Assessing Fracture Healing with Artificial Intelligence:

Using Transfer Learning to Predict the *Radiographic Union Score for Tibial Fractures*, in the Radiography of High-Energy Trauma

Shen Zhou Hong Goldsmiths, UoL

April 4th, 2023

Contents

Chapter 1

Introduction

Long bone fractures are a frequent effect of high-energy trauma [1], among which tibial fractures of the lower extremities are the most common. These fractures require long-term follow-up, where after initial fixation the fracture site must be re-examined at regular intervals for callus formation[1], bridging, and union [2]. Whelan et al's Radiographic Union Score for Tibial Fractures (RUST score) is a discrete 12-point scale that serves a common metric for the assessment of union from the lateral and antero-posterior[2] radiograph of a fracture [3]. This project proposes a means to automate the assessment of fracture healing, by using machine learning to classify radiographs of tibial fractures according to their RUST Scores.

1.1 Aims and Motivation

Non-union and delayed union are significant complications in fracture healing, one which results in heightened morbidity, loss-of-function, and infection risk [4]. As a result, it is important for physicians to determine non-union events so that further treatment and corrective surgery may be taken. Although non-union may be determined through a variety of clinical assessments (e.g. palpation, weight-bearing), the RUST score is emerging as a quantitative radiography-based measure with high consistency [5], biomechanical correlativity [6], and good guidance for postoperative rehabilitation [7]. However, in order to assess a fracture using the RUST score, a orthopaedic must examine at least two radiographs (one lateral, one anteroposterior) for callus formation — a non-trivial process.

Recent advances in deep learning, coupled with the increasing availability of large radiographic datasets (e.g. CheXpert, LERA, MURA) [8, 9, 10] offer the

[1] The development of cartilaginous material containing bone-forming cells.
[2] i.e. front-to-back.

possibility of automating the process of fracture classification. Certain research models such as Rajpurkar et al's DNN ConvNet are able to meet, or exceed radiologist-level performance for abnormality classification in specific anatomical domains [10], and as of 2021 commercial developments are beginning to see regulatory clearance[3] [11].

However, much of the current available literature[4] is focused on the mere detection and classification of fractures (i.e. abnormality detection). Such models either perform binary classification (e.g. "Is this a *normal* radiograph?"), multi-class (e.g. "Is this a *leg*, *arm*, or *knee* fracture?"), or localisation (e.g. "Where *is* the fracture on this radiograph?"). Comparatively less work has been done on the *characterisation* of radiographs, where the properties of a fracture are described [12]. This gap in the field offers opportunity for further investigation, especially as it is not the mere presence of a fracture which informs medical decision-making, but rather its severity and properties. By creating an machine learning model where the RUST-score of a fracture is inferred from a radiograph, we hope to advance the state of AI in medical imaging, and develop better diagnostic tooling.

1.2 Objectives and Evaluation

The objective of this project is to develop an AI model using transfer learning that is able to predict the RUST score of a pair of anteroposterior and lateral radiographs. We will evaluate two different transfer learning approaches based on the InceptionV3 model architecture[13]. The first model will use InceptionV3 trained with the general-purpose ImageNet dataset [14]. The second model will use the same InceptionV3 architecture, but trained with the domain-specific RadImageNet dataset [15]. The development and comparative evaluation of these two 'base' models for transfer learning will allow us to compare and contrast the use of a model pre-trained on a general-purpose dataset (ImageNet) versus a model pre-trained on a slightly smaller, but domain-specific dataset (RadImageNet).

[3] Authorisation for real-world clinical use by national health agencies like the Food and Drug Administration (FDA), Health Canada, *Conformité Européenne*, etc.

[4] A selection of which are analysed with commentary in chapter 2.

1.2.1 Project Specification

Thus, the aims of this project can be summarised as the following three objectives:

- Evaluate the performance of InceptionV3 trained with ImageNet and RadImageNet on a transfer learning task.
- Develop and optimise the best-performing transfer learning model for use in the automated assessment of fracture healing through RUST scores.
- Assess model performance through it's AUROC (Area Under Receiver Operating Characteristic) value.[5]

[5] See 2.2 for further information.

Chapter 2

Background Research

The use of artificial intelligence in the analysis of radiography predates deep learning, with early approaches reliant on predefined engineered features and handcrafted algorithms (e.g. edge detection, wavelet transform) [16]. With the advent of more powerful computer hardware and the democratisation of machine learning through open source frameworks, we see deep learning techniques being applied to the field of medical imagery. These studies often had to solve unique, domain-specific challenges — such as small datasets, the need for *evidence-based* labeling[1], and difficult image-preprocessing requirements. This project will face similar challenges. Hence, by taking a survey of existing literature, we may be better informed in overcoming these challenges.

2.1 Early Period: Small Datasets, Feature-Engineering

The first period of AI-based based fracture detection was characterised by the limitations of small datasets [2] , and innovative approaches aimed to overcome said limits. One example of work in this period was Cao et al's use of feature fusion in random-forests, which allowed multiple categories of features to be considered by the model [17]. Likewise, Dimililer's *Intelligent Bone Fracture Detection system* used meticulous image pre-processing, with Haar wavelet transforms and Sub-Variant Feature Transform (SIFT) [3] in order to extract invariances from the radiography image [18]. Both authors relied on a combination of domain-specific image pre-processing and feature-engineering, in order to compensate for limited datasets that they had. The limitations of their data further manifested in difficulties with model evaluation. In

[1] Labelling that is done by a clinical professional who is empowered to issue diagnoses.

[2] Often working with an individual hospital or medical center, these studies these studies generally had hand-annotated datasets of up to a hundred images, and seldom more than two hundred.

[3] Both functions are algorithms from the domain of signal processing, designed to compress spatio-temporal information in a manner that preserves invariances.

[18], only 100 labelled radiographs were available, making the use of a distinct hold-out validation set impossible. The lack of a separate validation set makes it difficult to judge whether or not the model performs without overfitting, weakening the study's conclusion. One possible mitigation that the author of [18] could have considered was k-fold cross-validation, which [17] does implement. In [17] Cao et al. implements 10-fold cross-validation over a data set of 145 radiographs, hereby yielding a more rigorous assessment of the model's performance. However, [17] did not use any data augmentation strategy, which may have been useful in light of the limited dataset available.

What lessons can we draw from these two studies? Radiographic imagery is highly hetereogenous, with individual x-rays mostly consisting of dark and light regions with sharp transitions in-between. Pre-processing steps such as scaling or down-sampling must capture these discontinuities faithfully, in order to prevent information-loss in the input. We can take particular inspiration from [18], who uses Haar wavelet transform in order to downsample inputs without loss of detail in the transition boundaries.[4] Likewise, we must use k-fold cross validation, and consider a data augmentation strategy for our own model. Overall, these two examples are representative of early, exploratory work in AI-based fracture detection, and they are useful to illustrate both solutions to working with small datasets (pre-processing, feature-engineering), as well as challenges (difficulties with validation). The context of the above studies help us better understand later work, which began with the advent of large, publicly-available datasets: the most important of them being the Stanford Musculoskeletal Radiography Dataset (MURA).

[4] Unlike the discrete cosine transform, the wavelet transforms are not fourier based and therefore discontinuities in image data can be handled with better results using wavelets." [18]

2.2 MURA: Large Musculoskeletal Radiography Datasets

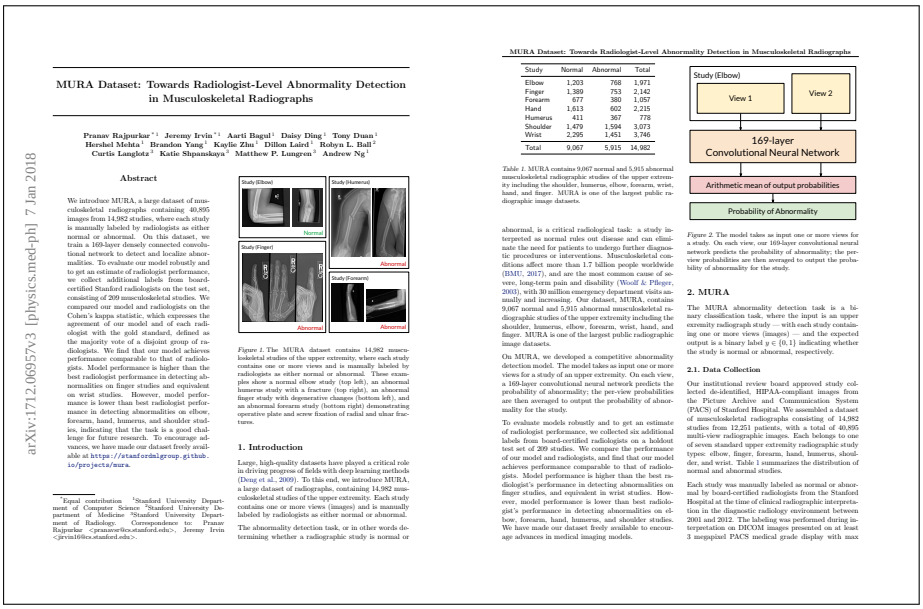


Figure 2.1: Thumbnail of article by Rajpurkar et al. [10]

