Author/s:
McMaster, C;Bird, A;Liew, DFL;Buchanan, RR;Owen, CE;Chapman, WW;Pires, DEV

Title:
Artificial Intelligence and Deep Learning for Rheumatologists

Date:
2022-12-01

Citation:
McMaster, C., Bird, A., Liew, D. F. L., Buchanan, R. R., Owen, C. E., Chapman, W. W. & Pires, D. E. V. (2022). Artificial Intelligence and Deep Learning for Rheumatologists. Arthritis and Rheumatology, 74 (12), pp.1893-1905. https://doi.org/10.1002/art.42296.

Persistent Link:
https://hdl.handle.net/11343/318269

Artificial Intelligence and Deep Learning for Rheumatologists: A Primer and Review of the Literature

Authors:

Christopher McMaster FRACP MBBS[1,2,3,4]
Alix Bird MBBS[6]
David FL Liew FRACP MBBS[1,2,5]
Russell R Buchanan FRACP MD MBBS[1,5]
Claire E Owen FRACP PhD MBBS[1,5]
Wendy W Chapman PhD[3]
Douglas EV Pires PhD[3,4]

Affiliations

1 Department of Rheumatology, Austin Health, Heidelberg VIC 3084
2 Department of Clinical Pharmacology and Therapeutics, Austin Health, Heidelberg VIC 3084
3 Centre for Digital Transformation of Health, University of Melbourne, Parkville VIC 3052
4 School of Computing and Information Systems, University of Melbourne, Parkville VIC 3052
5 Department of Medicine, University of Melbourne, Parkville VIC 3052
6 Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA 5000

ORCID
CM: 0000-0003-2432-5451
AJB: 0000-0002-8676-0822
DFLL: 0000-0001-8451-8883
RRB:
CEO: 0000-0002-2694-5411
WWC: 0000-0001-8702-4483
DEVP: 0000-0002-3004-2119

Address for correspondence:
Dr Christopher McMaster, Department of Rheumatology, Austin Health, 145 Studley Road, Heidelberg, VIC 3084
Email: christopher.mcmaster@austin.org.au

Disclosures:
Wendy W Chapman reports being a scientific adviser for Health Fidelity and Modern Trials. Alix J Bird reports funding from GlaxoSmithKline Australia

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/art.42296

Abstract

Deep Learning has emerged as the leading method in machine learning, spawning a rapidly-growing field of academic research and commercial applications across medicine, and could have particular relevance to rheumatology if utilized correctly. The greatest benefits of deep learning methods are seen with the unstructured data frequently found in rheumatology, such as images and text, where traditional machine learning methods have struggled to unlock the trove of information held within these data formats. The basis for this success comes from the ability of deep learning to learn the structure of the underlying data. It is no surprise that the first areas of medicine that have started to experience impact from deep learning rely heavily on interpreting visual data, such as triaging radiology workflows and computer-assisted colonoscopy. Applications in rheumatology are beginning to emerge, with recent successes in areas as diverse as detecting joint erosions on plain radiography, predicting future rheumatoid arthritis disease activity and identifying the halo sign on temporal artery ultrasound. Given the important role deep learning methods are likely to play in the future of rheumatology, it is imperative that rheumatologists appreciate the methods and assumptions that underlie the deep learning algorithms in widespread use today, their limitations and the landscape of deep learning research that will inform algorithm development and clinical decision support tools of the future. The best applications of deep learning in rheumatology must be informed by the clinical experience of rheumatologists, so that algorithms can be developed to tackle the most relevant clinical problems.

Introduction

Deep Learning refers to a group of algorithms that use artificial neural networks and an optimization algorithm called backpropagation (with gradient descent) to model complex problems by learning complex functions that describe them (see figure 1) (1). Whilst deep learning methods have been designed and applied for many decades, it is only in the last 10 years that computer hardware has been able to train these increasingly complex models to such a level that they now dominate the machine learning landscape, both in terms of publications and performance. In recent years, the applications of deep learning in medicine have not only gained prominence, but have begun entering clinical practice. At the time of writing, the FDA had approved 56 machine learning algorithms to support radiology practice (2), many of which use deep learning approaches. Deep learning methods power computers that beat grandmasters in Chess and Go (3), summarize documents as diverse as patents and academic papers (4), control autonomous cars (5), and can predict and design macromolecules (6). Although still in its infancy, applications of deep learning to rheumatology are increasing across a broad range of areas (see table 1). There are several ways to classify these various applications, however a natural way to categorize deep learning algorithms is based on the input data type. Time-series data is used for prediction tasks, written text is used for natural language processing and images are used for computer vision. Here we will explore rheumatological applications of deep learning across these three categories.

Problem	Data type	Model	Implications
Identifying giant cell arteritis features from temporal artery biopsy reports (7)	Text	Transformer	Accurate auditing of temporal artery biopsy reports can be performed using deep learning, although this performance drops when tested across centers.
Classification of Hep-2 cells based on ANA-IIF patterns (8)	Images	CNN	Automated ANA classification based on Hep-2 cells is approaching expert human performance.
EULAR-OMERACT Synovitis Scoring (OESS) from synovial ultrasound (9)	Images	CNN	Deep learning can identify synovitis on ultrasound with a high degree of accuracy.
Sharp/van der Heijde scoring using hand & foot radiographs (10)	Images	CNN	Radiographic scoring for rheumatoid arthritis is improving, but still requires work for clinical implementation.
Predicting progression (any increase in KL score) of knee OA based on baseline knee radiograph plus other clinical features (11)	Images	CNN	Radiographic progression in knee OA can be predicted with a combination of clinical features and baseline x-ray using deep learning, however there are definitely unmeasured factors we are missing in these models.
Identifying 'halo sign' on temporal artery ultrasound images (12)	Images	CNN	Deep learning has significant potential for automated identification of the halo sign, however ensuring standardized image acquisition is a major barrier to implementation.
Predicting future rheumatoid arthritis disease activity (controlled vs. uncontrolled) using clinical data from previous encounters (13)	Electronic Health Records	RNN	Deep learning can predict future disease activity from past disease activity and baseline factors, however performance drops significantly when this model is tested at a second center, suggesting that there is substantial heterogeneity between centers that must be accounted for in future models.

Table 1: Current applications of deep learning in rheumatology.
CNN: Convolutional Neural Network; RNN: Recurrent Neural Network; ANA: Antinuclear Antibodies; JSN: Joint Space Narrowing

Learning from Text

Natural Language Processing (NLP) is an interdisciplinary field of study, with the main aim of having computers "perform useful tasks involving human language" (14). Traditionally, NLP has relied heavily on linguistic models of syntax and grammar, complemented by statistical analysis. In contrast, deep learning approaches in contemporary NLP rely less on assumptions about rules of natural language, including expertly-curated words and phrases, building models capable of inferring those rules by learning from large corpuses of text (15). Most recently, deep learning models for NLP have moved towards attention-based models, in particular a group of models collectively referred to as *transformers* (16). Whilst previous state-of-the-art NLP algorithms relied on modeling text as a sequence of words read one-by-one directionally (left-to-right for English text), transformers allow the algorithm to view all the text at once and *pick out* the important words that provide context (see figure 2). Attention-based models, combined with pre-training on very large corpuses of text (see section on transfer learning), have allowed deep learning NLP algorithms to achieve state-of-the-art performance on many language tasks.

