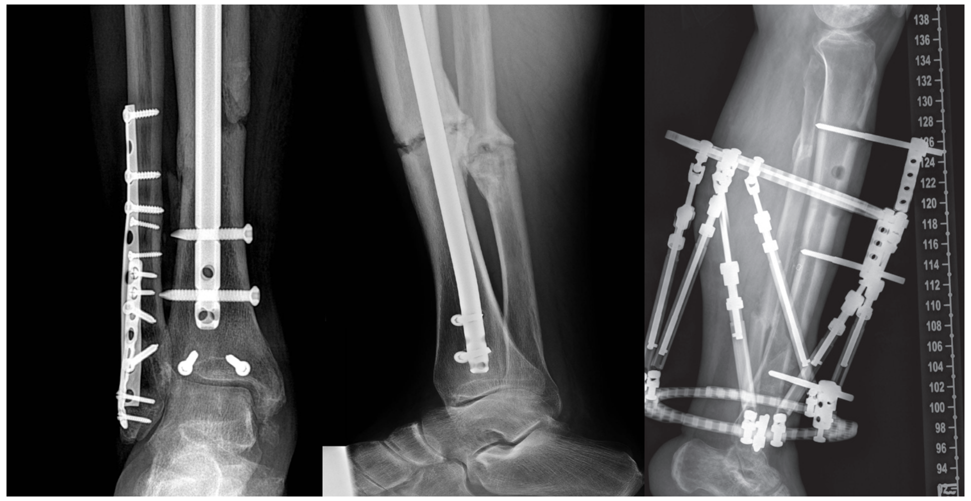
Using Transfer-Learning to Predict RUST Scores in X-Rays

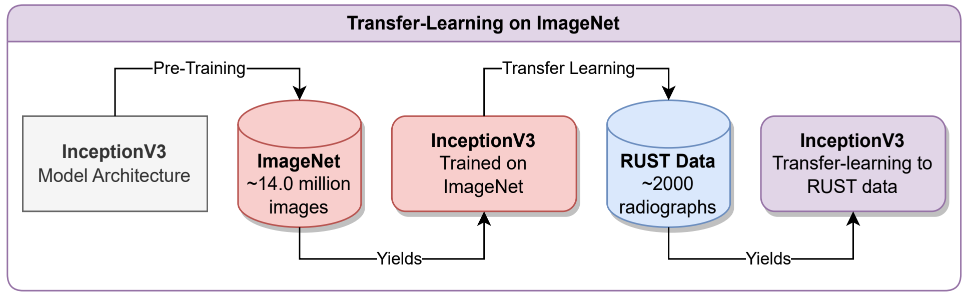
METRC
JOHNS HOPKINS BLOOMBERG SCHOOL *of* PUBLIC HEALTH
Goldsmiths UNIVERSITY OF LONDON

Shen Zhou Hong. Supervisor: Georgios Mastorakis, Renan Castillo



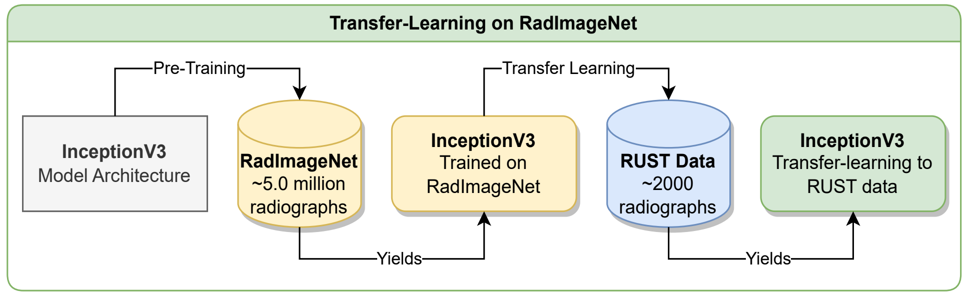
Introduction

Long-bone fractures are serious injuries which require longterm care and rehabilitation. After fixation, a fracture must be monitored for callus-formation, bridging, & union. Whelan et al's *Radiographic Union Score for Tibial Fractures* (RUST) is a 12-point clinical instrument used to assess union from the antereoposterior & lateral radiographs of a fracture. However scoring radiographs for RUST is a tricky process, requiring the attention of a trained orthopaedic clinician, potentially increasing workloads. **Is it possible to train an AI model to automatically infer RUST scores from radiographs?** This study explores the practicality of using transfer-learning as a technique to build an AI model on a limited dataset, to automatically predict RUST scores from radiographs of fractures.