MURA is one of the first large radiographic datasets focused exclusively on musculoskeletal imagery, as well as the name of a study conducted alongside said data by the Stanford Center for AI in Medical Imaging [10]. The MURA study marks a distinct landmark in the field of AI-assisted fracture classification, because it was the first to apply a deep-learning model on a large (40,561 radiographs), publicly available dataset. As a result, an examination of MURA allows us to contextualise our study in important ways: first, the success of MURA establishes the *possibility* of using deep learning to robustly analyse radiography. Rajpurkar et al. demonstrates near-radiologist levels of performance, validating the feasibility of our project in general. Secondly, the fact that the MURA model takes multi-view input imagery will help inform our own architectural design for processing both lateral and antero-posterior data. Finally, the limitations of MURA being a pure anomaly-detection system allows us to understand the need for a model which goes beyond binary classification, i.e. what our project is trying to accomplish.

To begin, it will be useful to look at MURA's architecture The MURA model is a 169-layer convolutional neural network which takes one or more views as input[5], and delivers a *probability of anomaly*. The final probability is the arithmetic mean of output probabilities from every view [10]. Every type of radiograph may have

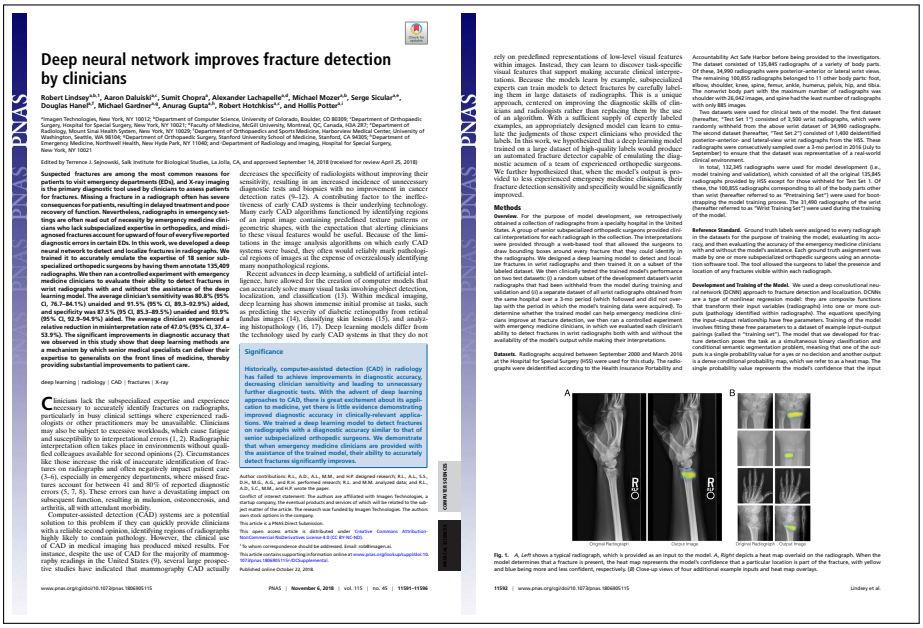
[5]Different views are radiographs of the same subject taken at different standard perspectives.

multiple standard views, and the model architect has a choice of either training a model on a single view, or combining information from views in some ensemble stage. For anomaly detection, MURA chose a fairly conservative approach of assessing a separate probability of anomaly for each view, and then finding their average. This approach will be similar to the one that our project must take: which is to look at both the lateral and anteroposterior view of the tibia.

Another aspect of MURA that is worth examining, is their evaluation process. The model was accessed in two different ways: first, the model's precision versus recall plotted out in a Receiver Operating Characteristic (ROC) plot, and then the Area-Under-Curve of the ROC (AUROC) was quantified. The AUROC is a common metric to assess diagnostic ability since it serves to quantify the precision-recall curve of the model. The MURA model had an AUROC of 0.929. Second, a panel of three radiologists were assembled to evaluate a set of radiographs, and their performance was compared against the model. This kind of competitive evaluation allowed the study to compare the model performance against human clinicians, yielding an informative baseline for the AUROC. With this information, we can contextualise the earlier AUROC value of 0.929, and see that the model performs slightly worse than humans.

The use of AUROC as an evaluative metric, as well as competitive evaluation with human clinicians, present an advancement in model evaluation compared with the two earlier studies. However, in practice any AI model in radiography will not aim to replace human radiologists entirely, but serve as an additional tool which augments a human clinician's own diagnostic ability. Thus, the MURA study's evaluation is not representative of what real-world deployment would look like. This is why we must turn to Lindsey et al's *Deep neural network improves fracture detection by clinicians*, for a more holistic example of model evaluation.

2.3 Lindsey et al: Refinements to Model Evaluation



Deep neural network improves fracture detection by clinicians

Figure 2.2: Thumbnail of article by Lindsey et al. [19]

In this study, a deep convolutional neural network is trained on a dataset of 31,490 wrist radiographs [19]. Like MURA, the study presents an AI-based fracture detection model, outputting both labels as well as a location heatmap. We include this study in the background research, for it's more holistic approach to evaluation, which simulates a real-world use-case. "Radiographic interpretation often takes place in environments without qualified colleagues available for second opinions" [19], the paper acknowledges, before proposing a model which aims to serve a possible 'second opinion.' After an initial evaluation which demonstrated an AUROC of 0.967 on certain datasets[6], Lindsey et al. proceeds to conduct a second experiment, where a group of clinicians were shown radiographs from the same test set, and tasked to evaluate the radiograph both with and without the model's assistance:

[6]For model testing, the study utilised two distinct categories of radiographs divided into 'Test Set 1' and 'Test Set 2' (a slight improvement over MURA), with the former containing a collection of singleton wrist radiographs, and the latter of wrist radiographs with anterio-posterior and lateral views.

> For each radiograph shown, the clinicians were asked whether or not a fracture was present. After a clinician made a response, the model's semantic segmentation prediction [i.e. heatmap] was shown overlaid on the radiograph; the model's clinical determination was shown as text, and the clinician was asked the same question again. [19]

This experiment simulates the workflow of a clinician when using a fracture-detection model as a part of a CAD workflow. By doing so, we evaluate the effectiveness of the model as an useful tool within a broader clinical practice. The study found that the "... sensitivity and specificity of the emergency medicine MDs were significantly improved with the assistance of the deep learning model." [19]

The key advantage of Lindsey et al's paper lies in its evaluative design, which our own project must take inspiration from. It may be impractical to setup a similar trial involving radiologists or clinical professionals,[7] but one feature that we do aim to replicate is the output of a heatmap which highlights the location of features that the model detects. For the Lindsey et al. and the MURA study, these heatmaps highlighted fracture-sights, whereas for our project they will highlight the sites of callus formation and bridging. By having this as a output, we will make the model's behaviour much more interpretable, and allow for integration into real clinical workflows.

[7] Such an addition to this project is, however, within the realm of possibility, in collaboration with the medical faculty at METRC.

2.4 Kim & MacKinnon: Cross-Domain Transfer Learning

Clinical Radiology 73 (2018) 439–445

Contents lists available at ScienceDirect

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net

Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks

D.H. Kim*, T. MacKinnon

Medical Imaging Department, Royal Devon and Exeter Hospital, Barrack Road, Exeter EX2 5DW, UK

ARTICLE INFORMATION

Article history:
Received 28 May 2017
Accepted 14 November 2017

AIM: To identify the extent to which transfer learning from deep convolutional neural networks (CNNs), pre-trained on non-medical images, can be used for automated fracture detection on plain radiographs.

MATERIALS AND METHODS: The top layer of the Inception v3 network was re-trained using lateral wrist radiographs to produce a model for the classification of new studies as either "fracture" or "no fracture". The model was trained on a total of 11,112 images, after an eightfold data augmentation technique, from an initial set of 1,389 radiographs (695 "fracture" and 694 "no fracture"). The training data set was split 80:10:10 into training, validation, and test groups, respectively. An additional 100 wrist radiographs, comprising 50 "fracture" and 50 "no fracture" images, were used for final testing and statistical analysis.