Classifying Temporal Artery Biopsy Reports

Classical NLP techniques have been used for identifying patients with rheumatic diseases from the electronic health record with success (18–20). These methods generally rely on the cultivation of a set of words and phrases strongly associated with the disease of interest. Given the wide phenotypic variability of rheumatic diseases, accurate identification and classification of patients based on clinical notes lends itself to deep learning techniques. Presently, deep learning on text has only been applied in a single conference abstract, using transformer models to classify temporal artery biopsy reports based on the presence of three histopathological features (adventitial inflammation, giant cells, intimal hyperplasia) and overall conclusion (giant cell arteritis or not) (7). This study used a model called DistilBERT (21), training on 161 biopsy reports from one center and testing both within that center and on 220 biopsy reports from a second center. The authors achieved excellent performance, particularly on the report conclusion (Area Under Receiver Operating Characteristic Curve, AUC 0.99) and the presence of giant cells (AUC 0.99), with performance reducing slightly in the second center (AUC 0.93 and 0.97, respectively). Pooling data between the two centers and training a new model resulted in significant improvements (AUC 0.99 and 0.99, respectively), when tested on reports across both sites, suggesting that a diverse dataset drawing on the language of multiple institutions will result in more generalisable models. It remains to be seen whether this model can generalize beyond two centers, particularly given both were within the same city in Australia – it is likely that variability in documentation practices, vocabulary and idiomatic expressions will result in reduced performance.

Learning from Electronic Health Records (EHRs)

Deep learning algorithms for predicting future events are varied in design, but often rely on the use of time-series data modeled as sequences. In this respect, predictive algorithms often use similar architectures to those used when learning from texts, which are also modeled as sequences. Many EHR prediction algorithms have been developed, most notably and commonly for inpatient outcomes such as length of stay and inpatient mortality (22). The nature of EHR data produces unique

challenges, reflecting the bias of clinical decisions as much as patient physiology. Which data points are missing and the presence of noise often reflects systematic decision making, rather than a random process (*e.g.,* the absence of invasive blood pressure data in a critically ill patient may reflect a decision about the goals of care, rather than the lack of critical illness). Additionally, care must be taken to ensure dataset leakage(23) doesn't occur, where the training dataset is inadvertently informed by the testing dataset (e.g. the same patient appears in both, but for different inpatient encounters).

Predicting Future Diagnoses from EHRs

The Transformer architecture that has been successfully applied to natural language processing has found similar success in sequence models. Such models may use time-sequence data to predict future events, drawing on the power of self-attention, efficiently learning to recognise long-range relationships between past events and future events (16). Li *et al.* developed a Transformer model trained on 1.6 million primary care patients with at least 5 EHR encounters (24). For each individual, they created a sequence of EHR encounters, with the diagnoses and patient age at the time of the encounter forming the components of the sequence. The authors assessed predictive performance for a number of diseases, including rheumatic diseases. Most notably, the model was able to predict the future development of polymyalgia rheumatica (PMR) with very high accuracy (AUC 0.96). This result might partially reflect data quality issues, with PMR diagnosis in primary care frequently occurring without exclusion of alternative diagnoses(25), thus the model is likely predicting the onset of a polymyalgic syndrome, rather than the specific diagnosis of polymyalgia rheumatica. Additionally, this algorithm may simply be identifying a pattern of clinical coding, rather than a sequence of clinical events - any model that uses clinical interpretation rather than patient physiology is prone to modeling not just patient outcomes, but also physician behavior.

Predicting Disease Activity from EHRs

Rheumatology, as a specialty dealing with chronic, relapsing-remitting disease, maintains a strong interest in the task of predicting future disease activity. Frequent outpatient follow-up with clinical and laboratory assessment is used to detect changes in disease state and allow for interventions to treat any disease relapse or deterioration, yet predicting precisely who will experience relapse or deterioration remains a difficult task. Apart from the Transformer architecture mentioned above, for a long time deep learning has approached the analysis of sequence data and prediction using recurrent neural networks (RNN).

Norgeot *et al.* (13) developed a model to predict whether rheumatoid arthritis patients would have controlled or uncontrolled disease at their next clinic visit, based on data from the EHR. Their definition of disease control was based on a threshold cut-off of the Clinical Disease Activity Index (CDAI) (26). The input data included baseline measures (demographics, rheumatoid factor, CCP antibody) and time-dependent variables (laboratory values, medication and CDAI) from each visit. The time-dependent variables were used to train a recurrent neural network (RNN) — designed to identify periodicity and trend — to account for long-range influences that might affect future disease states. The model was trained and tested on data from one hospital, before being further assessed on data

from a different hospital. Predictably, the performance on the second hospital was inferior, however they were able to improve the performance with a small amount of training on data from the new hospital, in a process known as transfer learning.

Learning from Images

Computer vision is a field of image processing interested in automating tasks of visual perception. Since the ground-breaking AlexNet architecture won the ImageNet competition in 2012 (27), computer vision has been dominated by deep learning algorithms. The basis for its success to date has been one specific neural network architecture: the convolutional neural network (CNN). The CNN has had several inventors with slight variations, however the precursor to modern CNN models is most frequently attributed to a 1999 paper by LeCun *et al.* on object detection (28).

The building block of CNNs is a convolution kernel, a grid or matrix of numbers. The convolution passes over an image (a grid of numbers representing the individual pixels of an image), transforming the image in particular ways. Hard-coded convolutions, like the one in figure 3, might perform tasks like vertical edge detection. Whilst convolutions have existed as an image processing technique for many decades, the innovation in deep learning is that the convolutions are not hard-coded, they are learnt. The layering of many convolutions allows a model to build progressively complex features. Whilst early layers might learn convolutions for simple tasks like edge detection, later layers may join different edge detectors together to make object detectors, face detectors and eventually solve complex tasks like facial expression detection.

HEp-2 Image Classification

Testing for antinuclear antibodies (ANA) using indirect immunofluorescence (IIF) assays has been a cornerstone of the diagnostic workup for systemic autoimmunity for many decades (30). These methods rely on the visual inspection of HEp-2 cells, mixed with fluorescently-labeled antibodies from patient sera. Because different antigens are distributed differently within the HEp-2 cells, the staining pattern produced by the fluorescently-labeled antibodies can provide important diagnostic information about the antigen target and associated disease (31).

Automated analysis of HEp-2 images has arisen as a field of research interest, motivated by concerns that visual assessment of IIF patterns is subjective and time-consuming (32). Recently, deep learning techniques have been applied to this task, with increasing success. Broadly, these techniques attempt to classify either individual HEp-2 cells or specimens (containing many cells), by applying deep learning to coarsely-labeled images. Rahman *et al.* reviewed the application of deep learning techniques for these tasks, identifying 24 published methods for the classification of individual HEp-2 cells and 7 methods for the classification of specimens (33).