Transfer-Learning on ImageNet

Pre-Training
Transfer Learning
InceptionV3 Model Architecture
ImageNet ~14.0 million images
InceptionV3 Trained on ImageNet
RUST Data ~2000 radiographs
InceptionV3 Transfer-learning to RUST data
Yields
Yields



Transfer-Learning on RadImageNet

Pre-Training
Transfer Learning
InceptionV3 Model Architecture
RadImageNet ~5.0 million radiographs
InceptionV3 Trained on RadImageNet
RUST Data ~2000 radiographs
InceptionV3 Transfer-learning to RUST data
Yields
Yields

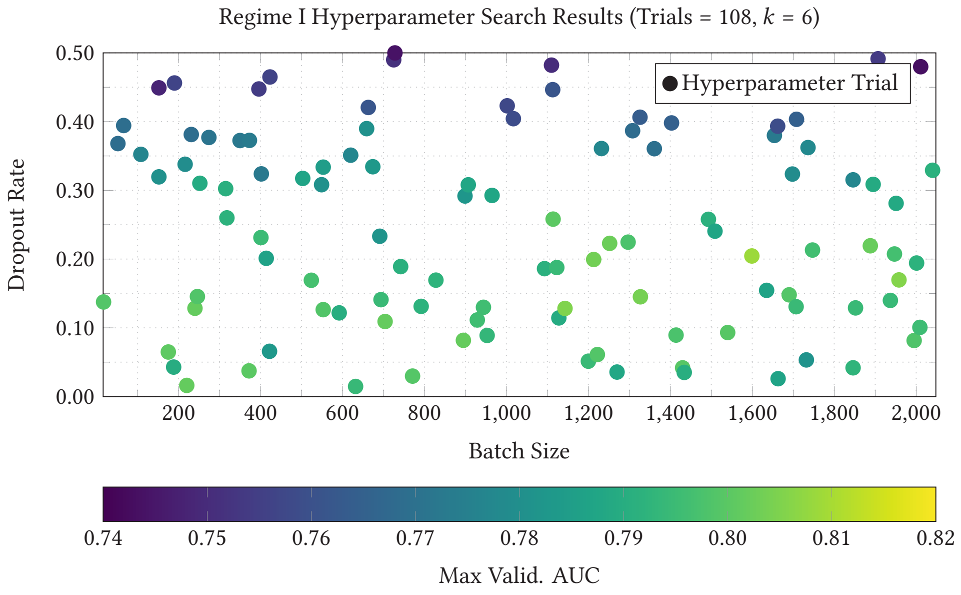
Methodology & Datasets

In partnership with the Major Extremity Trauma Research Consortium (METRC), a dataset was assembled of 3900 radiographs and labels. A key challenge in this project is dealing with the limitations of a small dataset. Small models with few parameters have limited ability to generalise, while large models with many parameters struggle to converge on limited data. To mitigate this issue, **we use the technique of transfer learning to train an AI model first on a large, general-purpose dataset**, before transfer-learning said model on our smaller radiography dataset.

