JAISCR, 2022, Vol. 12, No. 2, pp. 101 – 120
10.2478/jaiscr-2022-0007

A PROGRESSIVE AND CROSS-DOMAIN DEEP TRANSFER LEARNING FRAMEWORK FOR WRIST FRACTURE DETECTION

Christophe Karam[1], Julia El Zini[1], Mariette Awad[1], Charbel Saade[2],
Lena Naffaa[3], Mohammad El Amine[3]

[1] *Department of Electrical and Computer Engineering, American University of Beirut*

[2] *Department of Health Professions, Medical Imaging Sciences, Faculty of Health Sciences, American University of Beirut*

[3] *Department of Radiology, Faculty of Medicine, American University of Beirut*

**E-mail: mariette.awad@aub.edu.lb*

Submitted: 24th May 2021; Accepted: 9th August 2021

Abstract

There has been an amplified focus on and benefit from the adoption of artificial intelligence (AI) in medical imaging applications. However, deep learning approaches involve training with massive amounts of annotated data in order to guarantee generalization and achieve high accuracies. Gathering and annotating large sets of training images require expertise which is both expensive and time-consuming, especially in the medical field. Furthermore, in health care systems where mistakes can have catastrophic consequences, there is a general mistrust in the black-box aspect of AI models. In this work, we focus on improving the performance of medical imaging applications when limited data is available while focusing on the interpretability aspect of the proposed AI model. This is achieved by employing a novel transfer learning framework, *progressive transfer learning*, an automated annotation technique and a correlation analysis experiment on the learned representations.

Progressive transfer learning helps jump-start the training of deep neural networks while improving the performance by gradually transferring knowledge from two source tasks into the target task. It is empirically tested on the wrist fracture detection application by first training a general radiology network *RadiNet* and using its weights to initialize $RadiNet_{wrist}$, that is trained on wrist images to detect fractures. Experiments show that $RadiNet_{wrist}$ achieves an accuracy of 87% and an AUC ROC of 94% as opposed to 83% and 92% when it is pre-trained on the ImageNet dataset.

This improvement in performance is investigated within an *explainable AI* framework. More concretely, the learned deep representations of $RadiNet_{wrist}$ are compared to those learned by the baseline model by conducting a correlation analysis experiment. The results show that, when transfer learning is *gradually* applied, some features are learned earlier in the network. Moreover, the deep layers in the *progressive transfer learning* framework are shown to encode features that are not encountered when traditional transfer learning techniques are applied.

In addition to the empirical results, a clinical study is conducted and the performance of $RadiNet_{wrist}$ is compared to that of an expert radiologist. We found that $RadiNet_{wrist}$ exhibited similar performance to that of radiologists with more than 20 years of experience.

This motivates follow-up research to train on more data to feasibly surpass radiologists' performance, and investigate the interpretability of AI models in the healthcare domain where the decision-making process needs to be credible and transparent.

Keywords: deep learning, transfer learning, wrist fracture detection, medical informatics, progressive transfer learning

1 Introduction

Diagnosing wrist fractures from plain radiographs is not an easy task due to complex anatomical structures and the variability of fracture types [1]. An accurate diagnosis requires the expertise of trained and specialized orthopedic physicians who are not always available in any Emergency Rooms (ERs) to assess urgent wrist fracture cases. Misdiagnosed fractures can cause bone displacement or internal injuries which can severely affect the patient [2]. Wrist fractures are one of the most common fracture types in ERs, in both adults and children [3, 4], yet remain one of the most frequently misdiagnosed [5, 6]. Consequently, developing a more automated method of evaluating possible fractures in the wrist from plain radiographs can have a substantial positive impact on the ER by saving time and improving support for both the hospital and the patient. Beyond the ER, this could also be particularly attractive for universities' health services, and sports resorts where medical facilities and support are not regularly available. For such cases, an X-ray would be immediately read by an intelligent software to help in the triage of patients needing medical care, assurance and rest.

Advances in artificial intelligence have shown remarkable success in improving the performance of medical imaging applications [7]. Specifically, Deep Learning (DL) algorithms have been successfully applied to minimize the classification or diagnosis error in wrist fracture radiographs [8, 9, 10]. However, such networks require a significant amount of annotated data in order to produce accurate solutions, which poses a new challenge in the medical domain where data acquisition is expensive and time-consuming when available. Recently, researchers focused on developing approaches to enhance or accelerate the performance of DL, mainly when limited data is available for training. Transfer learning (TL) is a successful method to accelerate DL's integration into the medical field. TL would facilitate feature extraction for small datasets of medical images and improving model performance in diagnosing fractures [11, 12, 13, 14]. Furthermore, automated annotation has been used to increase the size of datasets used in the training of DL algorithms [15, 16]. Although such DL models have proven successful, their lack of interpretability has limited the adoption of AI models in critical decision making domains such as the medical field.

To remedy the aforementioned challenges, we propose employing a *two-tiered domain general transfer learning* that gradually transfers feature representations by first transferring general features learned on a general classification task then specific features learned on a more related task. In the context of wrist fracture detection, *progressive transfer learning* is used to transfer general features from ImageNet [17] to effectively train *RadiNet* on a large set of radiology images. *RadiNet* is later used to transfer radiology-specific features to *RadiNet*$_{wrist}$, a wrist fracture classifier. To further improve the performance of *RadiNet*$_{wrist}$, an automated annotation technique is developed to utilize Natural Language Processing (NLP) techniques in order to analyze radiologists' reports and annotate XRay images for use in the training process. Finally, to engender the medical community trust's in the *progressive transfer learning* framework, we thoroughly *interpret* the deep model representations by studying the learned features in the employed framework as well as the traditional transfer learning framework and their correlations. *RadiNet* is publicly available at [18] so that it can be used as a radiology-specific pre-trained network and help jump-start the performance of radiology networks.

RadiNet pre-training is shown to outperform other pre-training methods including the state-of-the-art ImageNet pre-trained network on the wrist fracture detection and on other radiology applications including finger, shoulder and elbow fracture detection by 4% in accuracy and 2% in AUC ROC. *RadiNet$_{wrist}$*'s performance is tested against expert radiologists where it achieved a comparable accuracy. Besides, the *interpretability* experiment shows that the progressive TL model is able to learn, in its early and mid layers, almost the same representations that the traditional TL model is able to learn in its very deep layers. Additionally, *progressive* TL is able to learn advanced deep representations that are not relatively encountered in the traditional TL model.

The contributions of this work are

- a novel transfer learning approach, *progressive transfer learning*, which *gradually* transfers general and domain-specific features from two source tasks to the target task in the medical imaging domain. Our proposed approach is model-agnostic, i.e. it does not make any assumptions on the underlying model. Moreover, our approach does not require any network architecture modification.

- A classifier that makes use of the proposed progressive learning to detect wrist fractures accurately. This classifier is further extended to improve the performance of general radiology prediction tasks and is publicly available at [18].

- NLP techniques to perform automated annotation of XRay images based on the reports of expert radiologists.

- A comparative clinical study where the performance of *RadiNet$_{wrist}$* is compared to expert radiologists.

The rest of the paper is organized as follows: the literature on fracture detection and automated annotation is presented and compared to this work in Section 2. Then, *progressive and cross domain transfer learning* and its application to the wrist fracture detection along with the automated annotation are described in Section 3. Finally, the experimental setup and results are described in Section 4 before Section 5 concludes the work.

2 Related Work

2.1 Wrist Fracture Detection

Recent years have witnessed a great success of DL algorithms [19] used in medical applications [7, 20, 21, 22, 23] raging from mammography [13, 24, 25, 26] to tomography [27, 28, 29, 30] and ultra-sound [31, 32, 33]. [34] and [35] present an overview of deep learning approaches applied on fracture detection from radiographs and CT scans. Moreover, [36] highlights that fracture detection systems offer poor generalization guarantees when tested on unseen data. Image enhancement techniques have been utilized in [37] to improve the performance of deep networks for arm fracture detection. TL [38, 39] has been also applied to deep networks in order to improve model generalization while reducing needed computational resources and data hunger, the latter of which is especially useful in the medical field where collecting and labeling radiography images can be challenging [11, 12, 13, 14].

As a result, most recent work in radiology [12, 40, 41, 42] has employed TL in order to boost the model's accuracy by initializing its weights from state-of-the-art pre-trained networks, such as ImageNet [17, 43] Inception-ResNet [44] and Faster R-CNN [45]. High accuracy results are reported on a variety of fracture detection in radiographs for various body parts such as hips [46], humeri, forearms, and other various parts [47]. Similar approaches have been applied to wrist fracture detection. For instance, in [8], Inception v3 network was fine-tuned on 11,112 images to predict "fracture" or "no fracture" and achieved an area under the receiver operator characteristic curve (AUC) of 0.954. Similar work has been done by Olczak et al. in [9] who combined multiple exam views to predict the possibility of wrist fracture. Compared to two orthopedic surgeons who reviewed the images at the same resolution, the model was able to achieve an 83% accuracy and a 0.76 as Cohen's kappa metric. Recently, in [48], Faster-R-CNN [45] was trained to extract the distal radius on wrist radiographs as the regions of interest before a CNN model detects the distal radius fracture achieving a 93% accuracy.