A wide variety of deep learning techniques have been applied to cell classification. Broadly, it has been applied in two ways to this task: 1) automatically extract features and classify cells; 2) automatically extract features, which are then passed to an alternative model for classification. State-of-the-art models using either technique have achieved accuracies exceeding 97% (33,34), which compares favorably to human accuracy (73.3%) (35). However, this comparison has been criticized, as the task of classifying a single HEp-2 cell, abstracted from the broader context of the specimen, is

not representative of how ANA-IIF is performed in real clinical practice (8). Moreover, these methods are developed, tested and validated on limited datasets (36–39). These datasets lack consistency, both in terms of whether the images contain single cells or specimens, and the number of different staining categories classified.

In response to these data issues, Wu *et al.* (8) curated a large dataset of 51,694 Hep-2 cell slides that more closely reflect real-world practice. These slides contain multiple cells per image, with up to 4 different patterns present in a single image. They tested multiple CNN architectures, ultimately finding that the InceptionResNetV2 (40) architecture had the best performance. On their testing dataset, they found interobserver agreement — as measured by Cohen's kappa (41) — was similar between expert readers (0.85) and between expert readers and the final model (0.84).

Synovial Ultrasound

Synovial ultrasound is an important and emerging technology in the diagnosis (42) and assessment (43) of inflammatory arthritis. Recently, the EULAR-OMERACT ultrasound taskforce has developed a scoring system, sometimes referred to as the EULAR-OMERACT Synovitis Scoring (OESS) system, designed to be used as an outcome measure in clinical trials (43). Whilst there is ongoing work to validate and refine its use (44), the standardized application of an ultrasound synovitis scoring system is amenable to deep learning methods. Andersen *et al.* (45) first applied two well-studied CNN architectures to this problem (VGG-16 (46) and Inception-v3 (47)), after first performing pre-training on the popular ImageNet dataset (48). They found overall good performance (AUC 0.93) on the task of discriminating healthy (OESS score 0-1) from unhealthy (OESS 2-3) joints, however precise scoring on the ordinal scale did not appear to match human performance (49). A follow-up method from the same group, this time using a so-called "cascade" CNN (figure 4a), showed similar performance to expert rheumatologists (9). In this algorithm, a CNN is given the task of classifying a power doppler image as either being at or above a given OESS grade. Each image that is classified as being of a higher grade is then passed to the next CNN, which performs an identical task for the next highest grade. These algorithms may be further enhanced with larger, multi-centre datasets and further refinement of the CNN architecture. As in many medical applications, the laborious task of manually labeling images could be augmented by semi-supervised learning and potentially synthetic data from methods such as generative adversarial networks (50).

Joint Damage in Inflammatory Arthritis

Progressive joint erosion early in the course of rheumatoid arthritis is one of the major predictors of future physical function (52). The prevention of progressive joint erosions is an important outcome measure in establishing the efficacy of any disease modifying agent in rheumatoid arthritis and other inflammatory arthritides. It is therefore important that erosive disease is measured consistently and with high sensitivity. The van der Heijde modified Total Sharp Score (mTSS) (53) is commonly used in clinical trials to assess for progressive joint erosions, consisting of both an erosion score (ES) and joint space narrowing score (JSN). The process of grading these components is a visual task performed by trained radiologists and rheumatologists, making it a good application for neural network models.

Hirano *et al.* developed a neural network model to grade hand joints according to the mTSS using hand radiographs (54). To perform joint-level scoring, their model first had to perform a task called image segmentation to create bounding boxes around individual joints so they could be assessed. Rather than employing a deep learning model for this task, such as the popular U-Net model (figure 4b) first developed for biomedical image segmentation (51), the authors used hard-coded convolutions known as Haar-like features, in a method first described in 2001 by Viola *et al.* (55). After image segmentation, the authors built JSN and ES models with two convolutional layers and three fully-connected layers. The JSN and ES models had similar performance to clinician assessment (correlations coefficients 0.72-0.88 and 0.54-0.75, respectively), however overall sensitivity to detect erosions was low (34.8–42.4%). The major limitation for this algorithm was a relatively small dataset (186 radiographs). With larger datasets, deeper models with more sophisticated architectures will perhaps make this a clinically-applicable scenario for deep learning.

Recently, deep learning methods using a combination of CNNs and attention mechanisms (10) have been used for mTSS scoring in rheumatoid arthritis. The authors also used a two-stage approach, first using a CNN architecture called RetinaNet (56) to detect joint groups, followed by another CNN architecture called EfficientNet (57) to score individual joints. Additionally, the authors applied an attention layer (figure 2) after the convolutional layers. The attention layer effectively constrains the area of interest to only those pixels in the image that provide a substantial contribution to the final prediction – the *attention* to all other pixels becomes negligible. By interrogating this attention layer, the authors produced heatmaps to demonstrate which regions contribute to the scoring of a joint, although interpreting these as an explanation of how the model works should be treated with great caution (see section on explainability below) (58,59).

Plain Films in Knee Osteoarthritis

Current EULAR recommendations for the diagnosis of osteoarthritis (OA) support imaging as a diagnostic tool in atypical presentations of suspected osteoarthritis only (60), however this recommendation is not supported by high-level evidence and the role of routine imaging remains a topic of debate (61). Outside of diagnosis, there is also debate about the prognostic role of imaging features to predict symptom severity and progression. The current EULAR recommendations do not support the use of imaging for prognostication, however this is on the basis of older studies using hand-crafted imaging features, not the systematic discovery of prognostic features from deep learning algorithms.

Tiulpin *et al.* developed deep learning models for diagnosis and prognosis of knee osteoarthritis from plain radiographs (11,62–64). For diagnosis, the authors made use of CNN architectures to train two models to grade images according to OA severity. In their first model, they used a Siamese Network to grade knee radiographs according to the Kellgren-Lawrence composite score (KL) (65), a global rating system for knee osteoarthritis on a 0 – 4 scale. The Siamese Network, as first proposed by Baldi and Chauvin (66), trains two identical neural networks simultaneously, with the task of determining whether two images meet some similarity threshold – in the original paper, the models compare two

fingerprints to determine whether they come from the same finger. In the present study, the input image pairs are automatically segmented from the original plain radiographs, consisting of the right half of the tibiofemoral joint and the horizontally-flipped left half. Because the tibiofemoral joint has horizontal symmetry with respect to the features that predict KL score (figure 4c), a single model can identify salient features from both sides (provided one half is horizontally flipped). These left- and right-sided predictions are then joined to give an overall KL score. Overall, they found good agreement between model and clinician scores, with a Cohen's kappa of 0.83.

The second diagnostic model from Tiulpin *et al.* used a conventional ResNet architecture with transfer learning from ImageNet to grade both individual OA features using the Osteoarthritis Research Society International (OARSI) atlas of OA radiographic features (67) and KL score. The authors were able to achieve state-of-the-art results on OARSI scoring, exceeding human accuracy on this task.