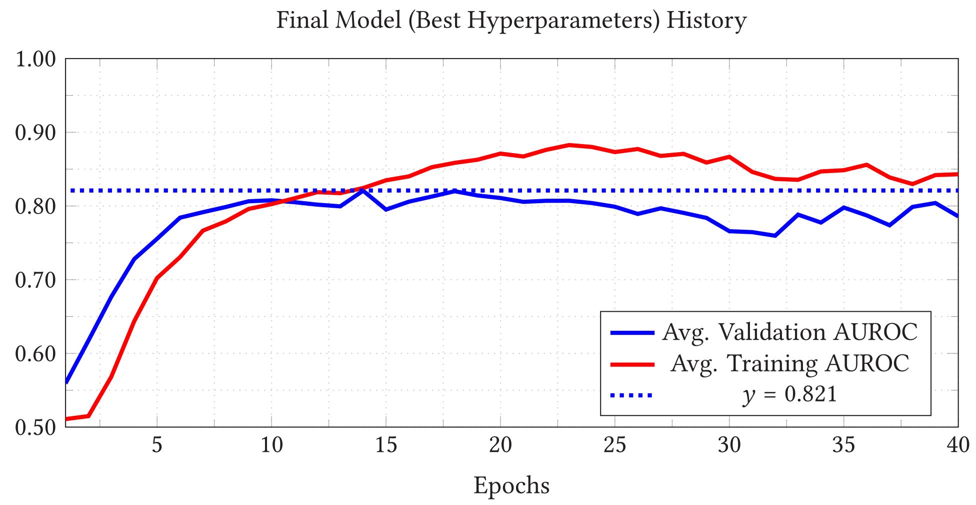
RUST labels are one-hot encoded into 18-value sparse vectors, and **model performance is assessed via the AUC** (Area Under Curve) of the precision/recall graph. We use the **InceptionV3** CNN architecture as our base model with the Adam optimizer. We evaluate two different pre-training datasets for our base model: Stanford's **ImageNet** dataset (14.0 million images), and Cornell's **RadImageNet** dataset (5.0 million images). RadImageNet is a smaller dataset but offers potential advantages by being domain-specific to our task. With the better-performing model, we proceed to search for optimal hyperparameters, iterating on *batch size*, *dropout rate*, *learning rate*, and *epsilon*. This hyperparameter search is performed in two regimes:

- Regime I: Random Search on *Batch Size* & *Dropout* (see bottom fig.)
- Regime II: Grid Search on *Learning Rate* & *Epsilon*

Each trial in the search space is evaluated using k-fold cross-validation with k = 6. After finding the best set of hyperparameters, an evaluation of the final model is performed using k-fold cross-validation, and then the best-performing fold is evaluated on the hold-out test set.



Regime I Hyperparameter Search Results (Trials = 108, k = 6)



Final Model (Best Hyperparameters) History

Key Results

By evaluating both ImageNet weights and RadImageNet, we were able to find that ImageNet weights had slightly better performance on our baseline model. Using ImageNet and the best-performing hyperparameters, we were able to yield a Final Model with a k-fold cross-validated validation AUC of 0.821 (see graph above). This best-performing model was then evaluated upon our hold-out test set, yielding a **AUC of 0.891.** The performance achieved by this model is promising, especially given the small size of the dataset that it is trained on.

The **precision** achieved over the one-hot encoded RUST labels on the hold-out test set is 0.880. However, the **recall** is only 0.409, highlighting difficulties with the highly imbalanced nature of our dataset. This may indicate that the model performance degrades with edge cases.

Discussion

As an exploration into a novel *multi-class, multi-label* image classification task with a small, domain-specific dataset, this project validates the feasibility of using transfer-learning as a technique in the domain of medical imaging. Medical image processing tasks present unique challenges to AI scientists: they are characterised by their small datasets, expensive labelling, & unbalanced class distributions. This project addresses these challenges and yields persuasive results, showing that general-purpose architectures like ImageNet can be applied to highly domain-specific tasks.

As radiography places an increasing role not just in fixation, but also in rehabilitation, the use of AI can become a powerful tool in alleviating workloads for clinicians, and reducing healthcare cost for patients. By developing AI models which can accomplish menial tasks like scoring radiographs, we hope to empower clinicians in their practice, ultimately helping *to fulfil the promise of medicine.*

Repository & Data URL: https://github.com/ShenZhouHong/radiography-ai-project

Project Report: https://s.hong.io/project.pdf