While this straight-forward use of TL is beneficial, some researchers have taken different approaches that are more domain-specific. For instance, in [49], a double transfer learning approach is used to classify malign and benign histopathological images for breast cancer. The first step is a feature representation transfer from the ImageNet dataset to the histopathological images dataset for convolutional neural networks, and the second step is to train an SVM classifier on a different dataset in order to filter out irrelevant patches in the target dataset. More specifically, Lindsey et al. [10] applied a domain-specific TL approach in wrist fracture detection. The authors used a dataset of 135,845 radiographs from various body parts, 34,990 of which were wrist radiographs. They used the remaining 105,855 as a pretraining set and fine-tuned their resulting model on the 34,990 wrist images to achieve a 47% decrease in the misinterpretation rate compared to emergency medicine clinicians. Other researchers tackle multi-source domain transfer learning approaches, transferring knowledge from multiple datasets at once. Christodoulidis et. al. [50] use 6 source datasets containing texture information to pre-train 6 CNN networks, which are then fine-tuned on the target dataset of lung tissue, and fused into one model through ensembling. Other methods also involve multiple classification models where some kind of selection or ensembling technique is used ([51, 52]). Yu, et al. [54] and Hu et al. [54] propose two independent methods dubbed "progressive transfer learning" as a way to modify existing transfer learning techniques. For instance, [54] introduce the "*progressive*" aspect of transfer learning on batch-related convolutional cells that are trained on batches of highly variable image data (e.g. viewpoints, occlusions, illumination). These cells encode the dataset information in a latent state used to correct the extracted feature for a given batch. The "progressive" TL of this work is applied on the training phases and is thus different than that of [54] which is applied on the network structure. More specifically, [54] assume a batched architecture of convolutional cells that control the TL, whereas our approach works on *any* deep network by considering a two-phase training paradigm.

[54] use the term "*progressive*" approach on the training dataset to encompass more information as the training progresses. The approach proposed in [54] is fundamentally different from our proposed "progressive" TL. First, their model is application-specific and it is the fusion of deep learning with feature engineering concepts. Second, their "progressive" learning assumes two models: a velocity model and a deep network, and updates their parameters in a complementary fashion toward convergence. On the contrary, our approach is not bound to a particular application and only requires a deep network without any further assumptions.

Gu et al. [54], on the other hand, suggested the "progressive transfer learning" term to describe a domain adaptation technique that works by first fine-tuning an already pre-trained network on target datasets and assess the model's performance on a variety of medically-related tasks. While their *two-tired* TL approach is somewhat similar to ours, the type of tasks and data used in the different steps of the progressive transfer learning differ. We call the reader's attention to the fact that [55] made use of synthetic data to improve the performance of their model and provide some robustness guarantees. While the application of GANs is not straightforward in the context of automating fracture generation, we utilize general radiology XRays to pretrain $RadiNet_{wrist}$. This general pre-training step results in a radiology-specific pretrained network *RadiNet*, and will be publicly available at [18] to jump-start the performance of wider medical applications. In contrast, instead of an intermediate general classification step preceding a final specific classification task, the progressive TL pipeline in [55] is in fact the same classification task on 7 classes, but with different datasets. Additionally, we reinforce our *progressive and cross domain transfer learning* model with experiments inspired from the *explainability* of artificial intelligence models. This will highlight the correlation in the learned representations between the single and progressive transfer learning approaches in order to *explain* the improved performance of progressive transfer learning models. Moreover, we support our work by a clinical study conducted by expert radiologists to compare the performance of our model to the medical assessment of these skilled specialists.

2.2 Automated Annotation

Besides TL, automated annotation is an attempt at overcoming problems caused by data shortages, more specifically, annotated data shortages. Even when data is available, it cannot be used to train supervised deep learning algorithms without task-specific annotations. In the medical field, the lack of annotated data is an even more prominent issue because of the training and expertise required to label medical datasets, which cannot be done by a layperson.

Existing work focuses on linguistic-based approaches to utilize previous annotations to annotate similar medical documents. For instance, in [15], linguistic-based, and reuse-based approaches were investigated to semantically annotate medical documents such as Electronic Health Records (EHR) with concepts of ontology and achieved an accuracy of 59%. In [56], Antolik proposed a novel approach to transferring the information written in medical records into structured EHRs. The preliminary results show that automatic annotation of medical records helps in building systems that can substantially reduce the effort physicians have to conduct when relatively small amounts of data are available. Moreover, Klassen, Xia, and Yetisgen-Yildiz utilized advanced Natural Language Processing (NLP) tools to mark change-of-state events, diagnosis events, coordination, and negation [16]. Their system automatically identifies named entities and medical events in clinical notes with an f-score of 94.7% and 91.8%, respectively.

More recently, automated annotation has been employed to label medical reports term in Serbian. Medication detection in primary care visit conversations was addressed in [57] and automated annotation was proven successful in improving the detection performance. Beyond textual data, automated annotation has been also used to annotate medical images such as in [58] through semi-supervised learning and achieved an 89.8% of papillary thyroid carcinoma regions detection accuracy.

For medical imaging applications, automatic annotation has to rely on existing reports, such as the radiologist report for XRays in fracture diagnosis. It is common practice for these reports to accompany medial imaging test results. Bouslimi and Akaichi [59] utilize a multi-modal approach for semantic annotation in medical imaging, by encoding medical reports with textual bag-of-words, and medical images with visual bag-of-words, before finally combining the two representations using Latent Semantic Analysis (LSA). In [60], researchers used radiology reports to extract medical terminology and map it using MetaMap [61] in order to generate links between pathology concepts and anatomical locations, which can later be used to segment images into regions of interest (ROIs).

2.3 Explainable AI

Explainability methods usually target the model outputs and attempt at generating explanations for a particular model's decision [62, 63, 64]. Image classification tasks rely heavily on these techniques, such as visualizing the attention over the input by using the convolutional layers' ability to localize objects [65], with methods such as Class Activation Maps (CAM) [66], Grad-CAM [67], U-CAM [68]. The usefulness of these techniques also stems from their model-agnosticism: by examining the outputs, they eliminate the need to be specific in their knowledge of the model's internals, as exemplified by LIME: Local Interpretable Model-Agnostic Explanations [69]. Other methods investigate how patterns are encoded in a deep network [70, 71, 72, 73, 74]. Layer-Wise Relevation Propagation (LRP) [75] traces back input contributions to the final output node on a layer-by-layer basis, and has been used to explain decisions in MRI-based Alzheimer's disease classification [76] and multiple sclerosis diagnoses [77]. Canonical Correlation Analysis (CCA) [78] is a particular example of methods that attempt to understand how encoded knowledge relates to human-understandable concepts. CCA has been used to measure the correlation between the brain activity as modeled in deep networks and measured in real-time scenarios [79]. CCA has been also used to train word embeddings in multi-lingual language models [80] and to correlation knowledge encoding in an interpretability framework [81].

3 Methodology

In this section, we first present the mathematical formulation of the *progressive transfer learning*, then we develop *RadiNet*, our deep pretrained network for radiology applications, and *RadiNet*$_{wrist}$, a deep network for wrist fracture detection. Lastly, we describe how the automated annotation is used to increase the size of the dataset on which *RadiNet*$_{wrist}$ is trained.

3.1 Progressive Transfer Learning

In their early layers, deep neural networks learn non-linear representations of the input as low-level and high-level features which are then used to learn a classification in the last layers [82, 83]. Instead of learning the input representation from scratch, transfer learning can relay knowledge from a previously learned task where deep representations are learned on large sets of data. Specifically, in the inductive transfer learning settings, a predictive model in the target task where annotated data exists is learned by transferring knowledge from a source task where annotated data may or may not be available [38]. In this work, we extend inductive transfer learning to a novel *progressive transfer learning* approach where knowledge is transferred from two source tasks to the target task *gradually*.

Definition 1 *(Progressive Transfer Learning)* Given two source domains $\mathcal{D}_g$ and $\mathcal{D}_r$ and their respective learning tasks $\mathcal{T}_g$ and $\mathcal{T}_r$, a target domain $\mathcal{D}_t$ and its corresponding learning task $\mathcal{T}_t$, *progressive transfer learning* aims at improving the learning of the target predictive function $f_T(.)$ in $\mathcal{D}_t$ using the general knowledge in $\mathcal{D}_g$ embedded within the restricted knowledge in $\mathcal{D}_r$.

Specifically, we define $\mathcal{D}_g = \{\mathcal{X}_g, P(\mathcal{X}_g)\}$, $\mathcal{D}_r = \{\mathcal{X}_r, P(\mathcal{X}_r)\}$ and $\mathcal{D}_t = \{\mathcal{X}_t, P(\mathcal{X}_t)\}$, where $\mathcal{X}_g$ ($\mathcal{X}_r$, $\mathcal{X}_t$ resp.) is the feature space of the general (restricted, traget resp.) domain and $P(X)_g$, $P(X)_r$ and $P(X)_t$ are the marginal probability distributions with:

- $X_t = \{(\boldsymbol{x}_t^{(1)}, \boldsymbol{y}^{(1)}), (\boldsymbol{x}_t^{(2)}, \boldsymbol{y}^{(2)}), \ldots, (\boldsymbol{x}_t^{(N)}, \boldsymbol{y}^{(N)})\}$ is a sample of labeled instances in the target domain where each $x_t^{(i)}$ is a vector of pixels intensities representing the image i and $\boldsymbol{y}^{(i)}$ is the label vector

- $X_g = \{(\boldsymbol{x}_g^{(1)}, \boldsymbol{y}^{(1)}), (\boldsymbol{x}_g^{(2)}, \boldsymbol{y}^{(2)}), \ldots, (\boldsymbol{x}_g^{(N_g)}, \boldsymbol{y}^{(N_g)})\}$ is a sample of *general* labeled instances not necessarily related to instances in X_t.

- and $X_r = \{(\boldsymbol{x}_s^{(1)}, \boldsymbol{y}^{(1)}), (\boldsymbol{x}_r^{(2)}, \boldsymbol{y}^{(2)}), \ldots, (\boldsymbol{x}_r^{(N_r)}, \boldsymbol{y}^{(N_r)})\}$ is a sample of labeled instances more *similar* to X_t.