For the task of prognostication, the authors trained a CNN on baseline knee radiographs to predict whether a repeat radiograph would show an increase in KL score (11). They compared this model to a model that used tabular data: age, sex, Body Mass Index, KL grade, Western Ontario and McMaster Universities Arthritis Index, injury and surgery history. Additionally, they combined these two models to test whether there was any additional benefit from a so-called "multi-modal" approach. Their CNN model outperformed the tabular data model, whilst the multi-modal approach outperformed the individual models, demonstrating that knee radiographs contain prognostic features not present in structured patient data, even reported KL grade. Despite accurate radiographic scoring, applicability will ultimately be limited given the poor correlation between radiographic scoring and clinical outcomes such as pain scores(68). Predicting progressive disease may only be helpful if the definition of progressive disease is a clinical, rather than radiographic outcome.

Temporal Artery Ultrasound

The diagnosis of Giant Cell Arteritis (GCA) is a vexing problem for rheumatologists, in no small part due to the lack of an accurate non-invasive diagnostic test. Temporal artery ultrasound is one emerging solution to this problem, with EULAR guidelines recommending ultrasound as the first-line imaging modality for the workup of suspected GCA (69). Despite its relatively good performance as a diagnostic test, temporal artery ultrasound suffers from only moderate inter-rater agreement, with significant training requirements that pose a barrier to the widespread adoption of this test (70). Deep learning therefore seems attractive as a tool to reduce the variability in interpretation and perhaps even lower the barrier to the adoption of this test.

Roncato *et al.* (12) developed a CNN model, specifically to identify one common feature in positive temporal artery ultrasounds: the halo sign (71). The first step in their algorithm was to perform semantic segmentation of transverse and longitudinal color or power doppler images of temporal arteries. This process involved drawing boxes around arteries and adjacent tissue, with individual pixels in these boxes labeled as either halo- sign positive or negative. The authors used a U-Net model, designed specifically for biomedical image segmentation. The final classification of each image as either positive or negative was based on the percentage of pixels within the bounding box classified as 'halo sign' positive — a higher percentage of 'halo sign' pixels means a higher probability that the image truly contains a 'halo sign'. The accuracy of the model was compared using two groups of

images: group 1 obtained by a single operator, using a standardized protocol; and group 2 obtained by multiple different operators using a variety of parameters. The performance on group 1 was significantly higher than group 2 (AUC 0.95 vs 0.82). Although training on more non-standard images may improve performance, deep learning can also be used for computer-assisted image acquisition to assist clinicians and sonographers in acquiring standardized views at the time of ultrasound.

The Future

Learning With Limited Data

One of the main features of deep learning is the ability to capitalize on the wealth of large datasets, with improvements in computer vision models seen even beyond massive datasets composed of 300 million images (72). Despite this, there are three established and emerging technical solutions to the problem of learning with limited data.

Transfer Learning

The first method, now well-established in deep learning research and applications, is a concept known as transfer learning. Transfer learning is the process whereby a model is trained on a large dataset (pre-training), and then only partially re-trained on a small dataset, even from a different domain (*e.g.*, a model trained on photographs is re-trained on radiographs). The motivation for this process is that, in training a large dataset, the model will have learnt generalisable properties. These generalisable properties are thought to be learnt in the early layers and so it is only the final layers that are retrained on the new, smaller dataset of interest. This has the effect of "specializing" a pre-trained model for a downstream task. The pre-training process is generally performed using a process called self-supervised learning, which does not require any hand-labeled data, but instead learns by predicting missing words, the next word in a sentence or similar tasks. In certain circumstances, these pre-trained models can be used in a few- or zero-shot setting, meaning they are fine-tuned on few or even no examples of the downstream task. This is particularly the case for NLP, where the task can be distinguished by a text prompt (e.g. a clinical question) – if the pre-trained model can interpret the prompt, then it may not necessarily need to see any examples in order to perform the task.

Self-supervised Learning

Recently, self-supervised learning has emerged as a method to learn from large, unlabelled datasets (73). Self-supervision is achieved by creating tasks that allow models to learn generalisable features. For example, in NLP, learning to predict the next word in a sentence requires learning common linguistic features, such as sentence structure and word meaning. In computer vision, learning to pair original and distorted versions of the same image requires learning invariant features, such as the rounded shape of the metacarpal head. Self-supervised models can then be used in a transfer learning process to train on smaller, labeled datasets. Given the relatively small datasets in rheumatology, it is highly likely that new applications of deep learning will be powered by self-supervised learning and transfer learning.

Increase Dataset Size

Whilst single institutions might generate only relatively small datasets, pooling data across many institutions can result in large datasets. Large datasets may be necessary, particularly where outcomes are rare. In rheumatology, the Rheumatology Informatics System for Effectiveness has begun pooling EHR data to further our understanding of rheumatic diseases (74), however further barriers exist where text and image data are needed. Concerns about data privacy can be a barrier to sharing data across institutional, regional and international borders – and aside from specific data sharing agreements, much effort has been made in producing technical solutions to this problem.

Recently, federated learning has emerged as a powerful tool to allow multiple sites to contribute training data without ever sharing raw data (75). Several federated learning techniques exist, however recent work has focused on methods whereby individual sites train small models on local data, with the coefficients of these models sent to a central site that uses these to train a cross-center model (76,77). Federated learning is not without significant challenges, including: high cost of setting up a central server and communication between sites; black box models, particularly if only model weights are shared and not raw data; and no clear best way to aggregate data from heterogeneous sites (78). In addition to these concerns, federated learning does not fully solve the data privacy problem; all models can "memorize" training data and therefore if privacy is a significant concern, models have to be trained using special techniques called differential privacy to prevent data memorization (79).

When there is not even enough data to pool, or data pooling is impractical, an alternative practice is to generate synthetic data. Generating artificial samples has been an area of intense research in deep learning, particularly after the development of Generative Adversarial Networks (GANs) (50) by Goodfellow et al. in 2014. GANs train two networks simultaneously: a generator network to generate new data and a discriminator network to discriminate real from synthetic data. Although these models can be difficult to train, as it becomes harder and harder for the discriminator to distinguish between real and synthetic data, the quality of synthetic data increases.

Data Integrity, Bias and Ethics

Although deep learning has a long history in computer science research, it has only been the relatively recent development of computers capable of training these algorithms that has resulted in an explosion of applications in medicine. In rheumatology, these methods have only now begun to show promise – although not without potential impediments. As deep learning methods gain more prominence, issues of data integrity and standardization will become more important. As we saw above, the standardized acquisition of temporal artery ultrasound images poses a potential barrier to the effective deployment of deep learning algorithms to diagnose giant cell arteritis. Outside of rheumatology, a similar problem in echocardiography has led to several FDA-approved ultrasound machines that not only interpret images, but guide the user on how to adjust the probe to obtain enhanced views (80). As rheumatologists and machine learning engineers begin applying deep learning techniques to clinical problems, we will need an even greater focus on data integrity and quality. Beyond simply ensuring high quality data, careful collection and curation of datasets is required when considering how data may reflect the systematic biases against marginalized and under-represented minorities in our communities, making resultant algorithms unsafe and inaccurate for many.

