transfer learning studies to better apply and adapt networks trained on non-medical datasets and also be used to motivate exact the choice of features in a medical imaging task.

We also visualized the features using dimension reduction technique, t-SNE, which has shown to be effective in visualizations of high-dimensional data. For features from off-the-shelf CaffeNet we notice the a slight overlap between the healthy and diseased subject groups, which is interestingly reduced post fine-tuning (Fig. 6). Similar to our quantitative analysis, features from the subject groups, Inception-v3 Net, cluster almost perfectly with only a little overlap seen. We would however like to point out that classification performance using SVMs and cluster separation in using t-SNE need not always produce perfect agreement. It should also be noted that t-SNE is an unsupervised dimension reduction approach that can provides provide valuable insight regarding the separability of the the data. From our visualizations, we see that reducing the dimension of features using t-SNE produces a good separation of the two groups, in features from some layer [[[[[SUPPP FIGURE]]]]]. This is indicative of the *possibility* of achieving good classification results, which agrees with our observation that SVM classifies the two groups well as shown by the high AUC values. Since, SVM is a supervised learning algorithm that uses training examples as support vectors for defining the decision boundary. As the two algorithms perform different task using different approaches we cannot expect its classification results to have a one to one correspondence with the t-SNE clusters, in all scenarios. This is seen in the t-SNE visualizations obtained for features from $inception_4$, $inception_6$, & $inception_7$, wherein the clusters are not separable however using the features with SVM still produces a high classification accuracy.

Overall, our results indicate that the features from CNN in fact can accurately capture differences between healthy and osteoarthritic patterns in the patellar cartilage. WHY SO HIGH??

These features also outperform traditional hand-crafted features (such as GLCM) as seen in our results 2. Researchers have suggested the use of transfer learning to serve as potential benchmarks for evaluating newer methods.[21] The high accuracies in classification obtain suggest the potential of such methods for developing novel imaging based biomarkers for osteoarthritis.

Although our results are promising, we would like to acknowledge a few limitations of this study, in the current form. Firstly, the specimens used for imaging were obtained from a small number (five) of patients. To avoid overfitting due to the availability of a limited number of samples, we have endeavored to ensure strict separation of training and testing data, during the machine learning step. It, however, remains a possibility that the ROIs extracted from the two classes (healthy and osteoarthritic) could be over-represented due to the limited variations of patterns found in these subjects. In future studies as the availability of PCI-CT systems increases, we aim to include more patients to perform of a more robust evaluation of the methods proposed here. This dataset however allows the exploration and study of chondrocyte patterns at an unprecedented resolution for imaging studies. It has also enabled us to perform an exhaustive investigation into the feature sets obtained from different layers of CNNs and study their applicability in a computer-aided diagnostics task. Secondly, we also note a practical limitation with the experimental setup used in this study concerning the reliance of the imaging technique on synchrotron radiation. This use of a stationary radiation source restricts PCI-CT imaging to *ex vivo* specimens rather than cartilage tissue *in vivo*. In this regard, current studies being undertaken, are studying the implementation of PCI methods with high brilliance and high-energy compact Xray sources that show significant promise for transferring PCI-CT imaging to a clinical environment.

6 CONCLUSION

This study shows the applicability of deep transfer learning techniques to classify healthy and osteoarthritic chondrcyte patterns acquired from PCI-CT imaging of the human patellar cartilage. We explored the utility of feature representations extracted from two different convolutional networks: a simpler and widely used network, CaffeNet, as well as a network with more advanced architecture, Inception-v3. Our results show that, features extracted last convolutional layer and last fully connected layer of CaffeNet perform significantly better than other layers, suggesting that an informed choice regarding layer for feature extraction is critical for the achieving good performance. Although, we have used a smaller dataset for this study, we have shown that fine-tuning, when applied appropriately, can aid in improving the performance of such networks. Features extracted from the modules of the Inception-v3 network produce excellent classification performance even without fine-tuning. Thus, there is a potential for using such deep-transfer learning approaches, for detecting the presence osteoarthritis, in a computer-aided diagnosis framework. However, larger studies need to be conducted in order to further validate the clinical plausibility of such methods.

Disclosures

The authors do not have any financial interests to disclose.