RESULTS: The area under the receiver operator characteristic curve (AUC) for this test was 0.954. Setting the diagnostic cut-off at a threshold designed to maximise both sensitivity and specificity resulted in values of 0.9 and 0.88, respectively.

CONCLUSION: The AUC scores for this test were comparable to state-of-the-art providing proof of concept for transfer learning from CNNs in fracture detection on plain radiographs. This was achieved using only a moderate sample size. This technique is largely transferable, and therefore, has many potential applications in medical imaging, which may lead to significant improvements in workflow productivity and in clinical risk reduction.

© 2017 The Royal College of Radiologists. Published by Elsevier Ltd. All rights reserved.

Introduction

Demands on radiology services have increased dramatically in recent years causing a considerable strain on the workforce. In the UK, the number of computed tomography (CT) examinations increased by 29% between 2012 and 2015, whilst recruitment lagged behind, leaving 9% of consultant radiology posts vacant in 2015. Furthermore, in the same year nearly all radiology departments declared a failure to meet their reporting requirements.[1] In February 2016, there was an estimated backlog of 200,000 plain radiographs and 12,000 cross-sectional studies.[2] It is clear from these figures that improvements in reporting efficiency and workflow management are desperately needed if we are to avoid patient harm from delayed or missed diagnosis.

Artificial intelligence (AI) has the potential to address these issues. The wide adoption of electronic picture archiving and communications systems (PACS) has resulted in the development of one of the largest image data sets in existence. In the UK, there were 41 million imaging studies performed in 2016 alone.[2] These data lend themselves perfectly to machine learning.

Machine learning is a form of AI, which uses algorithms that iteratively improve, or learn, in response to training data in order to make autonomous predictions.[3] Supervised machine learning is one subtype, which relies on the provision of pre-labelled training data. In the field of general imaging and computer vision, deep learning is the leading machine learning tool.[3] Deep learning refers to techniques that build on developments in artificial neural networks in which multiple network layers are added to increase the levels of abstraction and performance.[3]

A simplified overview of the mechanics behind deep convolutional neural networks (CNNs) is provided in Fig 1. Neural networks aim to mimic the structure of the human brain. They use a series of interconnected "neurons", which collect, summarise, and/or transform data before passing the values to the next neuron in the sequence. This process culminates in an output layer, which can be used to formulate a prediction.

Constructing and training an effective neural network from scratch requires huge amounts of data. State-of-the-art image classification networks are frequently trained on data sets containing millions of images, facilitated by multiple computer servers, running continuously for several weeks.[4] This is not feasible for the majority of medical researchers. One method for overcoming this is to use a process called transfer learning. This is the process of adopting powerful and highly refined features from large existing pre-trained CNNs and using these as a starting point in training a new model for a different task. These low-level features, such as those illustrated in Fig 2, can be thought of as the basic building blocks of images such as lines and curves, and have been shown to be applicable to many different image-recognition tasks.[20] This technique can vastly reduce the computational requirements needed for network training and can deliver substantial performance benefits compared with training a CNN from scratch.[21]

Transfer learning has been relatively underutilised in the clinical setting despite the availability of vast amounts of image data. Two recent publications have demonstrated that transfer learning from pre-trained CNNs can produce human-level diagnostic results in the categorisation of skin lesions and in defining disease on digital retinal images.[22,23] Similar levels of diagnostic accuracy have yet to be reported in the analysis of plain radiographs; however, a few promising studies have recently been published. For example, one study retrained the GoogLeNet CNN for the detection of pathology in plain frontal chest radiographs, resulting in an area under the curve of between 0.861 and 0.964 for different chest radiograph features.[24] A different study used transfer learning from a pre-trained ImageNet CNN for the automated categorisation of osteoarthritis in knee radiographs.[25]

Plain radiographs are the most common radiological test with over 22 million studies performed in the UK in 2016.[2] A high proportion of these are plain extremity radiographs in the context of trauma. There is a strong case for developing automated strategies to improve efficiency and workflow management in this area considering the backlog in unreported studies and the fact that over £88 million was spent on outsourcing radiology reports in 2014–2015.[1] It is surprising therefore that, to the author's knowledge, there are currently no studies in the literature that have successfully applied transfer learning from pre-trained CNNs to the problem of fracture detection on plain radiographs. This proof of concept study aims to establish to what extent this is possible.

Materials and methods

This National Health Service (NHS)-based study was granted approval by the Health Research Authority in England. This study was retrospective and used only anonymised data, and therefore, ethics approval was not required.

Imaging study selection

Anonymised lateral wrist radiographs were obtained from the Royal Devon & Exeter Hospital for studies performed between January 2015 and January 2016. Radiographs were excluded if there was a plaster cast in place, if the growth plates of the wrist had not yet fused, or if the study demonstrated any fracture other than a fracture of the distal radius or ulna. The images were classified as either "fracture" or "no fracture" on the basis of the radiological report. This classification was checked and verified by a radiology registrar competent in the reporting of plain radiographs and with 3 years radiology experience. Images were also excluded if the single lateral projection was inconclusive for the presence or absence of fracture.

Image pre-processing

This resulted in a preliminary data set of 695 wrist radiographs demonstrating a fracture and 694 wrist radiographs demonstrating no fracture. The images were converted to JPEG by a trained radiologist ensuring the most appropriate windowing was selected. Images were then pre-processed by removing additional annotations such as the commonly used "Red Spot" annotation applied by radiographers at the time of image acquisition in order to reduce the possibility of over-fitting. Over-fitting describes the process by which the learning algorithm learns features that are not truly representative of the image category in an attempt to reduce the error in the learning process.

A data augmentation technique was used to amplify the data. This involved making a number of non-exact copies, or transformations, of each image. This served to provide the CNN with more training examples by incorporating the salient features in multiple orientations. The aim was to better reflect the real-world population of wrist radiographs,

Figure 2.3: Thumbnail of article by Kim and MacKinnon. [20]

Studies [10, 19] both rely on large, labelled radiographic datasets. While such datasets exist for fracture *detection* and fracture *classification*, the classification of fracture healing by their RUST scores is without precedence in literature, and to our knowledge there are no openly available datasets with the aforementioned labels. For our project, we will be using radiographic data internally collected by the Major Extremity Trauma Research Consortium (METRC), a research unit at the Johns Hopkins Bloomberg School of Public Health. Although this dataset has the RUST-scores we require, the number of radiographs we have access to is limited. Hence our approach is to utilise *transfer learning*, a process where the model is first trained on a larger, general-purpose dataset, before being fine-tuned on the study data. Thus, we turn to Kim and MacKinnon's *Artificial intelligence in fracture detection* for their innovative use of transfer learning. [20]

In Kim and MacKinnon's study, a dataset of 1,389 radiographs was available for a binary classification task (i.e. labels were "fracture" or "no fracture"). Instead of training a model *ab initio*, a CNN trained on a general purpose, non-radiographic dataset was re-applied on the radiography data. The authors used the Inception v3 network, an object-detection and image classification network [21] with the topmost layer fine-tuned on the radiographic data. [20]. By using a pre-existing model and

data augmentation[8], Kim and MacKinnon were able to achieve an AUROC of 0.954, a value that is within the same standard deviation of [10, 19], studies which both had much larger datasets for their models.

Our inclusion of Kim and MacKinnon's study serves as an important validation for the aims of our project. The METRC dataset will consist of around two to three thousand radiographs sourced from a handful of METRC studies. If Kim and MacKinnon's model is already able to achieve a robust performance with only 1,389 samples, than it is quite possible for our own project to achieve it's aims. The theoretical reasoning behind why transfer learning can achieve high levels of model performance is because the first layers of a CNN are primarily responsible for high-level feature extraction, e.g. edge detection, pattern recognition, with only the latter layers responsible for task-specific classification [20]. This does, however open up a further avenue of inquiry. Would model performance improve further, if the base model that the transfer learning is conducted from, was itself originally trained on a domain-specific dataset, like the MURA data from [10]? Our project will explore this possibility, through it's own application of transfer learning to the prediction of radiographic union with RUST scores.

[8]The authors of [20] ultimately generated 11,112 augmented samples from their initial set of 1,389 radiographs.

Chapter 3

Methodology

The primary objective of our study is to develop an AI model using the technique of transfer-learning, to automatically predict RUST scores from an input radiograph. In this chapter, we will first discuss our experiment design, contextualising the choices we make in our methodology and evaluation. Next, we will briefly summarise the datasets used in the study, before proceeding to the main methodology (protocols) itself. Afterwards, the evaluation and endpoints will be presented, as well as a discussion on patient privacy and ethical considerations.