New guidelines for the reporting of clinical trials of artificial intelligence interventions provide a guide for the steps required to demonstrate safety and efficacy of these algorithms (81). Like clinical trials of conventional medical interventions, careful trial design is paramount in testing clinical algorithms. Unlike conventional trials, the risk of bias extends beyond the study design and into the algorithm design itself. The dataset(s) used to train deep learning algorithms can introduce substantial bias that may not only invalidate the results, but introduce racial, gender or other forms of discrimination if applied systematically. Promising methods to identify and minimize such bias include report cards for evaluating performance in particular groups (82), however static model checks are not enough in production and the predictions of any model should be examined periodically to ensure they are not introducing or perpetuating unacceptable bias. For more detail on bias in clinical machine learning, Chen et al. (83) provides a detailed overview of how bias is introduced at each step of the algorithm development process, while Gianfrancesco et al. (84) details the sources of bias in EHR-based models.

New technologies like wearable devices, assisted by automatic deep learning algorithms, have the potential to diagnose myriad conditions under circumstances never seen before at such scale (e.g. asymptomatic atrial fibrillation in young people), with the resultant risk of widespread overdiagnosis. In rheumatology, for example, the unsolved problem of treatments for clinically suspect arthralgia may be compounded should automated diagnostic tools become available, identifying subclinical reductions in morning mobility leading to a tidal wave of very early arthritis diagnoses. Whilst coordinated data collection will be needed to quantify the risks associated with these new diagnostic paradigms (85), this also presents a new opportunity to further our understanding of rheumatic diseases by studying longitudinal cohorts from very early disease stages. Compounding this, access to these devices is contingent on affordability (86) – one way in which technologically-enhanced medicine widens the socioeconomic disparities in the provision of healthcare.

Explainability in Rheumatology

What is explainability

Neural networks are often described as black boxes, in that their internal processes are inscrutable to humans. Many have argued that we need ways to explain why a model reaches a decision so that doctors are able to interpret it and apply it clinically (87). The most common methods for explainability in conventional machine learning – in particular SHAP and LIME (88) – are broadly classified as post-hoc perturbation methods. These algorithms perturb the input data and measure how these perturbations in the input alter the model output. Although post-hoc explainability methods have shortcomings, including susceptibility to hide model bias (88), even inadequate model explanations at least provide interpretable outputs concerning perspicuous variables (e.g. how does cardiovascular risk change if the patient has hypertension). Compared to conventional machine learning methods, explainability methods for deep learning are more difficult to interpret. In medical imaging, the most common method for explainability are saliency maps, which show visually the parts of the input image that most contributed to the final prediction (89).

Strengths

Explainability methods may have a role in evaluating models. As previously discussed, a 2019 paper predicted future clinical disease activity index (CDAI) and used permutation importance scores to determine the feature importance. They concluded that disease activity, lab values and medications were the best predictors of future CDAI (13). With simple input data, this method provides useful information regarding how the model operates, but when dealing with complex data (*e.g.,* text, images) interpretation is unclear.

Weaknesses

In a model diagnosing knee OA, saliency maps were used to show that the model was identifying relevant radiological features (63). They found that osteophytes were highlighted and concluded that attention maps would “build better trust in the clinical community”. However, they also acknowledge that the reasons these anatomically relevant areas were highlighted was because they constrained the model to only assess these regions. Additionally, a model developed to predict RA radiographic scores used saliency maps to determine which joints were predictive of the scores (10). While these images show some focus over bone and joint space, they also highlight parts of the image that are empty.

In both of these cases the meaning of these explanations is unclear. In the first, it was inevitable that these regions would be highlighted, while in the second the model uses areas that have no anatomical relevance in its prediction. Both papers use this as evidence that their model is performing reliably. Even more problematically, it has been shown that in models where the input image is modified to result in an incorrect prediction, the saliency map can still highlight clinically relevant regions of the image (90). This is falsely reassuring that the model is behaving appropriately, while still providing the wrong answer. A recent paper has highlighted these concerns about the danger of relying on such methods to engender trust in a system, and suggests we ought to depend on rigorous evaluation instead (58). Whether or not saliency maps produce sensible explanations, they should not be what we rely on to trust the behavior of neural networks.

Evaluation

Although explainability techniques can tell us something about model behavior, it can't tell us how to interpret the predictions. For safe implementation testing on external datasets is needed to ensure models generalize to different populations, evaluate performance in subgroups to eliminate biases (91) and ultimately test models with randomized control trials. It is vital to understand how models affect patient outcomes in clinical settings and it is this, not explainability techniques, that rheumatologists should be demanding before a model reaches real world implementation.

Translation into Practice

Deep learning is transforming many industries and at an ever-increasing pace. For example, Google’s language translation (92), Uber’s expected arrival time (ETA) prediction (93) and Microsoft’s code completion tool (94) are all deep learning algorithms that many people rely on daily. Uber switched their ETA algorithm to deep learning because of its ability to easily scale up with larger datasets and

larger models. But the same value proposition is not always clear in medicine, where the scale may not be so large and financial barriers, such as the cost of regulatory approval, may be a substantial impediment. The barriers to widespread adoption are necessarily set high, with minimum standards of safety and efficacy set not only by the regulatory authorities, but also the clinicians who must “buy in” to this new technology. Nevertheless, the deep learning revolution has been largely driven by falling costs in computing power(95), with no clear indication that this will plateau significantly. We can therefore expect further improvements, shifting the value proposition and encouraging even greater investment in research. Who takes advantage of this, whether it be academic or industry, is an open question.

Conclusion

Deep learning is an important method in medical machine learning applications and will likely become the dominant method in the future. Several applications in rheumatology have been published, with the promise of many more to come. Importantly, although deep learning methods offer the opportunity to improve the efficiency of some clinical tasks, they also provide a powerful technique for generating new knowledge and insights, particularly in the previously impenetrable analysis of unstructured data such as text and images. Presently, much of the published work applying deep learning to rheumatology has occurred on small, homogeneous public datasets that do not reflect the diversity of real data, including the interactions between different data modalities (e.g. EHR plus imaging). Further collaboration and interaction between machine learning researchers and rheumatology researchers will likely result in more clinically-applicable algorithms, with translation into clinical practice being the next great hurdle to overcome. Researchers should therefore be familiar with the potential applications and limitations of these methods in their own research, and clinicians should be familiar with some of the potential benefits and pitfalls as these methods make their way into clinical practice.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28;521(7553):436–44.