It is worth mentioning that $x_g^{(i)}$ and $x_r^{(i)}$ have the same modality as $\boldsymbol{x}_t^{(i)}$, i.e. images, but are not necessarily of the same nature.

Progressive and cross domain transfer learning works as follows: first, *general* feature representation is transferred from $\mathcal{D}_g$ and task $\mathcal{T}_g$ to efficiently learn restricted feature representation on task $\mathcal{T}_r$ in $\mathcal{D}_r$. Once the general knowledge is embedded in the predictive function $f_T(.)$ in $\mathcal{D}_r$, the feature representation in $\mathcal{D}_r$ is transferred to the target task $\mathcal{T}_t$ as shown in Figure 1.

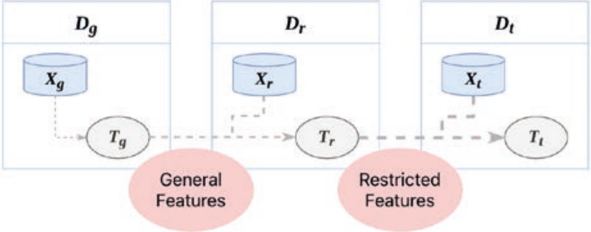


Figure 1. *Progressive transfer learning* workflow

Given the above, one can explain the difference between *inductive* and *progressive* TL as follows: while the former transfers the knowledge from a source task to the target task in a one-shot, the latter performs it in a stepwise fashion. In particular, *progressive transfer learning* first transfers general knowledge from a source task to another intermediate more restricted source task, then transfers the learned restricted knowledge to the target task. This technique can be viewed under the scope of multi-source domain transfer learning, applied to a single learner through step-wise, sequential weight-initialization as opposed to other methods where multiple learners pre-trained on different source domains are fused back into one through ensembling or boosting approaches.

3.2 *RadiNet* And *RadiNet$_{wrist}$* in Wrist Fracture Detection

In this work, the goal, i.e. target task, is to predict wrist fracture from plain radiographs. The general feature representation consists of image features, not necessarily specific to the medical domain, such as edges, corners, and lines and the restricted features are those specific to XRay images such as skewness and energy levels. For this purpose, we define $\mathcal{X}_g$ to be a collection of *general* images, such as ImageNet data, with $\mathcal{T}_g$ their classification task, $\mathcal{X}_r$ to be a collection of XRay images (not necessarily wrist data) and $\mathcal{T}_r$ their corresponding classification task and $\mathcal{X}_t$ to be the set of wrist

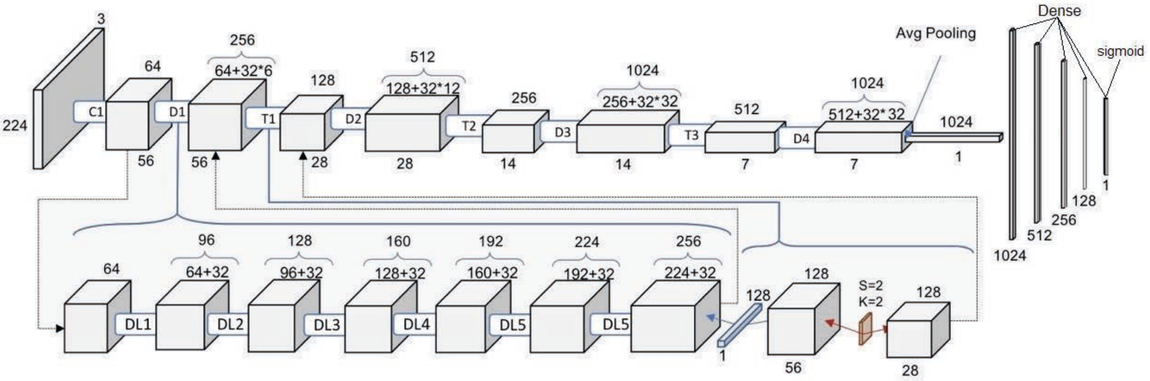


Figure 2. *RadiNet* architecture

XRay images with $\mathcal{T}_t$ being the fracture detection task defined on $\mathcal{X}_t$.

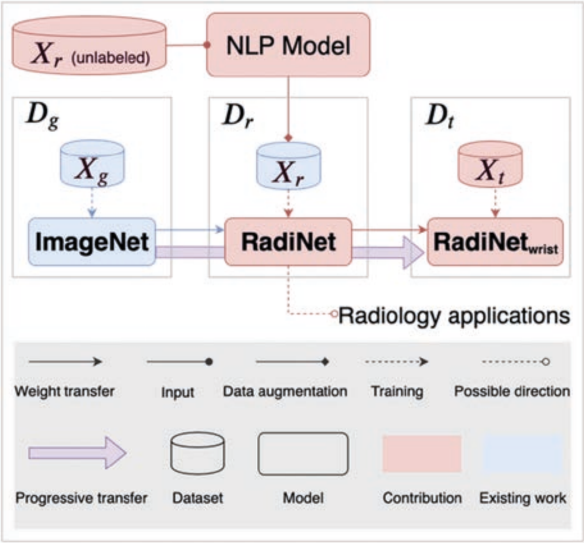


Figure 3. *Progressive transfer learning* workflow on the wrist fracture detection problem

The architecture of *RadiNet* is shown in Figure 2: it takes as input an XRay image and predicts which part of the body the XRay corresponds to: humerus, shoulder, finger, hand, forearm, or wrist. As shown in Fig 3, *RadiNet* is initialized by the weights learned in $\mathcal{D}_g$ to learn general feature representations and fine-tuned in $\mathcal{D}_r$ to learn knowledge representations that are *restricted* to the medical domain. The weights of *RadiNet* are further used in the initialization of *RadiNet$_{wrist}$* in $\mathcal{D}_t$ where *RadiNet$_{wrist}$* takes as input an XRay image of the wrist and predicts whether or not this wrist presents a fracture. When pre-training occurs, the target training set in any of the source datasets, that is: $X_t \not\subset X_g$ and $X_t \not\subset X_r$. It's worth noting that *RadiNet* could be further used as a domain-specific pre-trained network for different radiology applications.

3.3 Automated Labeling

Radiology reports accompany medical imaging in order to interpret the current results (i.e. suspected diagnosis), as well as provide a better understanding of the clinical context (e.g. past medical history) [84]. These reports written by trained radiologists can be used to label using NLP the corresponding radiographs. This would provide researchers annotated datasets without the need of manual intervention which is resource consuming specially if coupled with lack of radiology expertise. In this work, we are presented with N_1 annotated training samples, i.e. N_1 tuples $(\boldsymbol{x}_i, rep_i, y_i)$, where $\boldsymbol{x}_i$ is the vector representing the pixels' intensities of the XRay image, rep_i is the radiologist report written in English and y_i is 0 if there is no wrist fracture, 1 otherwise. We are also presented with N_2 tuples of the form $(\boldsymbol{x}_i, rep_i)$, i.e. XRay images that are not annotated but that are complemented with reports rep_i written by expert radiologists. To improve the performance of *RadiNet$_{wrist}$*, the additional N_2 images are annotated in an automated manner and used in the training process.

For this purpose, a classifier is trained on the N_1 annotated instances of the form (rep_i, y_i) where the complementary reports rep_i, written in English, are vectorized into arrays $a \in \mathbb{R}^c$ of token counts where c is the total number of single words, bi-grams and tri-grams in the training set. More specifically, the vectorization of rep_i produces a_i where $a_i[j]$ is the number of times the word, bi-gram or tri-gram j appears in rep_i as shown in Figure 4. The classifier is then fed the pair (a_i, y_i). After training and testing for accuracy on the annotated dataset, the classifier is then used to infer y_i for the N_2 non-annotated instances. The N_2 non-annotated instances represent around 35% of the final dataset used for training.

4 Experimental Results

In what follows, we start by describing the experimental setup, datasets and baseline models used in this work in Sections 4.1, 4.2 and 4.3 respectively. We then report the results of the following contributions: the performance of progressive transfer learning in 4.4, automated annotation in Section 4.5, generalization to other radiology applications in Section 4.6 and finally the XAI dissection in Section 4.7, the comparison with existing work and the clinical case in Sections 4.8 and 4.9.

4.1 Experimental Setup

The experiments are run on Nvidia Tesla M60 GPU. The algorithms are written in Python 3.6.5 using Keras as a high-level API running on top of a Theano backend [85]. The dataset is split into 80% training, 10% validation, and 10% testing.

All trained models use the DenseNet-169 architecture [86] for feature extraction, with five dense layers on top for the classification. The Adam optimizer is used with parameters $\beta_1 = 0.9, \beta_2 = 0.999$, with an initial learning rate of 10^{-4} that decays by a factor of 0.1 after each 3-epoch plateau for the validation loss, attaining a minimum of 10^{-7}. The models are trained in two phases: a 25-epoch warmup where no fine-tuning occurs, and a 100-epoch phase where fine-tuning does occur, if any, with early stopping implemented to stop the training if the validation loss plateaus for 10 epochs. Class weights are also added to counter the slight imbalance in the data classes.

4.2 Datasets

Two datasets are considered in this work: (1) a general radiology dataset on which *RadiNet* is trained and (2) a wrist fracture dataset on which *RadiNet$_{wrist}$* is further fine-tuned. The general radiology dataset is created by augmenting the Stanford Musculoskeletal Radiographs dataset (MURA) [47] with the American University of Beirut Medical Center's (AUBMC) wrist dataset. *The study complied with the tenets of the Declaration of Helsinki and it was approved by the Institutional Review Board at the American University of Beirut.* The original MURA dataset consists of instances belonging to the following body parts: finger, hand, wrist, forearm, elbow, humerus, shoulder, split into a training and a validation set, as described in Table 1. The wrist dataset consists of 7,776 records provided by AUBMC. Each record consists of the radiologist's report written in English along with a set of different wrist XRay images from multiple views, making up a total of 21,800 images. The records belong to patients from the ages of less than 1 to 99 years, with a mean age of 36 ± 24 years (Figure 5) and 29% pediatric cases (under 18 years), with 43% of the patients being females, 57% being males.