3.1 Model Design

Transfer learning is a technique which uses a model trained upon a larger dataset first, before being applied to a smaller, task-specific dataset. Assuming our task-specific dataset (in our case, the RUST data from METRC) is fixed, the performance of our transfer-learning model is determined by the following factors:

1. The architecture of the original model.
2. The initial, large dataset that the original model is trained on.

Hence, in the context of our study design, we must first choose (or create) a model architecture, and select an initial large dataset to train our model on. Only then are we able to apply the transfer learning procedure (e.g. freeze model weights, add new classifier, fine tuning, etc) upon our task-specific dataset.

3.1.1 Model Architecture Choices

The first decision that we must make in our experiment design is the choice of model architecture. According to a literature survey by Litjens, Kooi, Bejnordi et al., convolutional neural networks (CNNs) are the most common deep learning architecture deployed for medical image analysis, vastly outnumbering alternative methods such as

Stacked Autoencoders (SAE), or Recurrent Neural Networks (RNNs) [22, p. 77]. This is expected, as CNNs exhibit strong performance with image classification tasks, and as our project is an image classification task (albeit with radiographs), we will be using a CNN.

The choice now remains to either design our own CNN from scratch, or use an existing CNN architecture. Although designing a CNN *ab initio* allows the possibility of further experimentation and potentially finer-grained control, such a *de novo* model is difficult to assess holistically: there would be no prior work in literature to serve as a basis for comparison. In contrast, if we use a well-documented, existing CNN as our transfer-learning foundation, we may compare the performance of our implementation against other instances of the same model architecture used for transfer-learning in different domains.

Thus, we will be using the InceptionV3 model, a convolutional neural network developed by Szegedy, Vanhoucke, Ioffe, et al. from Google [13]. Our choice of InceptionV3 is based upon a 2022 literature review of transfer-learning models for the domain of medical image analysis. According to Kora, Ooi, Faust et al., CNNs with a broad (as opposed to deep) network topology perform well in transfer-learning tasks [23], with models like InceptionV3 outperforming more parameterized models like AlexNet [24]. Indeed out of the 54 studies included in the review, Inception-style models both the most common (14 out of 54) and among the highest-performing. [23]

3.1.2 'Top-Classifier' Choices

This subsection is unavailable in this version of the document.

3.1.3 Model Optimizer Choices

Finally, the last architectural decision we must make in our model design, is the choice of an optimizer. We have two possible approaches: we may either select an optimizer from *a priori* principles, or consider the selection of an optimizer to be a hyperparameter, and benchmark a variety of optimizers with our model on the data-set. Because we are already evaluating two variants of InceptionV3, the additional task of iterating through different optimizers will be infeasible given the time and compute constraints of the project.

Hence, our choice of an optimizer is determined by a review of available benchmarks and literature. The study and benchmarking of deep learning optimizers is a fairly recent field, initiated by the development of robust, reproducible benchmarks. Projects like Schneider, Balles, et Hennig's *DeepOBS: A Deep Learning Optimizer Benchmark Suite* allowed researchers to evaluate optimizers against an assay of realistic optimization problems, simulating common deep learning tasks and neural network architectures. [25] The availability of reproducible benchmarks allowed the first large-

scale empirical experiments to be conducted on optimizer performance, culminating in Schmidt, Schneider, et Hennig's 2021 paper in optimizer benchmarking. [26] In an evaluation of 15 popular optimizers[1] across a total of more than 50,000 epochs of training, significant data on optimizer performance was gathered.

Of the eight optimization problem assays in the benchmark suite, our task of radiographic image classification is bears closest resemblance to the CIFAR-10 benchmark: a CNN-based image classification task. According to the latest results available on the paper's website[2], the ADAM optimizer has a slight accuracy improvement over Momentum and SDG. However, the differences are so small that the author mentions:

> ... a practitioner with a new deep learning task can expect to do about equally well by taking almost *any method* from our benchmark and tuning it, as they would by investing the same computational resources into running a set of optimizers with their default settings and picking the winner. [26, p. 2]

Therefore, we will select the ADAM optimizer, and spend time on fine-tuning the model and hyperparameter choices, instead of devoting further resources to selecting an optimizer.

3.1.4 Pre-Training Dataset Choices

Now that we decided our model architecture, the next step is to select the initial, or pre-training dataset, that is used to train our model.

ImageNet Dataset

Originally, InceptionV3 was trained on the ImageNet dataset: a general-purpose collection of more than 14 million everyday images. [14] The overwhelming size of the ImageNet dataset serves as a robust foundation for the InceptionV3 model, which exhibits strong performance in classification tasks upon it. However, the statistical characteristics of data in ImageNet is quite different from data in a typical radiography dataset. Hence, it remains a valid research question to ask whether a InceptionV3 model trained on a smaller, but more domain-specific dataset will exhibit better performance when applied to our transfer-learning task.

RadImageNet Dataset

Thus, we will also investigate RadImageNet, a collection of 5 million medical images composing of radiographic (CT), MRI, and ultrasound images. [15] As an open radiologic dataset developed specifically for transfer learning applications, a base model trained on RadImageNet may exhibit better performance on our radiography data, as the original network is trained upon images in a similar domain. However, the

[1] AMSBound, AMSGrad, AdaBelief, AdaBound, AdaDelta, Adam, LookaheadMomentum, LookaheadRAdam, Momentum, NAG, NAdam, RAdam, RMSProp, and SGD, respectively. The full list of results are available in their supplementary appendix.

[2] https://deepobs.github.io/leaderboardP4.html

advantages of a similar domain is moderated by the corresponding smaller dataset size (5 million versus ImageNet's 14 million).

Therefore, in this project we will evaluate the use of a InceptionV3 model trained both on the ImageNet dataset, as well as the RadImageNet dataset as the base model for transfer learning. By comparing the performance of both models, we will be able to select the best-performing one for further development and refinement. Additionally, the additional data point afforded by a second model allows greater context for our evaluation: according to Kora et al., only 13% of transfer-learning studies benchmarked their model performance against a second model. [23, p. 94] By choosing to evaluate two models from the very start of our project's design, we get to avoid this scientific blind-spot: and hopefully achieve a better result overall.

3.2 Data Egress and Preprocessing

Now that we have defined the design of our experiment, we must define the pre-processing and data egress requirements. Our study data consists of anteroposterior and lateral view radiographs, as well as their corresponding RUST scores. This data is provided in a collaboration with METRC, Johns Hopkins University, through their archive of past and on-going studies. [27, 28, 29, 30] The data is held within REDCap: an electronic data and clinical trials database. [31] Because the data is held across multiple clinical trials and within multiple instruments, egressing data out of REDCap and pre-processing it into a useable form is a non-trivial software engineering task.

Dataset	Name	Samples
[27] RetroDEFECT	*Retrospective Study of the Treatment of Long Bone Defects*	741
[28] Outlet	*Outcomes Following Severe Distal Tibia, Ankle and/or Foot Trauma*	707
[29] Pain	*Pain Management & Long Term Outcomes Following High Energy Orthopedic Trauma*	370
[30] Pacs	*Predicting Acute Compartment Syndrome using Optimized Clinical Assessment*	195
Sum		2,013

Table 3.1: Sources of Radiographic Data with RUST labels in REDCap

3.2.1 Validate Radiography Data with Branch-Parser

The first challenge that we face in the data egress process, is finding RUST-radiography pairings that are valid for inclusion in our dataset. A small subset of radiographs

do not possess valid RUST scores: either because they were uninterpretable due to hardware occlusion (e.g. presence of a titanium orthopaedic fixture), or because the radiograph was not taken for the purpose of assessing fracture healing. In order to validate the radiography data, a Python package called redcap-branch-parser was developed to parse the conditional logic contained within REDCap patient records. [32] The development of this package took up a non-trivial amount of early project work.

Once valid RUST-radiograph pairs were identified, the study data was egressed out of REDCap using a set of ad-hoc Jupyter Notebooks and the PyCap API package. [33]

3.2.2 Automated de-skewing with ImageMagick

Following the egress of radiography data, the radiography image files were automatically de-skewed using ImageMagic. At this point, data-preprocessing is complete, as data augmentation will be performed dynamically via layers within Keras.

3.3 Data Augmentation Strategy

Because we have a small dataset, a robust data augmentation strategy is needed to combat overfitting. In order to develop our data augmentation strategy, we followed suggestions from Bejani et Ghatee's *systematic review on overfitting control in shallow and deep neural networks* [34]. We may categorise data augmentation strategies to three broad families of methods: [35]

1. **Geometric Methods:** rotation, cropping, flipping.
2. **Photometric Methods:** noise, color-shifting, edge modification.
3. **'Complex' Methods:** generating artificial data using GANs, style transfer.