Acknowledgments

This research was funded in part by the National Institute of Health (NIH) Award R01-DA-034977, the Harry W. Fischer Award of the University of Rochester, the Clinical and Translational Science Award 5-28527 within the Upstate New York Translational Research Network (UNYTRN) of the Clinical and Translational Science Institute (CTSI), University of Rochester, the Center for Emerging and Innovative Sciences (CEIS), a NYSTAR-designated Center for Advanced Technology. This work was performed as a practice quality improvement (PQI) project for maintenance of certificate (MOC) of Axel Wismllers American Board of Radiology (ABR) certification. The authors would like to thank the ESRF for providing the experimental facilities and the ESRF ID17 team for assistance in operating the facilities.

References

1 P. Coan, F. Bamberg, P. C. Diemoz, *et al.*, "Characterization of osteoarthritic and normal human patella cartilage by computed tomography x-ray phase-contrast imaging: a feasibility study," *Investigative radiology* **45**(7), 437–444 (2010).

2 A. Snigirev, I. Snigireva, V. Kohn, *et al.*, "On the possibilities of x-ray phase contrast microimaging by coherent high-energy synchrotron radiation," *Review of scientific instruments* **66**(12), 5486–5492 (1995).

3 T. Davis, D. Gao, T. Gureyev, *et al.*, "Phase-contrast imaging of weakly absorbing materials using hard x-rays," *Nature* **373**(6515), 595–598 (1995).

4 T. Takeda, A. Momose, Y. Itai, *et al.*, "Phase-contrast imaging with synchrotron x-rays for detecting cancer lesions," *Academic radiology* **2**(9), 799–803 (1995).

5 J. Keyrilainen, M. Fernández, M.-L. Karjalainen-Lindsberg, *et al.*, "Toward high-contrast breast ct at low radiation dose 1," *Radiology* **249**(1), 321–327 (2008).

6 T. Schneider, P. Coan, D. Habs, *et al.*, "[application of brilliant x-rays in mammography. development and perspectives of phase contrast techniques].," *Der Radiologe* **48**(4), 345–350 (2008).

7 P. Coan, J. Mollenhauer, A. Wagner, *et al.*, "Analyzer-based imaging technique in tomography of cartilage and metal implants: a study at the esrf," *European journal of radiology* **68**(3), S41–S48 (2008).

8 C. Muehleman, S. Majumdar, A. S. Issever, *et al.*, "X-ray detection of structural orientation in human articular cartilage," *Osteoarthritis and cartilage* **12**(2), 97–105 (2004).

9 D. Chapman, W. Thomlinson, R. Johnston, *et al.*, "Diffraction enhanced x-ray imaging," *Physics in medicine and biology* **42**(11), 2015 (1997).

10 A. Bravin, "Exploiting the x-ray refraction contrast with an analyser: the state of the art," *Journal of Physics D: Applied Physics* **36**(10A), A24 (2003).

11 D. M. Connor, H. Benveniste, F. A. Dilmanian, *et al.*, "Computed tomography of amyloid plaques in a mouse model of alzheimer's disease using diffraction enhanced imaging," *Neuroimage* **46**(4), 908–914 (2009).

12 H. J. Braun and G. E. Gold, "Diagnosis of osteoarthritis: imaging," *Bone* **51**(2), 278–288 (2012).

13 A. Bashir, M. Gray, R. Boutin, *et al.*, "Glycosaminoglycan in articular cartilage: in vivo assessment with delayed gd(DTPA)(2-)-enhanced MR imaging," *Radiology* **205**(2), 551–558 (1997).

14 R. Reddy, S. Li, E. Noyszewski, *et al.*, "In vivo sodium multiple quantum spectroscopy of human articular cartilage," *Magnetic Resonance in Medicine* **38**(2), 207–214 (1997).

15 R. Stahl, A. Luke, X. Li, *et al.*, "T1rho, T2 and focal knee cartilage abnormalities in physically active and sedentary healthy subjects versus early OA patients: a 3.0-Tesla MRI study," *European Radiology* **19**(1), 132–143 (2009).

16 B. Schmitt, S. Zbyn, D. Stelzeneder, *et al.*, "Cartilage quality assessment by using glycosaminoglycan chemical exchange saturation transfer and 23 Na MR imaging at 7 T," *Radiology* **260**(1), 257–264 (2011).