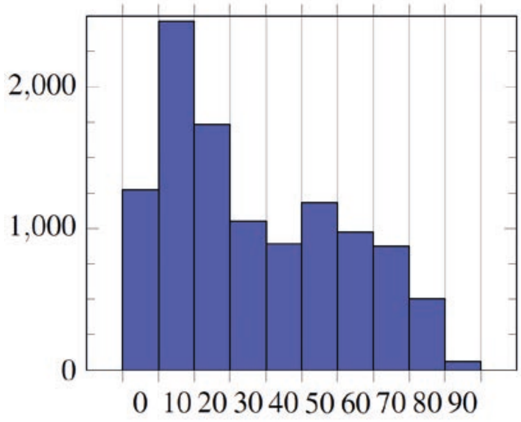


Figure 5. Patient age distribution

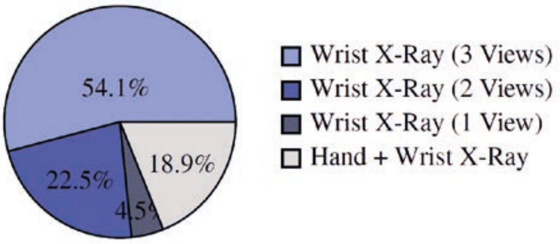


Figure 6. X-Ray exam types

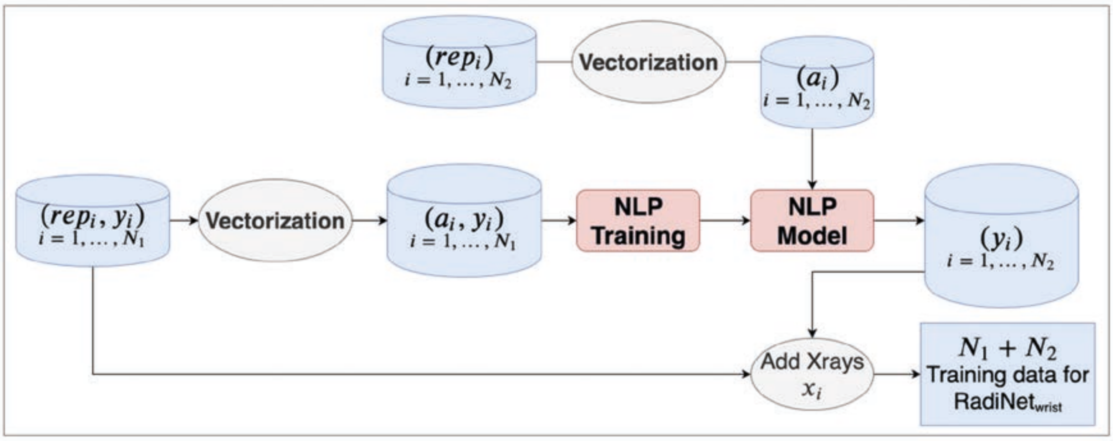

(rep_i) $i = 1, \ldots, N_2$
Vectorization
(a_i) $i = 1, \ldots, N_2$
(rep_i, y_i) $i = 1, \ldots, N_1$
Vectorization
(a_i, y_i) $i = 1, \ldots, N_1$
NLP Training
NLP Model
(y_i) $i = 1, \ldots, N_2$
Add Xrays x_i
$N_1 + N_2$ Training data for RadiNet$_{wrist}$

Figure 4. Automated annotation workflow

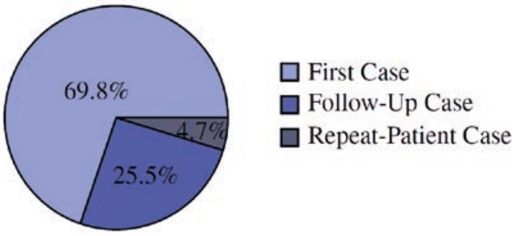


Figure 7. Repeat-patient case

The dataset is processed by converting the DICOM files to JPEG images, resized to 1324x1324 pixels while maintaining the aspect ratio through padding. Only two views are considered for each patient: the frontal and the lateral view, which are then concatenated horizontally to create one image per clinical case. These views were manually picked out from the rest of the dataset which contained multiple exam types resulting in different views for each case, varying from single to multiple-view X-Rays, as shown in Figure 6. It is interesting to note that around 70% of the cases in the dataset belong to first-time patients, while 25% of them are follow-up cases that occurred within a year of the patient's previous case as shown in Figure 10. A small portion of cases (5%) are cases for previously-seen patients, but occurring on average 3 years after the last case which is likely to be the outcome of a different injury for the same patient.

Overall, with regard to the target classes, the data is only slightly skewed, with 56% of the instances representing fractured wrists, and 44% for non-fractures. Finally, after training iterations and performance analyses on the various datasets and their combinations, a clinical case is setup with a test set of 299 patients who are not present in any of the other datasets, as described in section 4.9.

Table 1. General radiology dataset instances

MURA Part	Training Set	Validation Set
Finger	5,106	461
Hand	5,543	460
Wrist	9,748	679
Forearm	1,825	301
Elbow	4,931	465
Humerus	1,272	288
Shoulder	8,379	563
+ AUBMC Wrist	15,220	1,884
Total	52,024	5,101

Figure 8 (top-row) shows a sample of the frontal and lateral views of wrists with and without fractures. The data in this work is noisy, i.e. it contains different objects, such as rings or bracelets, that are not always part of wrist fracture datasets. Some of the XRays, especially those of newborns, might include the hand of a parent holding the kid's hand (bottom-right corner of Figure 8), which can be intuitive for radiologists to recognize but very misleading for a neural network. The data is also representative of real-life cases where patients can be wearing rings, bracelets, and even casts or splints.

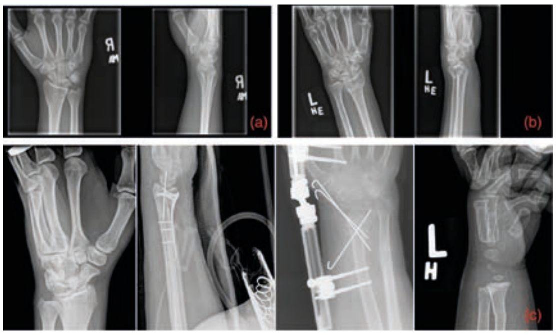


Figure 8. Dataset samples: (a) no wrist fracture, (b) wrist fracture, (c) noisy samples.

4.3 Baseline Models

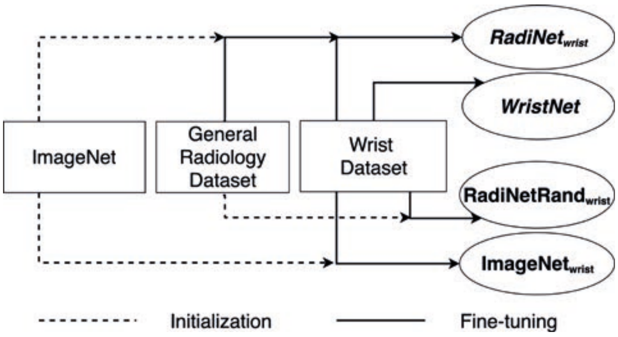


Figure 9. Baseline models

Table 2. Models used in this work

Network	Initialization	Training
WristNet	Random	Wrist
RadiNetRand	Random	Rad.
RadiNetRand$_{\text{wrist}}$	*RadiNetRand*	Wrist
ImageNet$_{\text{wrist}}$	ImageNet	Wrist
RadiNet$_{wrist}$	ImageNet $\mapsto$ Rad.	Wrist

To test the performance of the proposed *progressive transfer learning*, we compare *RadiNet$_{wrist}$* against three baseline models (having the same architecture as *RadiNet*), illustrated in Figure 9 and summarized in Table 2:

- *WristNet*: trained with random weights initialization, i.e. without any type of transfer learning.
- *RadiNetRand*$_{\text{wrist}}$: initialized with the weights of *RadiNetRand*, a network trained on radiology images, then fine-tuned on wrist images. This model relies on domain-specific pre-training but transfers the learned feature representations in one shot. Thus, it tests the importance of transferring weights gradually by comparing *progressive transfer learning* to domain-specific inductive transfer learning.
- *ImageNet*$_{\text{wrist}}$: initialized with ImageNet weights and fine-tuned on the wrist fracture data. Since no fine-tuning has been done on radiology-specific images, this model compares inductive transfer learning with our proposed *progressive transfer learning*.