Of the three, 'complex' data augmentation methods such as generating artificial data points is inappropriate for our use case, as it will compromise the evidence-based labelling of our dataset. Hence, we may recourse to either geometric methods, or photometric methods. Originally, our intuition was that photometric data augmentation methods would be of limited applicability to radiographs, which are entirely greyscale. Upon further investigation, it turns out that when evaluated individually, the most effective data augmentation strategy is cropping, followed by rotation. [36]

Hence, in order to avoid potentially compromising our small dataset with too much noise, our data augmentation strategy will be restricted to geometric methods such as rotation, cropping, and flipping.

3.4 Protocols

Now that we have finished discussing the model design, data egress, and augmentation, it is time to define the study protocols. The experimental portion of this study will consist of three protocols:

1. Protocol I: A 'naive' CNN without transfer-learning, for use as a baseline.
2. Protocol II: Transfer-learning w/ InceptionV3 trained on ImageNet.
3. Protocol III: Transfer-learning w/ InceptionV3 trained on RadImageNet.

All protocols will utilise the same data augmentation pipeline.

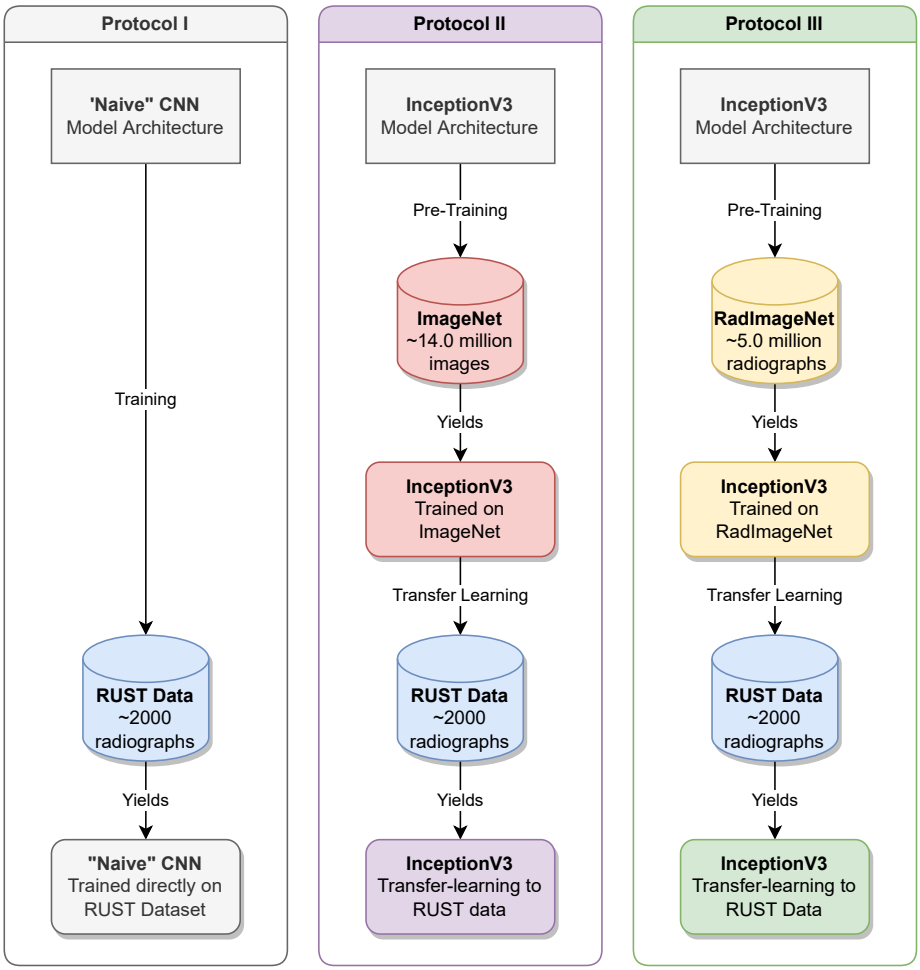


Figure 3.1: Overview illustrating the three model development protocols.

3.4.1 Protocol I: 'Naive' CNN Baseline

In protocol I, we will develop a 'naive' CNN without the use of transfer learning, that is trained directly on the RUST dataset. This model will serve as a performance baseline to benchmark our transfer-learning models against.

'Naive' CNN Architecture

This subsection is unavailable in this version of the document.

3.4.2 Protocol II: InceptionV3 with ImageNet

In protocol II, we will use InceptionV3 trained on the ImageNet dataset as a base model for transfer-learning.

This is an outline. More information will be added later.

1. Instantiate base model with pre-trained weights.
2. Freeze weights in base model. Remove output (classifier) from base model.
3. Create a new classifier on top of base model.
4. Create input augmentation pipeline using keras layers.
5. Normalise RUST dataset towards mean and standard deviation of ImageNet dataset
6. Compile model with optimizer settings.
7. Train classifier (i.e. top layer).
8. Unfreeze weights in base model.
9. Perform one additional round of training with a very small learning rate, for fine-tuning.
10. Evaluate model performance.

3.4.3 Protocol III: InceptionV3 with RadImageNet

In protocol II, we will use InceptionV3 trained on the RadImageNet dataset as a base model for transfer-learning.

This is an outline. More information will be added later.

1. Instantiate base model with pre-trained weights.
2. Freeze weights in base model. Remove output (classifier) from base model.
3. Create a new classifier on top of base model.
4. Create input augmentation pipeline using keras layers.
5. Normalise RUST dataset towards mean and standard deviation of ImageNet dataset
6. Compile model with optimizer settings.
7. Train classifier (i.e. top layer).
8. Unfreeze weights in base model.

9. Perform one additional round of training with a very small learning rate, for fine-tuning.
10. Evaluate model performance.

3.5 Hyperparameter and Learning Rate Tuning

The hyperparameter and learning rate tuning is only applicable to *either* Protocol II or Protocol III, depending on which was the better-performing one.

3.5.1 Hyperparameter Tuning Regime I

This subsection is unavailable in this version of the document.

3.5.2 Hyperparameter Tuning Regime II

This subsection is unavailable in this version of the document.

3.5.3 Learning Rate Schedule

This subsection is unavailable in this version of the document.

3.6 Evaluation and Endpoints

Assuming an available dataset of $\approx 2,000$ samples (after removing invalid entries), we will be setting aside 15% of the initial data (≈ 300 samples) for the testing dataset.

3.6.1 AUROC

Model performance will be assessed via AUROC, the Area-Under-Curve of the Receiver Operating Characteristic (ROC) graph. This is the area underneath the precision-versus-recall plot of the model, and is a common metric used to assess diagnostic accuracy.

3.6.2 K-Fold Cross-Validation

Due to the small size of our dataset, we will be using K-Fold Cross validation as a technique to get a more accurate assessment of our model performance. Instead of setting aside a hold-out validation set of 15% (i.e. ≈ 300 samples out of the initial $\approx 2,000$), we will be running $k = 5$ folds on the $1,700$ sample data. This yields ≈ 340 samples per validation fold, which approximates the usual ≈ 300 samples of a regular hold-out validation set.

3.6.3 Endpoints

First, prior to assessing model performance, we want to determine:

1. *Does InceptionV3 trained on the domain-specific RadImageNet dataset perform better as a transfer-learning base on our RUST dataset?*

Protocol I and Protocol II will allow us to answer that research question. Once we have

found the best-performing variant of InceptionV3, we will then begin the process of optimising model performance as measured through AUROC. For performance, we aim to achieve the following endpoints:

1. Endpoint 1: AUROC > 0.50
2. Endpoint 2: AUROC > 'Naive' CNN
3. Endpoint 3: AUROC > 0.75

Initially, we aim to achieve an AUROC that is greater than chance, which is the minimal backstop to model performance. Failure to reach this endpoint will indicate severe flaws with our study design, such as our dataset being too small for even transfer-learning to be applicable. Following that, we want to exceed the performance of our 'naive' CNN. Finally, we wish to validate the concept of using AI to infer RUST scores from radiographs, by achieving an AUROC that is greater than 0.75.

3.7 Feasibility and Proof of Concept

Right now, a significant and on-going challenge is completing the initial data egress and pre-processing. As the study dataset is still not assembled and ready to use yet, initial feasibility and proof of concept experiments can be conducted on a readily available dataset from a similar domain, such as the MURA musculoskeletal radiography data from Stanford. By beginning with an artificially constrained subset of images and labels from MURA, we may validate our transfer learning protocol and gather some preliminary information.