17 A. J. Sophia Fox, A. Bedi, and S. A. Rodeo, "The basic science of articular cartilage: structure, composition, and function," *Sports health* **1**(6), 461–468 (2009).

18 M. B. Nagarajan, P. Coan, M. B. Huber, *et al.*, "Computer-aided diagnosis in phase contrast imaging x-ray computed tomography for quantitative characterization of ex vivo human patellar cartilage," *IEEE Transactions on Biomedical Engineering* **60**(10), 2896–2903 (2013).

19 A. Z. Abidin, M. B. Nagarajan, W. A. Checefsky, *et al.*, "Volumetric characterization of human patellar cartilage matrix on phase contrast x-ray computed tomography," in *SPIE Medical Imaging*, 94171F–94171F, International Society for Optics and Photonics (2015).

20 G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *arXiv preprint arXiv:1702.05747* (2017).

21 A. Sharif Razavian, H. Azizpour, J. Sullivan, *et al.*, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 806–813 (2014).

22 Y. Bar, I. Diamant, L. Wolf, *et al.*, "Chest pathology detection using deep learning with non-medical training," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 294–297, IEEE (2015).

23 B. van Ginneken, A. A. Setio, C. Jacobs, *et al.*, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 286–289, IEEE (2015).

24 J. Yosinski, J. Clune, Y. Bengio, *et al.*, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, 3320–3328 (2014).

25 T. Schlegl, J. Ofner, and G. Langs, "Unsupervised pre-training across image domains improves lung tissue classification," in *International MICCAI Workshop on Medical Computer Vision*, 82–93, Springer (2014).

26 K. Pritzker, S. Gay, S. Jimenez, *et al.*, "Osteoarthritis cartilage histopathology: grading and staging," *Osteoarthritis and cartilage* **14**(1), 13–29 (2006).

27 V. Ingal and E. Beliaevskaya, "X-ray plane-wave topography observation of the phase contrast from a non-crystalline object," *Journal of Physics D: Applied Physics* **28**(11), 2314 (1995).

28 S. Fiedler, A. Bravin, J. Keyriläinen, *et al.*, "Imaging lobular breast carcinoma: comparison of synchrotron radiation dei-ct technique with clinical ct, mammography and histology," *Physics in medicine and biology* **49**(2), 175 (2004).

29 J. G. Brankov, M. N. Wernick, Y. Yang, *et al.*, "A computed tomography implementation of multiple-image radiography," *Medical physics* **33**(2), 278–289 (2006).

30 P. Coan, A. Peterzol, S. Fiedler, *et al.*, "Evaluation of imaging performance of a taper optics ccdfrelon'camera designed for medical imaging," *Journal of synchrotron radiation* **13**(3), 260–270 (2006).

31 F. Dilmanian, Z. Zhong, B. Ren, *et al.*, "Computed tomography of x-ray index of refraction using the diffraction enhanced imaging method," *Physics in medicine and biology* **45**(4), 933 (2000).

32 A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 1097–1105 (2012).

33 J. Donahue, Y. Jia, O. Vinyals, *et al.*, "Decaf: A deep convolutional activation feature for generic visual recognition.," in *ICML*, 647–655 (2014).

34 Y. Jia, E. Shelhamer, J. Donahue, *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, 675–678, ACM (2014).

35 C. Szegedy, V. Vanhoucke, S. Ioffe, *et al.*, "Rethinking the inception architecture for computer vision," *arXiv preprint arXiv:1512.00567* (2015).

36 M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400* (2013).

37 C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9 (2015).

38 R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE* **67**(5), 786–804 (1979).

39 H. Anys and D. He, "Evaluation of textural and multipolarization radar features for crop classification," *IEEE Transactions on Geoscience and Remote Sensing* **33**(5), 1170–1181 (1995).

40 F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

41 J. Yosinski, J. Clune, A. Nguyen, *et al.*, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579* (2015).

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

43 A. R. Jamieson, M. L. Giger, K. Drukker, *et al.*, "Exploring nonlinear feature space dimension reduction and data representation in breast cadx with laplacian eigenmaps and t-sne," *Medical physics* **37**(1), 339–351 (2010).

44 K. Bunte, B. Hammer, T. Villmann, *et al.*, "Neighbor embedding xom for dimension reduction and visualization," *Neurocomputing* **74**(9), 1340–1350 (2011).

45 K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).

List of Figures

List of Tables