2. FDA Cleared AI Algorithms [Internet]. [cited 2020 Jun 27]. Available from: https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms

3. Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm [Internet]. arXiv [cs.AI]. 2017. Available from: http://arxiv.org/abs/1712.01815

4. Zhang J, Zhao Y, Saleh M, Liu PJ. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization [Internet]. arXiv [cs.CL]. 2019. Available from: http://arxiv.org/abs/1912.08777

5. Chen K, Oldja R, Smolyanskiy N, Birchfield S, Popov A, Wehr D, et al. MVLidarNet: Real-Time Multi-Class Scene Understanding for Autonomous Driving Using Multiple Views [Internet]. arXiv [cs.CV]. 2020. Available from: http://arxiv.org/abs/2006.05518

6. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.

7. McMaster C, Yang V, Sutu B, Oon S, Ngian G, Wicks I, Buchanan R, Liew D. Temporal Artery Biopsy Reports Can Be Accurately Classified by Artificial Intelligence [abstract]. Arthritis Rheumatol [Internet]. 2021;73(suppl 10). Available from: https://acrabstracts.org/abstract/temporal-artery-biopsy-reports-can-be-accurately-classified-by-artificial-intelligence/

8. Wu Y-D, Sheu R-K, Chung C-W, Wu Y-C, Ou C-C, Hsiao C-W, et al. Application of Supervised Machine Learning to Recognize Competent Level and Mixed Antinuclear Antibody Patterns Based on ICAP International Consensus. Diagnostics (Basel) [Internet]. 2021 Apr 1;11(4). Available from: http://dx.doi.org/10.3390/diagnostics11040642

9. Christensen ABH, Just SA, Andersen JKH, Savarimuthu TR. Applying cascaded convolutional neural network design further enhances automatic scoring of arthritis disease activity on ultrasound images from rheumatoid arthritis patients. Ann Rheum Dis [Internet]. 2020 Jun 5; Available from: http://dx.doi.org/10.1136/annrheumdis-2019-216636

10. Chaturvedi N. DeepRA: Predicting Joint Damage From Radiographs Using CNN with Attention [Internet]. arXiv [cs.CV]. 2021. Available from: http://arxiv.org/abs/2102.06982

11. Tiulpin A, Klein S, Bierma-Zeinstra SMA, Thevenot J, Rahtu E, van Meurs J, et al. Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data. Sci Rep. 2019 Dec 27;9(1):20038.

12. Roncato C, Perez L, Brochet-Guégan A, Allix-Béguec C, Raimbeau A, Gautier G, et al. Colour Doppler ultrasound of temporal arteries for the diagnosis of giant cell arteritis: a multicentre deep learning study. Clin Exp Rheumatol. 2020 Mar;38 Suppl 124(2):120–5.

13. Norgeot B, Glicksberg BS, Trupin L, Lituiev D, Gianfrancesco M, Oskotsky B, et al. Assessment of a Deep Learning Model Based on Electronic Health Record Data to Forecast Clinical Outcomes in Patients With Rheumatoid Arthritis. JAMA Netw Open. 2019 Mar 1;2(3):e190606.

14. Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall; 2009. 988 p.

15. Young T, Hazarika D, Poria S, Cambria E. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. IEEE Comput Intell Mag. 2018 Aug;13(3):55–75.

16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 5998–6008.

17. Vig J. A Multiscale Visualization of Attention in the Transformer Model. arXiv [csHC] [Internet]. 2019 Jun 12; Available from: https://arxiv.org/abs/1906.05714

18. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res . 2010 Aug;62(8):1120–7.

19. Zhao SS, Hong C, Cai T, Xu C, Huang J, Ermann J, et al. Incorporating natural language processing to improve classification of axial spondyloarthritis using electronic health records. Rheumatology . 2020 May 1;59(5):1059–65.

20. Tedeschi SK, Cai T, He Z, Ahuja Y, Hong C, Yates KA, et al. Classifying Pseudogout using Machine Learning Approaches with Electronic Health Record Data. Arthritis Care Res [Internet].

2020 Jan 7; Available from: http://dx.doi.org/10.1002/acr.24132

21. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [Internet]. arXiv [cs.CL]. 2019. Available from: http://arxiv.org/abs/1910.01108

22. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018 May 8;1:18.

23. Elangovan A, He J, Verspoor K. Memorization vs. Generalization: Quantifying Data Leakage in NLP Performance Evaluation [Internet]. arXiv [cs.CL]. 2021. Available from: http://arxiv.org/abs/2102.01818

24. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: Transformer for Electronic Health Records. Sci Rep. 2020 Apr 28;10(1):7155.

25. Helliwell T, Hider SL, Mallen CD. Polymyalgia rheumatica: diagnosis, prescribing, and monitoring in general practice. Br J Gen Pract. 2013 May;63(610):e361-6.

26. Aletaha D, Nell VPK, Stamm T, Uffmann M, Pflugbeil S, Machold K, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. Arthritis Res Ther. 2005 Apr 7;7(4):R796-806.

27. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems 25. Curran Associates, Inc.; 2012. p. 1097–105.

28. LeCun Y, Haffner P, Bottou L, Bengio Y. Object Recognition with Gradient-Based Learning. In: Forsyth DA, Mundy JL, di Gesú V, Cipolla R, editors. Shape, Contour and Grouping in Computer Vision. Berlin, Heidelberg: Springer Berlin Heidelberg; 1999. p. 319–45.

29. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. Madison, WI, USA: Omnipress; 2010. p. 807–14. (ICML'10).

30. Mahler M, Meroni P-L, Bossuyt X, Fritzler MJ. Current concepts and future directions for the assessment of autoantibodies to cellular antigens referred to as anti-nuclear antibodies. J Immunol Res. 2014 Apr 27;2014:315179.

31. Damoiseaux J, Andrade LEC, Carballo OG, Conrad K, Francescantonio PLC, Fritzler MJ, et al. Clinical relevance of HEp-2 indirect immunofluorescent patterns: the International Consensus on ANA patterns (ICAP) perspective. Ann Rheum Dis. 2019 Jul;78(7):879–89.

32. Infantino M, Meacci F, Grossi V, Manfredi M, Benucci M, Merone M, et al. The burden of the variability introduced by the HEp-2 assay kit and the CAD system in ANA indirect immunofluorescence test. Immunol Res. 2017 Feb;65(1):345–54.

33. Rahman S, Wang L, Sun C, Zhou L. Deep Learning based HEp-2 Image Classification: A Comprehensive Review [Internet]. arXiv [cs.CV]. 2019. Available from: http://arxiv.org/abs/1911.08916

34. Vununu C, Lee S-H, Kwon K-R. A Strictly Unsupervised Deep Learning Method for HEp-2 Cell Image Classification. Sensors [Internet]. 2020 May 9;20(9). Available from: http://dx.doi.org/10.3390/s20092717

35. Foggia P, Percannella G, Soda P, Vento M. Benchmarking HEp-2 cells classification methods.

IEEE Trans Med Imaging. 2013 Oct;32(10):1878–89.