3.7.1 Experiments with the MURA Dataset

This subsection is unavailable in this version of the document.

3.7.2 Preliminary Results

This subsection is unavailable in this version of the document.

3.8 Ethical Considerations

Any research project or study that involves human subjects will necessitate ethical consideration. This section is only a brief discussion of ethical risks and mitigations. The aim of this project is to validate transfer learning as a technique to infer RUST scores from a small dataset of radiographs. Although we aim to achieve robust performance in order to demonstrate the feasibility of this technique as an avenue for further research, at no time do we imply to develop a *diagnostic* tool for clinical use. The overarching spirit of the study is to explore interdisciplinary applications of AI in medical imaging, in hopes of reducing clinical caseload for medical practitioners, leading to better standards of care for all patients overall. This overarching goal is a positive one, which aims to improve health and healthcare for all.

3.8.1 Human Radiographic Data

This study works with radiographs collected from human subjects, that were a part of past or on-going METRC studies in high-energy trauma. This data is fully anonymised, and does not contain any personally identifiable information.

3.8.2 Human Subject Research, and HIPAA Compliance

As a part of Johns Hopkins Bloomberg School of Public Health's IRB requirements, all researchers working with human data, even anonymised data, must complete a Human Subject Research certification, and a Information Privacy Security certification.

Chapter 4

Resources

4.1 Programming Languages, Frameworks, and Libraries

The AI models in this project will be developed in Python using the Tensorflow and Keras frameworks. Additional pre-processing and numerical analysis will be done with Pandas and NumPy.

4.2 Inference and Compute Requirements

This project will conduct inference through a remote Jupyter kernel provided by Paperspace Gradient, a hosted machine learning platform.

Compute costs will be self-funded.

Chapter 5

Outcomes and Deliverables

5.1 Chapter Structure

It is the intention of this project to deliver a report with the following chapters:

- Abstract
- Introduction
- Literature Review
- Methodology
- Research Results
- Conclusion
- Bibliography
- Appendices

5.2 Software and Git Repository

The project aims to deliver the following software artefacts:

1. A documented Jupyter notebook report, containing model information.
2. The final fine-tuned model with weights and parameters.[1]
3. A standalone Jupyter notebook or Python script for conducting inference.

Within the best of our ability, all project-related software and code will be released under an appropriate open source license, such as the GPLv3. A Git repository containing the project code will be available.

[1] Due to HIPAA (Health Insurance Portability and Accountability Act) and regulatory requirements, it may not be possible to deliver the original radiographic datasets.

5.3 Timeline

This project will endeavour to conform to the following timeline.

Date	Tasks
2022-12-05	Selection of base models for transfer learning
2022-12-12	Initial evaluation of base models on METRC dataset
2022-12-16	**Deadline:** Design Specification
2023-01-02	Pre-processing of METRC Radiographs
2023-01-00	Training and fine-tuning of models on METRC Radiographs
2023-02-27	Further time dedicated to model development.
2023-03-06	Complete model evaluation, AUROC, etc.
2023-03-31	**Deadline:** Implementation and Analysis
2023-04-10	Work on Report, Evaluation, Further Studies.
2023-04-24	Final cleaning and documentation of project code.
2023-05-02	**Deadline:** Poster Presentation
2023-05-12	**Deadline:** Evaluation and Executive Summary

Table 5.1: Proposed Project Timeline.

5.4 Potential Risks and Contingency Planning

Due to the scope of this project, we left ample room in the timeline in case of unexpected difficulties. Although the goal of predicting RUST scores from radiographs has not been attempted before, the task of using transfer learning on radiographic data is well documented in literature, and has resulted in models with near-human performance using data sets of similar magnitudes [20]. Hence we are confident of the goals that the project has set.

For our minimal viable product, we aim to deliver a model that is capable of generating heat maps and predicting RUST Scores. Only once we validate the architecture, will we aim for achieving higher AUROC scores.

Chapter 6

Implementation

```
class TransferLearningModel(tf.keras.Model):
    def __init__(self, dropout_rate: float, **kwargs):
        super().__init__(**kwargs)

        self.input_layer: tf.Tensor = layers.InputLayer(input_shape=(299, 299, 3))
        self.data_augmentation: tf.keras.Sequential = tf.keras.Sequential([
            layers.RandomFlip(seed=RNG_SEED),
        ])

        self.inceptionv3: tf.keras.Model = tf.keras.applications.InceptionV3(
            include_top=False,
            weights='imagenet'
        )
        self.inceptionv3.trainable = False

        self.classifier: tf.keras.Sequential = tf.keras.Sequential([
            layers.GlobalMaxPooling2D(),
            layers.Dense(1024, activation='relu'),
            layers.Dropout(dropout_rate),
            layers.Dense( 512, activation='relu'),
            layers.Dropout(dropout_rate),
            layers.Dense( 256, activation='relu'),
            layers.Dropout(dropout_rate),
            layers.Dense(  18, activation='sigmoid')
        ])

        self.model: tf.keras.Sequential = tf.keras.Sequential([
            self.input_layer,
            self.data_augmentation,
            self.inceptionv3,
            self.classifier
        ])

    def call(self, inputs):
        return self.model(inputs)
```

Listing 1: Sharding dataset for K-Fold Cross Validation

```python
def cross_validate(ModelClass: tf.keras.Model, ds: tf.data.Dataset, epochs: int = 50,
↪  batch_size: int = 128, k: int = 10) -> list[tf.keras.callbacks.History]:

    history_list: list[tf.keras.callbacks.History] = []
    train_valid_pairs: list[tf.data.Dataset] = k_fold_dataset(ds, k)

    for i, (ds_train, ds_valid) in enumerate(train_valid_pairs):

        tf.keras.backend.clear_session()
        model = ModelClass()
        model.compile(
            optimizer=tf.keras.optimizers.Adam(),
            loss=tf.keras.losses.BinaryCrossentropy(),
            metrics=metrics
        )
        history = model.fit(
            ds_train,
            validation_data=ds_valid,
            epochs=epochs,
            batch_size=batch_size,
        )
        history_list.append(history.history)

    return history_list
```

Listing 2: K-Fold Cross Validation Implementation

```python
def k_fold_dataset(ds: tf.data.Dataset, k: int = 10) -> list[tuple[tf.data.Dataset,
↪  tf.data.Dataset]]:
    # First shard the given dataset into k individual folds.
    list_of_folds: list[tf.data.Dataset] = []
    for i in range(k):
        fold: tf.data.Dataset = ds.shard(num_shards=k, index=i)
        list_of_folds.append(fold)

    # Next, generate a list of train and validation dataset tuples
    list_of_ds_pairs: list[tuple[tf.data.Dataset, tf.data.Dataset]] = []
    for i, holdout_fold in enumerate(list_of_folds):
        ds_valid: tf.data.Dataset = holdout_fold

        # Select every fold except holdout_fold as the training folds
        training_folds: list[tf.data.Dataset] = list_of_folds[:i] +
        ↪  list_of_folds[i+1:]

        # ds_train size is (k-1)/k of the original dataset
        ds_train: tf.data.Dataset = training_folds[0]
        for fold in training_folds[1:]:
            ds_train = ds_train.concatenate(fold)

        ds_pair: tuple[tf.data.Dataset, tf.data.Dataset] = (ds_train, ds_valid)
        list_of_ds_pairs.append(ds_pair)

    return list_of_ds_pairs
```

Listing 3: Sharding dataset for K-Fold Cross Validation

```
def hyperparameter_search(trials: int, kfolds: int = 6, epochs: int = 20) ->
↪  list[dict[str, Union[int, float, list[tf.keras.callbacks.History]]]]:
    search_results: list[dict[str, any]] = []

    for trial in range(trials):
        # Randomly pick hyperparameter options
        rng = np.random.default_rng()
        batch_size  : int   = rng.integers(16, 2048, endpoint=True)
        dropout_rate: float = rng.uniform(0.0, 0.5)

        # Conduct K-Fold cross-validation with given hyperparameters
        results: list[tf.keras.callbacks.History] = cross_validate(
            TransferLearningModel,
            ds_train_and_valid,
            epochs=epochs,
            batch_size=batch_size,
            dropout_rate=dropout_rate,
            k=kfolds
        )

        search_results.append({
            "max_val_auc" : calc_kfold_max(results, "val_auc"),
            "batch_size"  : batch_size,
            "dropout_rate": dropout_rate,
            "history_list": k_fold_results
        })

    return search_results
```

Listing 4: Hyperparameter Search Implementation

Bibliography

[1] H. Stein, I. Weize, D. Hoerer, A. Lerner, N. Rozen, and G. Nierenberg, "Musculoskeletal trauma: High- and low-energy injuries," *Orthopedics*, vol. 22, no. 10, pp. 965–967, 1999. DOI: 10.3928/0147-7447-19991001-14. eprint: https://journals.healio.com/doi/pdf/10.3928/0147-7447-19991001-14. [Online]. Available: https://journals.healio.com/doi/abs/10.3928/0147-7447-19991001-14.