4.4 *RadiNet$_{wrist}$* Performance

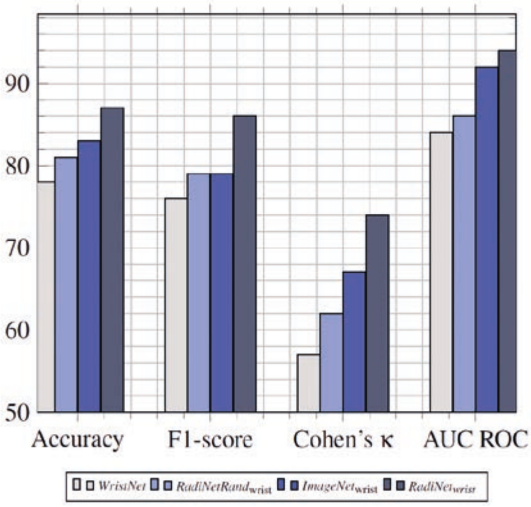


Figure 10. Performance of *RadiNet$_{wrist}$* and the baseline models

Figure 10 illustrates the performance of *RadiNet$_{wrist}$* and the three baseline models. *RadiNet$_{wrist}$*, which implements *progressive transfer learning*, outperforms the baseline models in terms of accuracy, F1-score, Cohen's κ measure, and AUC ROC. Specifically, *RadiNet$_{wrist}$* achieves an accuracy of 87%, an F1-score of 86%, a Cohen's κ measure of 74%, and an AUC ROC of 94%. *RadiNet$_{wrist}$* outperforms *WristNet* by 12%, 13%, 30%, and 12% in the aforementioned metrics showing the importance of transfer learning in improving the classification performance. Moreover, *RadiNet$_{wrist}$* achieves a 7% improvement in accuracy, 9% in F1-score, 19% in κ and 9% in AUC ROC over *RadiNetRand*$_{\text{wrist}}$which applies domain-specific pre-training. Thus, transferring the knowledge *gradually* as in *RadiNet$_{wrist}$*, seems to have merits achieving better performances within the domain-specific inductive transfer learning paradigm.

Finally, *RadiNet*$_{wrist}$ outperforms *ImageNet*$_{\text{wrist}}$ by 5%, 9%, 10% and 2%. Given that *ImageNet*$_{\text{wrist}}$ transfers knowledge from a general pre-trained network one, one can conclude the importance of domain-specific transfer learning over the usual transfer from state-of-the-art pre-trained networks.

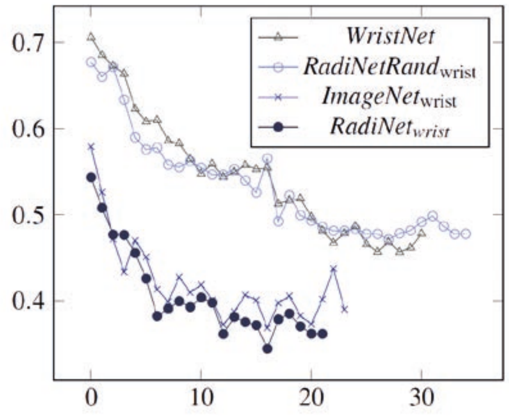


Figure 11. Training loss versus the number of epochs for *RadiNet*$_{wrist}$ and the baseline models

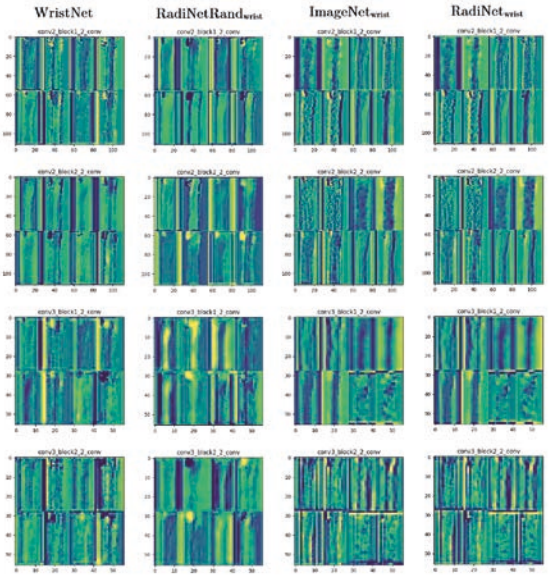


Figure 12. Feature maps of *RadiNet*$_{wrist}$ and the three baseline models

Figure 11 emphasizes the previous results by showing that, compared to the three baseline models, *RadiNet*$_{wrist}$'s training starts with and converges to the least loss. Figure 12 further shows how the features of *RadiNet*$_{wrist}$ are different than those of the three baseline models for different layers (early and late in the network).

Despite being trained on domain-specific data, *RadiNetRand*$_{\text{wrist}}$ fails to outperform the *ImageNet*$_{\text{wrist}}$ model, because the ImageNet dataset is about 20 times bigger than the Radiology dataset, and there is a trade-off between dataset size and domain-specificity. *RadiNet*$_{wrist}$ instead combines the best of these two approaches to achieve an enhanced performance.

4.5 Automated Annotation Results

4.5.1 Automated Annotation Peformance

In this work, the AUBMC wrist dataset used is partially unlabelled. 7414 records are annotated as fractures or non-fractures, while 3698 records do not have corresponding labels. Using the accompanying radiology reports, a pipeline for automated annotation was developed to complete the annotation of this dataset. The pipeline is trained and tested on the labelled portion of the dataset, and is then evaluated on the unlabelled portion of the dataset, effectively generating labels to be used for training the image classification network, as depicted in Figure 13. The report text was cleaned, vectorized into token counts, taking into account unigrams, bigrams, and trigrams, before being fed to a classifier.

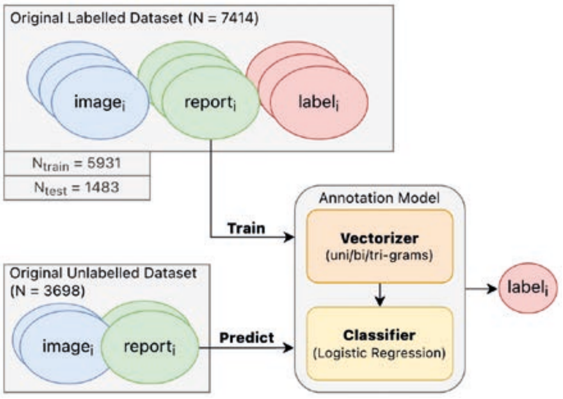


Figure 13. Automated annotation process

Table 3 shows the confusion matrix of the trained logistic regression classifier with an accuracy of 98.4%, averaged over 5 runs. The false positive and false negative rates are both low at around 1% which guarantees accurate labeling for training images used in *RadiNet$_{wrist}$*. The logistic regression outperformed other models, namely Gradient Boosting (mean accuracy of 98.0%), Random Forest (97.6%), and Naïve Bayes (97.0%).

Table 3. Confusion matrix reporting the per-class accuracies (%) of the automated annotation

	Predicted: Fracture	Predicted: No Fracture
Actual: Fracture	679	11
Actual: No Fracture	12	781

4.5.2 Impact on System Performance

An ablation study was performed to study the contribution of the automated annotation to the system. For this purpose, we consider the Wrist$_{\text{minimal}}$ dataset consisting of the manually labeled wrist fracture data where the data annotated in this work was removed. The ablation study trains *RadiNet$_{wrist}$* and the three baseline models on Wrist$_{\text{minimal}}$. As shown in Table 4, *RadiNet$_{wrist}$* performs better than the three baseline models. Moreover, *RadiNet$_{wrist}$* when trained on Wrist$_{\text{minimal}}$ achieves a 3% lower accuracy and a 4% lower ROC than the same network trained on the complete dataset, i.e. with the data provided by the automated annotation.

4.6 *Progressive Transfer Learning* Performance on Other Radiology Applications

RadiNet is trained on a large set of XRay images and is thus able to serve as a pre-trained network for diverse radiology applications. In this work, we consider six fracture types, and we perform the classification by transferring knowledge from *RadiNet*. Table 4 shows different metrics reported on the different fracture detection applications trained using *RadiNet* as a pre-trained network. *RadiNet* pre-training performs consistently better than training from scratch and than fine-tuning RadiNetRand and ImageNet networks. Specifically, *RadiNet* pre-training achieves 76.38% accuracy on the shoulder fracture detection compared to 73.71%, 64.3%, and 66.96% when ImageNet pre-training, RadiNetRand pre-training, and no pre-training are performed respectively. *RadiNet* pretraining gives better results in almost all metrics except for the precision on the *finger* dataset. However, on the same dataset ImageNet pre-training has a recall of 31%, which is low compared to 70.44% recall achieved by *RadiNet* pre-training. These results show that *progressive transfer learning* performs better than the inductive and the domain-specific transfer learning on different fracture detection types.

4.7 Correlation Analysis

The empirical results show that *RadiNet$_{wrist}$* outperforms *ImageNet*$_{\text{wrist}}$on the wrist fracture detection task. In what follows, we try to investigate the hidden representations to explain *why* progressive transfer learning improves such performance from an explainability perspective. Motivated by the interpretability of learned representations, we study the similarities between the representations that *RadiNet$_{wrist}$* and *ImageNet*$_{\text{wrist}}$ have learned by analyzing the activation vectors of selected layers in both networks. For this purpose, we rely on the Canonical Correlation Analysis (CCA) method presented in [78] to compare the deep representations learned in different transfer learning approaches. CCA is used in the literature to compute the similarity between the model's features and the brain activity [79], word embeddings in multi-lingual domains [80] and to analyze the representations of deep models [81].