36. AIDA Project | un progetto transfrontaliero Italia-Tunisia per l'applicazione di ICT alla diagnosi delle malattie autoimmuni - Downloads [Internet]. [cited 2020 Jun 10]. Available from: http://www.aidaproject.net/index.php/it/downloads

37. DOWNLOAD – HEp-2 contest @ ICPR2016 [Internet]. [cited 2020 Jun 10]. Available from: https://hep2.unisa.it/dbtools.html

38. Mivia. ICPR 2012 – Contest on HEp-2 Cells Classification – MIVIA [Internet]. [cited 2020 Jun 10]. Available from: https://mivia.unisa.it/contest-hep-2/

39. Classification of Human Epithelial Type 2 Cell Indirect Immunofluoresence Images via Codebook Based Descriptors [Internet]. [cited 2020 Jun 10]. Available from: http://staff.itee.uq.edu.au/lovell/snphep2/

40. Kassani SH, Kassani PH, Khazaeinezhad R, Wesolowski MJ, Schneider KA, Deters R. Diabetic Retinopathy Classification Using a Modified Xception Architecture. In: 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). 2019. p. 1–6.

41. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas. 1960 Apr 1;20(1):37–46.

42. Kaeley GS, Bakewell C, Deodhar A. The importance of ultrasound in identifying and differentiating patients with early inflammatory arthritis: a narrative review. Arthritis Res Ther. 2020 Jan 2;22(1):1.

43. D'Agostino M-A, Terslev L, Aegerter P, Backhaus M, Balint P, Bruyn GA, et al. Scoring ultrasound synovitis in rheumatoid arthritis: a EULAR-OMERACT ultrasound taskforce-Part 1: definition and development of a standardised, consensus-based scoring system. RMD Open. 2017 Jul 11;3(1):e000428.

44. Terslev L, Christensen R, Aga A-B, Sexton J, Haavardsholm EA, Hammer HB. Assessing synovitis in the hands in patients with rheumatoid arthritis by ultrasound: an agreement study exploring the most inflammatory active side from two Norwegian trials. Arthritis Res Ther. 2019 Jul 5;21(1):166.

45. Andersen JKH, Pedersen JS, Laursen MS, Holtz K, Grauslund J, Savarimuthu TR, et al. Neural networks for automatic scoring of arthritis disease activity on ultrasound images. RMD Open. 2019 Mar 30;5(1):e000891.

46. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [Internet]. arXiv [cs.CV]. 2014. Available from: http://arxiv.org/abs/1409.1556

47. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision [Internet]. arXiv [cs.CV]. 2015. Available from: http://arxiv.org/abs/1512.00567

48. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 2015 Dec 1;115(3):211–52.

49. Terslev L, Naredo E, Aegerter P, Wakefield RJ, Backhaus M, Balint P, et al. Scoring ultrasound synovitis in rheumatoid arthritis: a EULAR-OMERACT ultrasound taskforce-Part 2: reliability and application to multiple joints of a standardised consensus-based scoring system. RMD Open. 2017 Jul 11;3(1):e000427.

50. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks [Internet]. arXiv [stat.ML]. 2014. Available from: http://arxiv.org/abs/1406.2661

51. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing; 2015. p. 234–41.

52. Odegard S, Landewe R. van der HD, Kvien TK, Mowinckel P, Uhlig T. Association of early radiographic damage with impaired physical function in rheumatoid arthritis: a ten-year, longitudinal observational study in 238 patients. Arthritis Rheum. 2006;54:68–75.

53. van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol. 2000 Jan;27(1):261–3.

54. Hirano T, Nishide M, Nonaka N, Seita J, Ebina K, Sakurada K, et al. Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis. Rheumatol Adv Pract. 2019 Nov 22;3(2):rkz047.

55. Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2001. 2001. p. I–I.

56. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection [Internet]. arXiv [cs.CV]. 2017. Available from: http://arxiv.org/abs/1708.02002

57. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks [Internet]. arXiv [cs.LG]. 2019. Available from: http://arxiv.org/abs/1905.11946

58. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021 Nov;3(11):e745–50.

59. Lipton ZC. The Mythos of Model Interpretability [Internet]. arXiv [cs.LG]. 2016. Available from: http://arxiv.org/abs/1606.03490

60. Sakellariou G, Conaghan PG, Zhang W, Bijlsma JWJ, Boyesen P, D'Agostino MA, et al. EULAR recommendations for the use of imaging in the clinical management of peripheral joint osteoarthritis. Ann Rheum Dis. 2017 Sep;76(9):1484–94.

61. Wang X, Oo WM, Linklater JM. What is the role of imaging in the clinical diagnosis of osteoarthritis and disease management? Rheumatology . 2018 May 1;57(suppl_4):iv51–60.

62. Tiulpin A, Saarakkala S. Automatic Grading of Individual Knee Osteoarthritis Features in Plain Radiographs using Deep Convolutional Neural Networks [Internet]. arXiv [eess.IV]. 2019. Available from: http://arxiv.org/abs/1907.08020

63. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. Sci Rep. 2018 Jan 29;8(1):1727.

64. Tiulpin A, Klein S, Bierma-Zeinstra S, Thevenot J, van Meurs J, Oei E, et al. Deep learning predicts knee osteoarthritis progression from plain radiographs. Osteoarthritis Cartilage. 2019 Apr 1;27:S397–8.

65. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis. 1957

Dec;16(4):494–502.

66. Baldi P, Chauvin Y. Neural Networks for Fingerprint Recognition. Neural Comput. 1993 May 1;5(3):402–18.

67. Altman RD, Gold GE. Atlas of individual radiographic features in osteoarthritis, revised. Osteoarthritis Cartilage. 2007;15 Suppl A:A1-56.

68. Finan PH, Buenaver LF, Bounds SC, Hussain S, Park RJ, Haque UJ, et al. Discordance between pain and radiographic severity in knee osteoarthritis: findings from quantitative sensory testing of central sensitization. Arthritis Rheum. 2013 Feb;65(2):363–72.

69. Dejaco C, Ramiro S, Duftner C, Besson FL, Bley TA, Blockmans D, et al. EULAR recommendations for the use of imaging in large vessel vasculitis in clinical practice. Ann Rheum Dis. 2018 May;77(5):636–43.

70. Luqmani R, Lee E, Singh S, Gillett M, Schmidt WA, Bradburn M, et al. The Role of Ultrasound Compared to Biopsy of Temporal Arteries in the Diagnosis and Treatment of Giant Cell Arteritis (TABUL): a diagnostic accuracy and cost-effectiveness study. Health Technol Assess. 2016 Nov;20(90):1–238.

71. Schmidt WA, Kraft HE, Völker L, Vorpahl K, Gromnica-Ihle EJ. Colour Doppler sonography to diagnose temporal arteritis. Lancet. 1995 Apr 1;345(8953):866.

72. Zhai X, Kolesnikov A, Houlsby N, Beyer L. Scaling Vision Transformers [Internet]. arXiv [cs.CV]. 2021. Available from: http://arxiv.org/abs/2106.04560

73. Zbontar J, Jing L, Misra I, LeCun Y, Deny S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. arXiv [csCV] [Internet]. 2021 Mar 4; Available from: https://arxiv.org/abs/2103.03230

74. Yazdany J, Bansback N, Clowse M, Collier D, Law K, Liao KP, et al. Rheumatology Informatics System for Effectiveness: A National Informatics-Enabled Registry for Quality Improvement. Arthritis Care Res . 2016 Dec;68(12):1866–73.