[2] M. S. Jones and B. Waterson, "Principles of management of long bone fractures and fracture healing," *Surgery (Oxford)*, vol. 38, no. 2, pp. 91–99, 2020, ISSN: 0263-9319. DOI: https://doi.org/10.1016/j.mpsur.2019.12.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263931919302649.

[3] D. B. Whelan, M. Bhandari, D. Stephen, *et al.*, "Development of the radiographic union score for tibial fractures for the assessment of tibial fracture healing after intramedullary fixation," *Journal of Trauma and Acute Care Surgery*, vol. 68, no. 3, 2010, ISSN: 2163-0755. [Online]. Available: https://journals.lww.com/jtrauma/Fulltext/2010/03000/Development_of_the_Radiographic_Union_Score_for.24.aspx.

[4] J. Nicholson, N. Makaram, A. Simpson, and J. Keating, "Fracture nonunion in long bones: A literature review of risk factors and surgical management," *Injury*, vol. 52, S3–S11, 2021, ISSN: 0020-1383. DOI: https://doi.org/10.1016/j.injury.2020.11.029. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020138320309554.

[5] P. Panchoo, M. Laubscher, M. Held, *et al.*, "Radiographic union score for tibia (rust) scoring system in adult diaphyseal femoral fractures treated with intramedullary nailing: An assessment of interobserver and intraobserver reliability," *European Journal of Orthopaedic Surgery & Traumatology*, vol. 32, no. 8, pp. 1555–1559, Dec. 1, 2022, ISSN: 1432-1068. DOI: 10.1007/s00590-021-03134-6. [Online]. Available: https://doi.org/10.1007/s00590-021-03134-6.

[6] M. E. Cooke, A. I. Hussein, K. E. Lybrand, *et al.*, "Correlation between rust assessments of fracture healing to structural and biomechanical properties," *Journal of Orthopaedic Research*, vol. 36, no. 3, pp. 945–953, 2018. DOI: https://doi.org/10.1002/jor.23710. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jor.23710. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jor.23710.

[7] E. Debuka, N. S. Kushwaha, D. Kumar, A. Singh, and V. Sharma, "Rust score–an adequate rehabilitation guide for diaphyseal femur fractures managed by tens," *Journal of Clinical Orthopaedics and Trauma*, vol. 10, no. 5, pp. 922–927, Sep. 1, 2019, ISSN: 0976-5662. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0976566218302054.

[8] J. Irvin, P. Rajpurkar, M. Ko, *et al.*, *Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison*, 2019. DOI: 10.48550/ARXIV.1901.07031. [Online]. Available: https://arxiv.org/abs/1901.07031.

[9] S. University. "Lera — lower extremity radiographs." en-us, Center for Artificial Intelligence in Medicine & Imaging. (2014), [Online]. Available: https://aimi.stanford.edu/lera-lower-extremity-radiographs (visited on 11/13/2022).

[10] P. Rajpurkar, J. Irvin, A. Bagul, *et al.*, *Mura: Large dataset for abnormality detection in musculoskeletal radiographs*, 2017. DOI: 10.48550/ARXIV.1712.06957. [Online]. Available: https://arxiv.org/abs/1712.06957.

[11] S. J. Adams, R. D. E. Henderson, X. Yi, and P. Babyn, "Artificial intelligence solutions for analysis of x-ray images," *Canadian Association of Radiologists Journal*, vol. 72, no. 1, pp. 60–72, Feb. 1, 2021, ISSN: 0846-5371. DOI: 10.1177/0846537120941671. [Online]. Available: https://doi.org/10.1177/0846537120941671.

[12] L. Tanzi, E. Vezzetti, R. Moreno, and S. Moos, "X-ray bone fracture classification using deep learning: A baseline for designing a reliable approach," *Applied Sciences*, vol. 10, no. 4, 2020, ISSN: 2076-3417. DOI: 10.3390/app10041507. [Online]. Available: https://www.mdpi.com/2076-3417/10/4/1507.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015. arXiv: 1512.00567. [Online]. Available: http://arxiv.org/abs/1512.00567.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," pp. 248–255, 2009.

[15] X. Mei, Z. Liu, P. M. Robson, *et al.*, "Radimagenet: An open radiologic deep learning research dataset for effective transfer learning," *Radiology: Artificial Intelligence*, vol. 0, no. ja, e210315, 0. DOI: 10.1148/ryai.210315. eprint: https://doi.org/10.1148/ryai.210315. [Online]. Available: https://doi.org/10.1148/ryai.210315.

[16] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, no. 8, pp. 500–510, Aug. 1, 2018, ISSN: 1474-1768. DOI: 10.1038/s41568-018-0016-5. [Online]. Available: https://doi.org/10.1038/s41568-018-0016-5.

[17] Y. Cao, H. Wang, M. Moradi, P. Prasanna, and T. F. Syeda-Mahmood, "Fracture detection in x-ray images through stacked random forests feature fusion," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 2015, pp. 801–805. DOI: 10.1109/ISBI.2015.7163993.

[18] K. Dimililer, "Ibfds: Intelligent bone fracture detection system," *Procedia Computer Science*, vol. 120, pp. 260–267, 2017, 9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW 2017, 22-23 August 2017, Budapest, Hungary, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2017.11.237. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050917324493.

[19] R. Lindsey, A. Daluiski, S. Chopra, *et al.*, "Deep neural network improves fracture detection by clinicians," *Proceedings of the National Academy of Sciences*, vol. 115, no. 45, pp. 11 591–11 596, 2018. DOI: 10.1073/pnas.1806905115. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1806905115. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1806905115.

[20] D. H. Kim and T. MacKinnon, "Artificial intelligence in fracture detection: Transfer learning from deep convolutional neural networks," *Clinical Radiology*, vol. 73, no. 5, pp. 439–445, May 1, 2018, ISSN: 0009-9260. DOI: 10.1016/j.crad.2017.11.015. [Online]. Available: https://doi.org/10.1016/j.crad.2017.11.015.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 1, 2016.

[22] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2017.07.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841517301135.

[23] P. Kora, C. P. Ooi, O. Faust, *et al.*, "Transfer learning techniques for medical image analysis: A review," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 1, pp. 79–107, 2022, ISSN: 0208-5216. DOI: https://doi.org/10.1016/j.bbe.2021.11.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0208521621001297.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 1, 2017, ISSN: 0001-0782. DOI: 10.1145/3065386. [Online]. Available: https://doi.org/10.1145/3065386.

[25] F. Schneider, L. Balles, and P. Hennig, *Deepobs: A deep learning optimizer benchmark suite*, 2019. DOI: 10.48550/ARXIV.1903.05499. [Online]. Available: https://arxiv.org/abs/1903.05499.

[26] R. M. Schmidt, F. Schneider, and P. Hennig, "Descending through a crowded valley - benchmarking deep learning optimizers," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, Jul. 18, 2021, pp. 9367–9376. [Online]. Available: https://proceedings.mlr.press/v139/schmidt21a.html.

[27] W. T. Obremskey, P. Tornetta 3rd, J. Luly, *et al.*, "Outcomes of patients with large versus small bone defects in open tibia fractures treated with an intramedullary nail: A descriptive analysis of a multicenter retrospective study," en, *J Orthop Trauma*, vol. 36, no. 8, pp. 388–393, Aug. 2022.

[28] Major Extremity Trauma Research Consortium (METRC), "Outcomes following severe distal tibial, ankle, and/or Mid/Hindfoot trauma: Comparison of limb salvage and transtibial amputation (OUTLET)," en, *J Bone Joint Surg Am*, vol. 103, no. 17, pp. 1588–1597, Sep. 2021.

[29] R. C. Castillo, S. N. Raja, K. P. Frey, *et al.*, "Improving pain management and Long-Term outcomes following High-Energy orthopaedic trauma (pain study)," en, *J Orthop Trauma*, vol. 31 Suppl 1, S71–S77, Apr. 2017.

[30] A. Leroux, K. P. Frey, C. M. Crainiceanu, *et al.*, "Defining incidence of acute compartment syndrome in the research setting: A proposed method from the PACS study," en, *J Orthop Trauma*, vol. 36, no. Suppl 1, S26–S32, Jan. 2022.