We consider the layer representation to be its finite set of responses over a finite set of input instances drawn from the validation dataset. We study the correlation between the features outputted by *RadiNet$_{wrist}$*and *ImageNet$_{wrist}$* at 16 different layers of their architecture and we show the CCA results in Figure 14. We can see that the correlation between the early stages of *RadiNet$_{wrist}$*and the middle stages of *ImageNet$_{wrist}$* are moderately correlated, inferring *RadiNet$_{wrist}$*'s ability to extract radiological features right from the start that are not encountered with single transfer learning until deeper into the network. If we examine the correlations between the features produced at the end of *RadiNet$_{wrist}$*as compared to the different stages of *ImageNet$_{wrist}$* , we observe low values that could hint at the fact that the progressive transfer learning results in the encoding of features that are not encountered during traditional transfer learning at any stage in the net-

Table 4. Performance of the baseline models and *RadiNet*$_{wrist}$ on different validation sets with the best results in each dataset are highlighted in bold

Dataset	Pre-training	Validation Accuracy	Precision	Recall	F1-score	Cohen's κ	ROC AUC
Wrist$_{\text{MURA}}$	-	82.45	83.44	75.98	79.53	77.12	86.76
	RadiNetRand	84.20	85.35	74.86	79.76	77.92	87.21
	ImageNet	86.33	85.71	**80.45**	83.00	81.39	89.65
	RadiNet	**87.09**	**87.12**	79.33	**83.04**	**81.76**	**90.79**
Finger	-	72.40	76.14	67.20	71.39	42.52	75.23
	RadiNetRand	72.62	81.67	59.51	68.85	43.19	76.63
	ImageNet	71.96	**92.77**	31.17	46.66	26.98	80.6
	RadiNet	**75.49**	76.60	**70.44**	**73.41**	**45.39**	**81.22**
Shoulder	-	68.92	71.10	55.76	62.50	33.74	71.99
	RadiNetRand	67.01	67.74	52.88	59.39	44.48	71.03
	ImageNet	75.86	75.19	69.78	72.39	47.37	81.75
	RadiNet	**79.71**	**79.84**	**69.78**	**74.47**	**52.67**	**83.90**
Elbow	-	68.56	75.34	47.83	58.51	49.51	71.87
	RadiNetRand	70.43	73.53	54.35	62.50	52.23	72.88
	ImageNet	71.35	75.14	56.52	64.52	55.59	74.77
	RadiNet	**74.33**	**79.88**	**56.96**	**66.50**	**60.36**	**76.92**

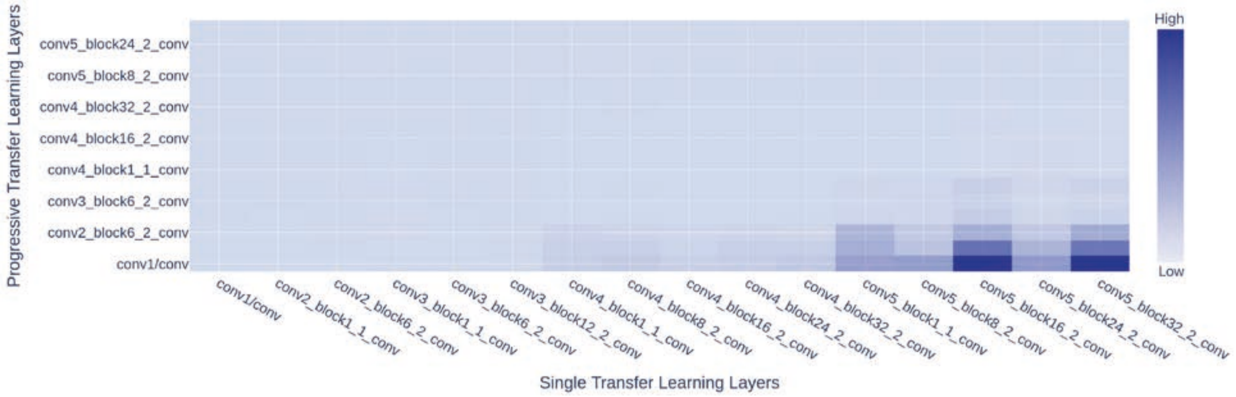


Figure 14. Feature correlations produced by CCA between the single transfer learning model (*ImageNet*$_{\text{wrist}}$ in the x-axis) and the progressive model (*RadiNet*$_{wrist}$ in the y-axis)

work. The last layers of the *RadiNet*$_{wrist}$model are not correlated to any other layers, supporting the argument that suggests it might be learning new features not encountered by the *ImageNet*$_{wrist}$ model.

4.8 Comparison With Literature

Unfortunately, a thorough comparison with results in the literature relating to wrist fractures is not entirely possible due to the unavailability of their datasets (with the exception of the Stanford MURA dataset). However, we can argue that our results are comparable with other studies. As previously reported, [9] obtained an 83% accuracy and a 0.76 Cohen's kappa in classifying wrist fractures which is on par with our results, as is Kim and MacKinnon's work [8] which achieved an AUC value of 0.954.

Table 5. *RadiNet*$_{wrist}$ and MURA$_{\text{baseline}}$ accuracies (%) on MURA validation sets

	Finger	Elbow	Shoulder	Wrist
MURA$_{\text{baseline}}$	38.9	71.0	72.9	**93.1**
RadiNet$_{wrist}$	**75.5**	**74.3**	**79.7**	85.2

Authors in [48] reported a 93% accuracy in detecting wrist fractures, but their work only focuses on the fractures of the distal radius area and relies on Faster-RCNN [45] to extract the ROI. We are aware that an accuracy reading is not sufficient for a fair qualitative assessment because it doesn't reflect the model specificity and sensitivity. However, we do not have access to the models and datasets which makes the assessment of the model's misses hard. Consequently, a dissection of the model's false predictions is not feasible to assess whether they are, clinically, on the easy or difficult level.

Lastly, compared to the work the MURA dataset in [47], we can see our model improves on their baseline model's accuracy for abnormality in the finger, elbow, and shoulder, but not for the wrist, as seen in Table 5. It is important to note that [47] used ensembling to boost performance, and a different testing scheme whereby the final prediction for a patient is an aggregate of the predictions on multiple images for that same patient.

4.9 Clinical Case

To evaluate how well *RadiNet*$_{wrist}$ generalizes to unseen data, we test its performance on 299 patients that were not included in the training or validation sets. 10% of the images were of low-resolution whereas only 3% of the training dataset consists of low resolution images.

Each patient is represented by two XRay images: one for the frontal view of the wrist and another one for the lateral view. The predictions of *RadiNet*$_{wrist}$ when trained on the Wrist$_{\text{minimal}}$ dataset, i.e. without the data produced by the automated labeling, and on the complete wrist dataset are compared against the ground truth labels and the accuracy, precision, and recall (sensitivity) are reported in Table 6. Two expert radiologists were asked to predict wrist fractures from the frontal views and the lateral views separately. Two rules are later applied to the separate results to predict a fracture in the wrist from both views. The conjunction rule predicts a wrist fracture if the radiologist labels *both* views as fracture and the disjunction rule predicts fracture if the radiologist labels *at least one* view as a fracture. Table 6 shows that disjunction gives better results in terms of accuracy and recall but lower results in terms of precision. Moreover, *RadiNet*$_{wrist}$ achieves an accuracy of 83.61% which is slightly higher than the average radiologist accuracy (83.28%) when the conjunction is applied and lower than the average radiologist accuracy when a disjunction is applied by 3.68%. Notably, *RadiNet*$_{wrist}$'s performance significantly decreased when the automated annotation data is not used for training. This shows that increasing the dataset size by automating the annotation process provides better generalization guarantees for *RadiNet*$_{wrist}$and this can be clearly seen through the poor performance metrics under Wrist$_{\text{minimal}}$, as shown in Table 6. Table 6 also shows that *RadiNet*$_{wrist}$ generally outperforms our baseline models in terms of accuracy, precision and recall. Only ImageNet$_{wrist}$'s precision was shown to be higher than that of *RadiNet*$_{wrist}$ by 0.73% while the overall accuracy and recall are higher with *RadiNet*$_{wrist}$.

However, to explain these results, it is necessary to examine the misclassified cases, for both the model and the radiologist. The radiologists have reassessed the difficulty of these misclassified cases

Table 6. Results of the clinical case: performance of *RadiNet*$_{wrist}$ when trained with and without the data of the automated labeling versus the performance of two expert radiologists.

	Radiologist A		**Radiologist B**		**Average Radiologist**			***RadiNet*$_{wrist}$**		***ImageNet*$_{\text{wrist}}$**	***RadiNetRand*$_{\text{wrist}}$**	***WristNet***
	Conjunction	Disjunction	Conjunction	Disjunction	Conjunction	Disjunction	Average	Wrist$_{\text{minimal}}$	Wrist			
Accuracy	81.94	85.62	84.62	88.96	83.28	87.29	85.29	66.56	83.61	79.93	74.91	71.24
Precision	87.88	84.27	88.16	84.47	88.02	84.37	86.20	56.25	81.00	81.73	72.11	73.01
Recall	55.77	72.12	64.42	83.65	60.09	77.88	68.99	17.31	71.96	67.46	61.98	56.72

on a simple binary-scale, easy or difficult. This assessment is of course biased and based on the reviewer's clinical experience. We thus define a binary difficulty score of d_i for each case. Consequently, for each "agent" participating in the study (model or radiologist), we can calculate an average difficulty score for the N misclassified cases by that agent as follows:

$$D_{agent} = \frac{1}{N}\sum_{i=1}^{n} d_i = \frac{d_1 + d_2 + \cdots + d_n}{N}$$

Then, we construct the Venn diagram that shows the overlap of incorrectly-classified cases between the different agents, along with their corresponding average difficulty score D_{agent}, and the ratio R_{agent} of misclassified cases by the agent to the total number of misclassified cases in Figure 15. The diagram shows that the model failed to correctly classify cases that were mostly easily judged by the radiologists. We refer to an easy case one that usually translates into an obvious fracture present in the XRay. Given the variation in X-Ray quality - lower than average resolution, that was present in the training images and the limited data the model had access to as compared to the radiologists' experiences both in the number of years and quantity of studied XRays, these failures point out the performance shortcomings of the model which could be improved by more training on more data or using different networks. More importantly, this highlights the ever-growing need for more explainable AI that is very important in healthcare applications given the consequences and impacts of AI decision on human lives.