75. Yang Q, Liu Y, Chen T, Tong Y. Federated Machine Learning: Concept and Applications. ACM Trans Intell Syst Technol. 2019 Jan 28;10(2):1–19.

76. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med. 2021 Oct;27(10):1735–43.

77. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA y. Communication-Efficient Learning of Deep Networks from Decentralized Data. Singh A, Zhu J, editors. 20--22 Apr 2017;54:1273–82.

78. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020 Sep 14;3:119.

79. Abadi M, Chu A, Goodfellow I, Brendan McMahan H, Mironov I, Talwar K, et al. Deep Learning with Differential Privacy [Internet]. arXiv [stat.ML]. 2016. Available from: http://arxiv.org/abs/1607.00133

80. Office of the Commissioner. FDA Authorizes Marketing of First Cardiac Ultrasound Software That Uses Artificial Intelligence to Guide User [Internet]. U.S. Food and Drug Administration. 2020 [cited 2020 Jun 12]. Available from: https://www.fda.gov/news-events/press-announcements/fda-authorizes-marketing-first-cardiac-ultrasound-software-uses-artificial-intelligence-guide-user

81. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020 Sep;26(9):1364–74.

82. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model Cards for Model Reporting [Internet]. arXiv [cs.LG]. 2018. Available from: http://arxiv.org/abs/1810.03993

83. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical Machine Learning in Health Care [Internet]. arXiv [cs.CY]. 2020. Available from: http://arxiv.org/abs/2009.10576

84. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern Med. 2018 Nov 1;178(11):1544–7.

85. Capurro D, Coghlan S, Pires DEV. Preventing Digital Overdiagnosis. JAMA. 2022 Feb 8;327(6):525–6.

86. Smuck M, Odonkor CA, Wilt JK, Schmidt N, Swiernik MA. The emerging clinical role of wearables: factors for successful implementation in healthcare. NPJ Digit Med. 2021 Mar 10;4(1):45.

87. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? [Internet]. arXiv [cs.AI]. 2017. Available from: http://arxiv.org/abs/1712.09923

88. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery; 2020. p. 180–6.

89. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps [Internet]. arXiv [cs.CV]. 2013. Available from: http://arxiv.org/abs/1312.6034

90. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity Checks for Saliency Maps [Internet]. arXiv [cs.CV]. 2018. Available from: http://arxiv.org/abs/1810.03292

91. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. Proc ACM Conf Health Inference Learn (2020). 2020 Apr;2020:151–9.

92. Caswell I, Liang B. Recent Advances in Google Translate [Internet]. Google AI Blog. [cited 2022 Apr 28]. Available from: https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html

93. Hu X, Cirit O, Binaykiya T, Hora R. DeepETA: How Uber Predicts Arrival Times Using Deep Learning [Internet]. Uber Engineering. [cited 2022 Apr 28]. Available from: https://eng.uber.com/deepeta-how-uber-predicts-arrival-times/

94. Chen M, Tworek J, Jun H, Yuan Q, de Oliveira Pinto HP, Kaplan J, et al. Evaluating Large Language Models Trained on Code [Internet]. arXiv [cs.LG]. 2021. Available from: http://arxiv.org/abs/2107.03374

95. Slominski A, Muthusamy V, Ishakian V. Future of Computing is Boring (and that is exciting!) or How to get to Computing Nirvana in 20 years or less [Internet]. arXiv [cs.CY]. 2019. Available from: http://arxiv.org/abs/1906.10398

Figure 1: Neural network architectures. The first layer of a neural network consists of the data. This data is then passed to the first "hidden layer". Each node, represented by a circle, is a weighted linear combination of all the nodes in the layer before. It is the weights that the model "learns". Apart from a classic neural network where all nodes from one layer are connected to the next (otherwise known as a multi-layer perceptron), other common architectures include recurrent networks with connections between nodes within a layer, usually used for sequence data (*e.g.,* time-series or text); and residual networks, where information from one layer can "skip" the next layer, giving the network a way to bypass inefficient layers.

Figure 2: Attention model visualized (17). At the top are two attention layers with text input for NLP. The original input text reads: "There was swelling and redness of the joint. The joint was also stiff and tender, with reduced range of motion." This text is converted into tokens, sometimes splitting words into more than one token (here "redness" is split into "red" and "##ness" - the "##" signifying that this token belongs with the preceding token. On the left, a lower layer of the attention-based model has relied on the words "range" and "motion" to interpret the word "reduced". On the right, at a higher layer, the word "reduced" also depends strongly on the word "swelling" in the previous sentence.
Attention can be used for any sequence data, as seen in the bottom panel. Here these numbers could be laboratory values, with the task of predicting the next value in the sequence. The attention layer has used the values of 16 and 90 to predict the next value in the sequence. In this instance, attention is used to focus on a similar pattern to anticipate a future value.

Figure 3: A vertical edge detector convolution kernel. Edges that transition from dark to light (like that seen in the input image) will be light in the output image. The pixel values (representing light intensity) in this image are given as the pink numbers. Notice that no values in the output image are below zero. This is because all values less than zero are turned into zero by a function known as a rectified linear unit (29) – this is known as an activation function and is a common technique in deep learning. Also note the rim of zeroes around the input image – this is known as "padding", which allows the output image to retain the same dimensions as the input image.

Figure 4: Three unique deep learning methods used in rheumatology literature.
(A) A cascade of CNNs is used to classify power doppler images. At each step, the CNN classifies the image either to a certain OESS class or any class above. If the CNN determines that it belongs to a higher class, it is passed along to the next CNN which performs the same task for the next highest class. Eventually, the final CNN simply classes between class 2 and 3. **(B)** A simplified diagram of the U-Net architecture (51). An image begins as some NxNxC shape, where NxN is the image size (e.g. 224x224 pixels) and C is the number of channels (typically 3 channels of red green and blue for a color image). The model gradually reduces the size, whilst increasing the number of channels, until the bottom of the architecture is reached and then the reverse occurs. Connections across the architecture (dashed lines) act as a "memory". The image recovered at the end is a segmented image, partitioning the original into the relevant parts. In this example, the bones of two metacarpal joints are segmented from the plain radiograph. **(C)** A single coronal radiograph of the knee joint is split into two images: the right half of the knee and the horizontally-flipped left half. Both images are passed through the same convolutional neural network, before joining up to produce a Kellgren-Lawrence Score as the model output.

Figure 5: Three methods to overcome limited datasets. Transfer learning takes a model trained on a large dataset and repurposes it for a new task, replacing only the final layer. Self-supervised learning is a type of transfer learning, however the dataset used for pre-training does not need to have labels – here the task is simply to recognise that two versions of the same image are indeed the same image, and in doing so the model learns to recognise invariant features. Increasing dataset size can be done in a number of ways, however pooling data across institutions has technical, logistical and privacy issues that must be overcome.