[31] P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J. G. Conde, "Research electronic data capture (redcap)—a metadata-driven methodology and workflow process for providing translational research informatics support," *Journal of Biomedical Informatics*, vol. 42, no. 2, pp. 377–381, 2009, ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2008.08.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046408001226.

[32] S. Z. Hong. "Redcap branching logic parser, Python library to automatically parse redcap's branching logic." en-us. version 0.1.0. (Jan. 2023), [Online]. Available: https://github.com/metrc/redcap-branch-parser (visited on 01/18/2023).

[33] S. S. Burns, A. Browne, G. N. Davis, S. L. Rimrodt, and L. E. Cutting. "Pycap api package." (2023), [Online]. Available: https://github.com/redcap-tools/PyCap (visited on 01/18/2023).

[34] M. M. Bejani and M. Ghatee, "A systematic review on overfitting control in shallow and deep neural networks," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 6391–6438, Dec. 1, 2021, ISSN: 1573-7462. DOI: 10.1007/s10462-021-09975-1. [Online]. Available: https://doi.org/10.1007/s10462-021-09975-1.

[35] L. Perez and J. Wang, *The effectiveness of data augmentation in image classification using deep learning*, 2017. DOI: 10.48550/ARXIV.1712.04621. [Online]. Available: https://arxiv.org/abs/1712.04621.

[36] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018, pp. 1542–1547. DOI: 10.1109/SSCI.2018.8628742.

[37] D. B. Whelan, M. Bhandari, D. Stephen, *et al.*, "Development of the radiographic union score for tibial fractures for the assessment of tibial fracture healing after intramedullary fixation," *Journal of Trauma and Acute Care Surgery*, vol. 68, no. 3, 2010, ISSN: 2163-0755. [Online]. Available: https://journals.lww.com/jtrauma/Fulltext/2010/03000/Development_of_the_Radiographic_Union_Score_for.24.aspx.

[38] C. Court-Brown, S. Rimmer, U. Prakash, and M. McQueen, "The epidemiology of open long bone fractures," *Injury*, vol. 29, no. 7, pp. 529–534, 1998, ISSN: 0020-1383. DOI: https://doi.org/10.1016/S0020-1383(98)00125-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020138398001259.

[39] D. Arpit, S. Jastrzębski, N. Ballas, *et al.*, *A closer look at memorization in deep networks*, 2017. DOI: 10.48550/ARXIV.1706.05394. [Online]. Available: https://arxiv.org/abs/1706.05394.

[40] C. M. Jones, Q. D. Buchlak, L. Oakden-Rayner, *et al.*, "Chest radiographs and machine learning – past, present and future," *Journal of Medical Imaging and Radiation Oncology*, vol. 65, no. 5, pp. 538–544, Aug. 1, 2021, ISSN: 1754-9477. DOI: 10.1111/1754-9485.13274. [Online]. Available: https://doi.org/10.1111/1754-9485.13274.

[41] M. Reyes, R. Meier, S. Pereira, *et al.*, "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," *Radiology: Artificial Intelligence*, vol. 2, no. 3, e190043, May 1, 2020. DOI: 10.1148/ryai.2020190043. [Online]. Available: https://doi.org/10.1148/ryai.2020190043.

[42] C. Tzioupis and P. V. Giannoudis, "Prevalence of long-bone non-unions," *Injury*, vol. 38, S3–S9, 2007, Management of Long-Bone Non-Unions, ISSN: 0020-1383. DOI: https://doi.org/10.1016/S0020-1383(07)80003-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020138307800039.

Appendix A

Additional Materials

A.1 Project Proposal Presentation

Project proposal presentation given on 2022-01-11, at the Johns Hopkins University, East Baltimore Campus.

Evaluating Fracture Healing with Artificial Intelligence

Using Transfer Learning to Predict RUST Scores from Radiographs.
A Major Extremity Trauma Research Consortium Project.

Shen Zhou Hong <shong@jhu.edu>

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH
METRC Major Extremity Trauma and Rehabilitation Consortium
Goldsmiths UNIVERSITY OF LONDON
UNIVERSITY OF LONDON

Introduction and Table of Contents

- Table of Contents:
  - §1. Background Information
  - §2. Project Design, Methodology, & Endpoints
  - §3. Current Progress, Roadmap, & Challenges
  - Q&A Session.

Background Information

Using AI to Evaluate Fracture Healing

- We want to develop an AI that can infer RUST scores from radiographs.
- RUST: Radiographic Union Score for Tibial Fractures. Also used in other long bone fractures.
- RUST measures the progression of fracture healing via callus formation and bridging.

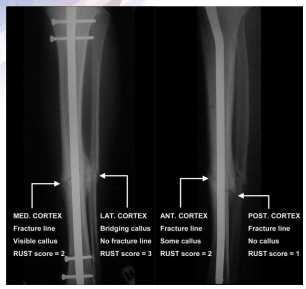


Image source: D. B. Whelan, M. Bhandari, D. Stephen, et al. 2010

Why is this project useful?

- The AI model can automate the analysis of archived study data.
- RUST scores have good biomechanical correlation, can serve as guidance for rehab.
- Project is novel: prior work in AI-radiography mainly focused on fracture detection.

Prior Research in AI-Radiography

- MURA: Anomaly detection. 40.5k radiographs, achieved AUROC of 0.929
- Lindsey et al: Anomaly detection & localization. 31.0k radiographs, achieved AUROC of 0.967
- Kim et MacKinnon: 1.3k radiographs, achieved AUROC of 0.954

Why has nobody done this before?

- Most individual institutions do not have access to datasets large enough to perform training.
- Most large radiography datasets do not have specialized, *evidence-based, adjudicated* labels.
- Most AI research organizations do not have access to medical research organizations.

METRC is uniquely positioned to conduct this research.

METRC
Major Extremity Trauma and Rehabilitation Consortium

Design, Methodology, & Endpoints

AI Models versus Statistical Models

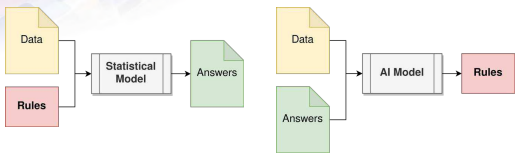


Image source: METRC 2022

METRC Radiography Datasets

- RetroDEFECT: 741 radiographs
- OUTLET: 707 radiographs
- PAIN: 370 radiographs
- PACS: 195 radiographs
- Total: ~2,013 radiographs
- Our study architecture is data-constrained.

Model Architecture

- *Transfer learning*: train an AI model on a large, general-purpose dataset. Then *fine-tune* on the smaller, task-specific dataset.
- Data Augmentation & K-Fold Validation
- We will use ImageNet and MURA as base models for transfer-learning.

Anatomy of a Deep Neural Network

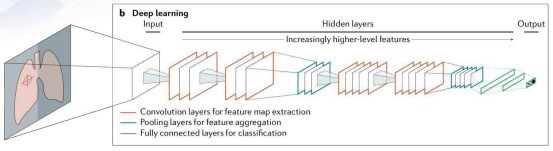


Image source: Hosny, Parmar, Quackenbush et al. 2018

Endpoints

- Model performance measured via AUROC: Area Under Curve (of) Receiver Operating Characteristic.
- Endpoint 1: Model AUROC > 0.50
- Endpoint 2: Model AUROC > "naive" ConvNet
- Endpoint 3: Model AUROC > 0.75

Progress, Roadmap, & Challenges

Roadmap and Current Progress

- 2022-11-15: Complete background research
- 2022-12-01: Study and DNN architecture design
- 2022-12-12: Initial exploration of METRC datasets
- 2023-01-02: Preprocessing of METRC radiographs. **(We are here)**
- 2023-01-31: Initial model development
- 2023-02-27: Fine-tuning, hyperparameter search.
- 2023-03-31: **Deadline: Impl. & Analysis**
- 2023-05-02: **Deadline: Poster Presentation at UoL Goldsmiths**
- 2023-05-12: **Deadline: Submission to UoL Goldsmiths**

Challenges

- Dataset validation and "cleaning."
- Combating model overfitting.
- Programmatically parsing data from REDCap.

REDCap Libraries and Tools

- REDCap Branch Parser: https://github.com/metrc/redcap-branch-parser
- REDCap Schema: https://github.com/metrc/redcapschema

Future Pathways

- Heat-maps and "Explainable AI"
- Further assessment of model performance with human orthopedic specialists/surgeons
- Collect more RUST data. Train a more robust model, conduct serious evaluations of using AI as a diagnostic tool.

The End. Thank you for your attention!