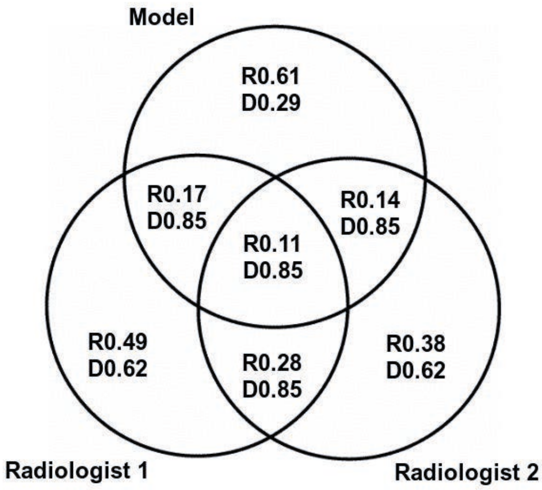


Figure 15. Misclassification Venn diagram

5 Conclusion

In this work, we proposed a progressive and cross domain transfer learning approach for wrist fracture detection application by transferring general features learned from *ImageNet* to learn more specific features from a related radiology dataset in *Radinet* before fine-tuning the specialized features on the target wrist dataset in *RadiNet*$_{wrist}$. This *stepwise* approach was able to provide an increase in the performance of the model over regular *inductive* transfer learning methods, allowing us to detect wrist fractures in radiographs with an accuracy of 87% and AUC ROC of 94% (Figure 10). Furthermore, *Radinet*, when used as a pre-trained network for radiology applications, was shown to outperform state-of-the-art pre-training by 4% in accuracy. This novel transfer learning method was also combined with NLP techniques to automate the annotation and increase the dataset size for an added improvement in accuracy and generalization.

This improvement is studied within the *explainable AI* framework to leverage the mistrust of the medical field in AI applications. The correlation analysis shows that the better performance of *progressive transfer learning* can be explained by its ability to learn domain-specific features in its early layers, i.e. faster than the traditional approaches that learn such features in their last layers, while distinctly extracting new features in its deepest layers.

Radinet could be also used as a domain-specific network for further fine-tuning other radiology applications, thus offering an increase in performance compared to regular transfer learning from ImageNet. A clinical case study showed that our model's performance is close to that of the radiologist, whose reading can be affected by subjectivity and human fatigue. Conducting a larger clinical study can help obtain a more accurate comparison between the deep learning model and radiologists with different backgrounds and different experience levels.

Future work can focus on the classification of displaced fractures, the localization of fractures, and the automation of report generation.

Acknowledgment

This work is funded by the University Research Board and the Maroun Semaan Faculty of Engineering and Architecture at the American University of Beirut. The authors would like to thank Alexandra Sakkah, Samir El Khatib and Abdel Rahman El Kurdi for their earlier manipulation of the data, Prof. Nassir Sabah for his early engagement in the project, and Walid Rafeh for his support.

References

[1] W. Cooney, R. Bussey, J. Dobyns, and R. Linscheid, "Difficult wrist fractures. perilunate fracture-dislocations of the wrist.," Clinical Orthopaedics and Related Research, no. 214, pp. 136–147, 1987.

[2] R. Lindsey, A. Daluiski, S. Chopra, A. Lachapelle, M. Mozer, S. Sicular, et al., "Deep neural network improves fracture detection by clinicians," Proceedings of the National Academy of Sciences, vol. 115, no. 45, pp. 11591–11596, 2018.

[3] C. M. Court-Brown and B. Caesar, "Epidemiology of adult fractures: A review," Injury, vol. 37, pp. 691–697, Aug 2006.

[4] C. A. Goldfarb, Y. Yin, L. A. Gilula, A. J. Fisher, and M. I. Boyer, "Wrist fractures: What the clinician wants to know," Radiology, vol. 219, no. 1, pp. 11–28, 2001. PMID: 11274530.

[5] H. R. Guly, "Injuries initially misdiagnosed as sprained wrist (beware the sprained wrist)," Emergency Medicine Journal, vol. 19, no. 1, pp. 41–42, 2002.

[6] B. Petinaux, R. Bhat, K. Boniface, and J. Aristizabal, "Accuracy of radiographic readings in the emergency department," The American Journal of Emergency Medicine, vol. 29, pp. 18–25, Jan 2011.

[7] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, et al., "A survey on deep learning in medical image analysis," Medical image analysis, vol. 42, pp. 60–88, 2017.

[8] D. Kim and T. MacKinnon, "Artificial intelligence in fracture detection: Transfer learning from deep convolutional neural networks," Clinical Radiology, vol. 73, 12 2017.

[9] J. Olczak, N. Fahlberg, A. Maki, A. Razavian, A. Jilert, A. Stark, et al., "Artificial intelligence for analyzing orthopedic trauma radiographs: Deep learning algorithms—are they on par with humans for diagnosing fractures?," Acta Orthopaedica, vol. 88, pp. 1–6, 07 2017.

[10] R. Lindsey, A. Daluiski, S. Chopra, A. Lachapelle, M. Mozer, S. Sicular, et al., "Deep neural network improves fracture detection by clinicians," Proceedings of the National Academy of Sciences, vol. 115, no. 45, pp. 11591–11596, 2018.

[11] D. Soekhoe, P. van der Putten, and A. Plaat, "On the impact of data set size in transfer learning using deep neural networks," Lecture Notes in Computer Science Advances in Intelligent Data Analysis XV, pp. 50–60, 2016.

[12] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, et al., "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1285–1298, 2016.

[13] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," Journal of Medical Imaging, vol. 3, no. 3, p. 034501, 2016.

[14] A. Van Opbroek, M. A. Ikram, M. W. Vernooij, and M. De Bruijne, "Transfer learning improves supervised image segmentation across imaging protocols," IEEE transactions on medical imaging, vol. 34, no. 5, pp. 1018–1030, 2014.

[15] V. Christen, A. Groß, and E. Rahm, "Approaches for annotating medical documents.," in LWDA, pp. 227–232, 2016.

[16] P. Klassen, F. Xia, and M. Yetisgen-Yildiz, "Annotating and detecting medical events in clinical notes," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 3417–3421, 2016.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, Ieee, 2009.

[18] C. Karam, J. El Zini, and M. Awad, "X-ray wrist fracture classification," 2019.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, p. 436, 2015.

[20] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," Radiology, vol. 284, no. 2, pp. 574–582, 2017.

[21] M. P. McBee, O. A. Awan, A. T. Colucci, C. W. Ghobadi, N. Kadom, A. P. Kansagra, et al., "Deep learning in radiology," Academic radiology, vol. 25, no. 11, pp. 1472–1480, 2018.

[22] J. H. Thrall, X. Li, Q. Li, C. Cruz, S. Do, K. Dreyer, et al., "Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success," Journal of the American College of Radiology, vol. 15, no. 3, pp. 504–508, 2018.

[23] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on mri," Journal of Magnetic Resonance Imaging, vol. 49, no. 4, pp. 939–954, 2019.

[24] A. S. Becker, M. Marcon, S. Ghafoor, M. C. Wurnig, T. Frauenfelder, and A. Boss, "Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer," Investigative radiology, vol. 52, no. 7, pp. 434–440, 2017.

[25] J. Wang, X. Yang, H. Cai, W. Tan, C. Jin, and L. Li, "Discrimination of breast cancer with microcalcifications on mammography by deep learning," Scientific reports, vol. 6, p. 27327, 2016.

[26] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," Scientific reports, vol. 8, no. 1, p. 4165, 2018.

[27] M. Araya-Polo, J. Jennings, A. Adler, and T. Dahlke, "Deep-learning tomography," The Leading Edge, vol. 37, no. 1, pp. 58–66, 2018.

[28] K.-L. Hua, C.-H. Hsu, S. C. Hidayati, W.-H. Cheng, and Y.-J. Chen, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," OncoTargets and therapy, vol. 8, 2015.

[29] T. Würfl, F. C. Ghesu, V. Christlein, and A. Maier, "Deep learning computed tomography," in International conference on medical image computing and computer-assisted intervention, pp. 432–440, Springer, 2016.

[30] H. Zhang, L. Li, K. Qiao, L. Wang, B. Yan, L. Li, et al., "Image prediction for limited-angle tomography via deep learning with convolutional neural network," arXiv preprint arXiv:1607.08707, 2016.

[31] M. H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís, R. Zwiggelaar, et al., "Automated breast ultrasound lesions detection using convolutional neural networks," IEEE journal of biomedical and health informatics, vol. 22, no. 4, pp. 1218–1226, 2017.

[32] K. Lekadir, A. Galimzianova, À. Betriu, M. del Mar Vila, L. Igual, D. L. Rubin, et al., "A convolutional neural network for automatic characterization of plaque composition in carotid ultrasound," IEEE journal of biomedical and health informatics, vol. 21, no. 1, pp. 48–55, 2016.

[33] P. Burlina, S. Billings, N. Joshi, and J. Albayda, "Automated diagnosis of myositis from muscle ultrasound: Exploring the use of machine learning and deep learning methods," PloS one, vol. 12, no. 8, p. e0184059, 2017.

[34] P. H. Kalmet, S. Sanduleanu, S. Primakov, G. Wu, A. Jochems, T. Refaee, A. Ibrahim, L. v. Hulst, P. Lambin, and M. Poeze, "Deep learning in fracture detection: a narrative review," Acta orthopaedica, vol. 91, no. 2, pp. 215–220, 2020.

[35] R. M. Jones, A. Sharma, R. Hotchkiss, J. W. Sperling, J. Hamburger, C. Ledig, R. O'Toole, M. Gardner, S. Venkatesh, M. M. Roberts, et al., "Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs," NPJ digital medicine, vol. 3, no. 1, pp. 1–6, 2020.

[36] A. M. Raisuddin, E. Vaattovaara, M. Nevalainen, M. Nikki, E. Järvenpää, K. Makkonen, P. Pinola, T. Palsio, A. Niemensivu, O. Tervonen, et al., "Critical evaluation of deep neural networks for

wrist fracture detection," Scientific reports, vol. 11, no. 1, pp. 1–11, 2021.

[37] B. Guan, G. Zhang, J. Yao, X. Wang, and M. Wang, "Arm fracture detection in x-rays based on improved deep convolutional neural network," Computers & Electrical Engineering, vol. 81, p. 106530, 2020.

[38] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2010.

[39] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in Proceedings of the 24th international conference on Machine learning, pp. 759–766, ACM, 2007.

[40] H. Ravishankar, P. Sudhakar, R. Venkataramani, S. Thiruvenkadam, P. Annangi, N. Babu, et al., "Understanding the mechanisms of deep transfer learning for medical images," in Deep Learning and Data Labeling for Medical Applications, pp. 188–196, Springer, 2016.

[41] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, et al., "Convolutional neural networks for medical image analysis: Full training or fine tuning?," IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1299–1312, 2016.

[42] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," Radiographics, vol. 37, no. 2, pp. 505–515, 2017.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, pp. 1097–1105, 2012.

[44] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[45] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, pp. 91–99, 2015.

[46] T. Urakawa, Y. Tanaka, H. Matsuzawa, K. Watanabe, and N. Endo, "Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network," Journal of the International Skeletal Society A Journal of Radiology, Pathology and Orthopedics, vol. 42, pp. 239–244, 2019.

[47] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, et al., "Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs," 2017. cite arxiv:1712.06957.

[48] K. Gan, D. Xu, Y. Lin, Y. Shen, T. Zhang, K. Hu, et al., "Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments," Acta orthopaedica, pp. 1–12, 2019.

[49] J. de Matos, A. de Souza Britto Jr., L. E. S. Oliveira, and A. L. Koerich, "Double transfer learning for breast cancer histopathologic image classification," CoRR, vol. abs/1904.07834, 2019.

[50] S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, and S. G. Mougiakakou, "Multisource transfer learning with convolutional neural networks for lung pattern analysis," CoRR, vol. abs/1612.02589, 2016.

[51] J. Li, W. Wu, D. Xue, and P. Gao, "Multi-source deep transfer neural network algorithm," Sensors (Basel, Switzerland), vol. 19, p. 3992, Sep 2019. 31527437[pmid].

[52] R. Gupta and L.-A. Ratinov, "Text categorization with knowledge transfer from heterogeneous data sources," in AAAI, pp. 842–847, 2008.

[53] Z. Yu, Z. Jin, L. Wei, J. Guo, J. Huang, D. Cai, X. He, and X.-S. Hua, "Progressive transfer learning for person re-identification," Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Aug 2019.

[54] W. Hu, Y. Jin, X. Wu, and J. Chen, "Progressive transfer learning for low frequency data prediction in full waveform inversion," 2019.

[55] Y. Gu, Z. Ge, C. P. Bonnington, and J. Zhou, "Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 5, pp. 1379–1393, 2020.

[56] J. Antolík, "Automatic annotation of medical records," Studies in health technology and informatics, vol. 116, pp. 817–822, 2005.

[57] C. Ganoe, W. Wu, P. Barr, W. Haslett, M. Dannenberg, K. Bonasia, J. Finora, J. Schoonmaker, W. Onsando, J. Ryan, et al., "Natural language processing for automated annotation of medication mentions in primary care visit conversations," medRxiv, 2021.

[58] H. Li, B. Zhang, Y. Zhang, W. Liu, Y. Mao, J. Huang, and L. Wei, "A semi-automated annotation algorithm based on weakly supervised

learning for medical images," Biocybernetics and Biomedical Engineering, vol. 40, no. 2, pp. 787–802, 2020.

[59] R. Bouslimi and J. Akaichi, "New approach for automatic medical image annotation using the bag-of-words model," in 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 1088–1093, 2015.

[60] T. Gong, S. Li, J. Wang, C. L. Tan, B. Pang, T. Lim, C. Lee, Q. Tian, and Z. Zhang, "Automatic labeling and classification of brain ct images," pp. 1581–1584, 09 2011.

[61] A. R. Aronson and F.-M. Lang, "An overview of metamap: historical perspective and recent advances," Journal of the American Medical Informatics Association : JAMIA, vol. 17, no. 3, pp. 229–236, 2010. PMC2995713[pmcid].

[62] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," CoRR, vol. abs/1311.2901, 2013.

[63] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," arXiv preprint arXiv:1703.01365, 2017.

[64] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[65] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," 2015.

[66] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba, "Learning deep features for discriminative localization.," CVPR, 2016.

[67] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Mar 2018.

[68] B. N. Patro, M. Lunayach, S. Patel, and V. P. Namboodiri, "U-cam: Visual explanation using uncertainty based class activation maps," in Proceedings of the IEEE International Conference on Computer Vision, pp. 7444–7453, 2019.

[69] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[70] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6541–6549, 2017.

[71] R. Fong and A. Vedaldi, "Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8730–8738, 2018.

[72] K. Leino, S. Sen, A. Datta, M. Fredrikson, and L. Li, "Influence-directed explanations for deep convolutional networks," in 2018 IEEE International Test Conference (ITC), pp. 1–8, IEEE, 2018.

[73] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5188–5196, 2015.

[74] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4829–4837, 2016.

[75] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PLoS ONE, vol. 10, 2015.

[76] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification," Frontiers in aging neuroscience, vol. 11, pp. 194–194, Jul 2019. 31417397[pmid].

[77] F. Eitel, E. Soehler, J. Bellmann-Strobl, A. U. Brandt, K. Ruprecht, R. M. Giess, J. Kuchling, S. Asseyer, M. Weygandt, J.-D. Haynes, M. S. l, F. Paul, and K. Ritter, "Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional mri using layer-wise relevance propagation," NeuroImage: Clinical, vol. 24, p. 102003, 2019.

[78] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," Neural computation, vol. 16, no. 12, pp. 2639–2664, 2004.

[79] D. Sussillo, M. M. Churchland, M. T. Kaufman, and K. V. Shenoy, "A neural network that finds a naturalistic solution for the production of muscle activity," Nature neuroscience, vol. 18, no. 7, pp. 1025–1033, 2015.

[80] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," in Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 462–471, 2014.

[81] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," in Advances in Neural Information Processing Systems, pp. 6076–6085, 2017.

[82] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[83] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in Advances in neural information processing systems, pp. 3320–3328, 2014.

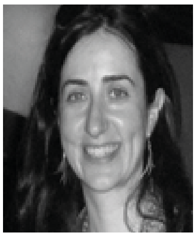
[84] C. Castillo, T. Steffens, L. Sim, and L. Caffery, "The effect of clinical information on radiology reporting: A systematic review," Journal of Medical Radiation Sciences, vol. 68, no. 1, pp. 60–74, 2021.

[85] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," arXiv e-prints, vol. abs/1605.02688, May 2016.

[86] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," CoRR, vol. abs/1608.06993, 2016.

Christophe Karam obtained his Bachelor in Mechanical Engineering from the American University of Beirut in 2019 and is currently pursuing his masters in Image and Signal Processing at the Grenoble INP. His research interests include robotics, object detection, and generative adversarial networks.

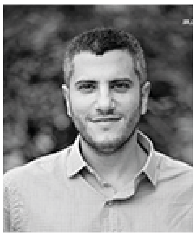
Julia El Zini is a Ph.D. student enrolled in the electrical and computer engineering department at the American University of Beirut (AUB). She has received her B.S. and M.S. in computer science from AUB, Lebanon, in 2015 and 2017, respectively. Her research interests include distributed optimization, parallel computing, reinforcement learning, multi-task and transfer learning, and scalable machine learning applications.



Prof. Mariette Awad is a tenured associate professor in the Electrical and Computer Engineering Department of the American University of Beirut. Her book "Efficient Learning Machines: Theories, Concepts and Applications for Engineers and Systems", published in 2015, is among the most open source downloaded books for Summer 2020 according to Springer Nature. She has more than 100 patents, conferences and journals articles and she is managing multimillions grants. Her current research interests include machine learning, data analytics and internet of things. She can be reached at mariette.awad@aub.edu.lb

Charbel Saade graduated from the University of Sydney in Medical Imaging Sciences, Masters in CT and MRI, and a Ph.D. in contrast media delivery. Charbel is currently working as Assistant Professor and Program Coordinator of Medical Imaging Sciences at the American University of Beirut. He is a frequent lecturer in international congresses like RSNA, ECR and Arab Health and has co-authored more than 90 original papers.

Lena Naffaa graduated with M.D. degree from the Lebanese University (1991-1998). She is certified by the American Board of Radiology in Diagnostic Radiology, Pediatric Radiology and as an Authorized User in Nuclear Medicine. Dr. Naffaa has been on staff at Akron Children's Hospital since 2006, promoted to Assistant Professor of Radiology at Northeastern Ohio Medical College of Medicine in 2010 and to Associate Professor of Radiology in 2015. Dr. Naffaa has served as Director of Pediatric Radiology rotation and Director of Nuclear Medicine at Akron children's Hospital since 2006.



Mohammad El Amine graduated with M.D. degree from the Lebanese University in 2015. Currently a fifth and final year Diagnostic Radiology resident in the American University of Beirut Medical Center, and expected to graduate in June 2020. Future plans include fellowship at Memorial Sloan Kettering Cancer Center, New York. Interests: body and oncologic imaging, neuroradiology, musculoskeletal radiology.