Musculoskeletal Radiograph Classification Using Deep Learning

A diagnostic AI system for anomaly detection

Álvaro Sánchez Martín, March 2025

10cm

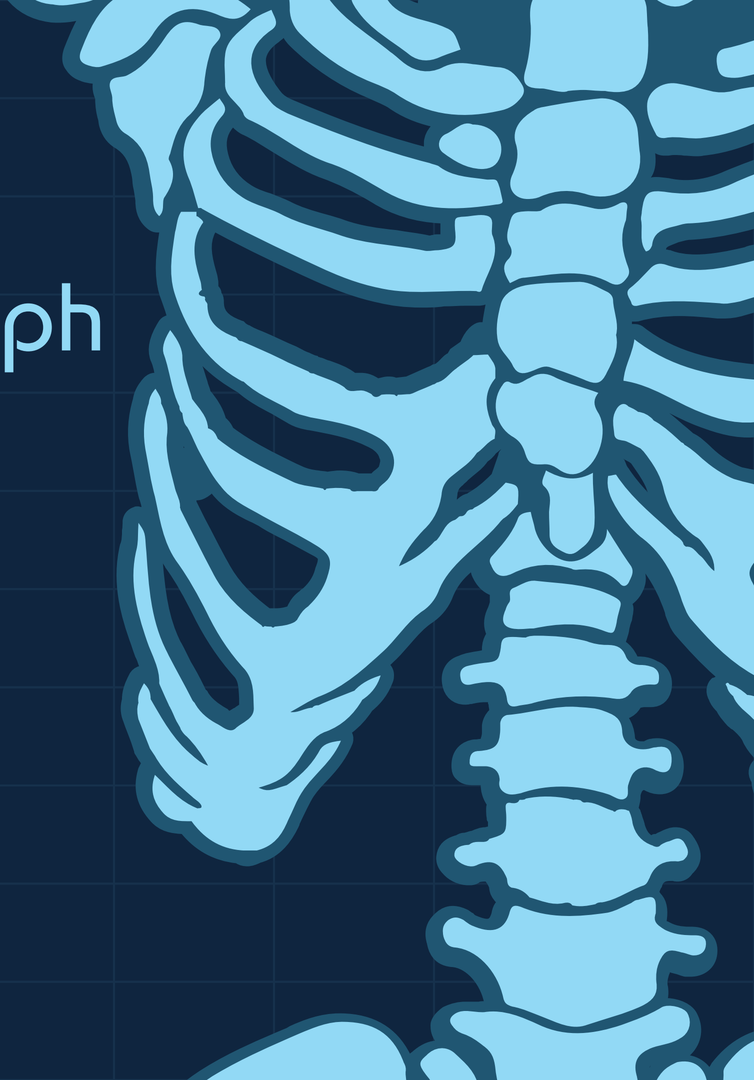


Table of contents

10cm

01 The disease

Business problem:

Automating Anomaly Detection in MSK Radiographs

Diagnostic errors occur mostly when radiologists exceed their average daily production

Marty Stempniak | August 11, 2023 | Radiology Business | Leadership

Work overload and diagnostic errors in radiology

Ömer Kasalak*, Haider Alnahwi, Romy Toxopeus, Jan P Pennings, Derya Yakar, Thomas C Kwee

*Corresponding author for this work

Basic and Translational Research and Imaging Methodology Development in Groningen (BRIDGE)

iew

Using AI to Improve Radiologist Performance in Detection of Abnormalities on Chest Radiographs

Souhail Bennani, Nor-Eddine Regnard, Jeanne Ventre, Louis Lassalle, Toan Nguyen, Alexis Ducarouge, Lucas Dargent, Enora Guillo, Elodie Gouhier, Sophie-Hélène Zaimi, Emma Canniff, Cécile Malandrin, Philippe Khafagy, Hasmik Koulakian, Marie-Pierre Revel, Guillaume Chassagnon

Author Affiliations

Published Online: Dec 12 2023 | https://doi.org/10.1148/radiol.230860

10cm

02 Diagnosis

Technical framing:

Understanding the Case & the Data

Dataset: MURA* (Stanford ML Group)

Total Images: ~40,561 musculoskeletal radiographs

Total Studies: 14,863 from **12,173 patients**

Body Parts: Shoulder, Humerus, Elbow, Forearm, Wrist, Hand, Finger

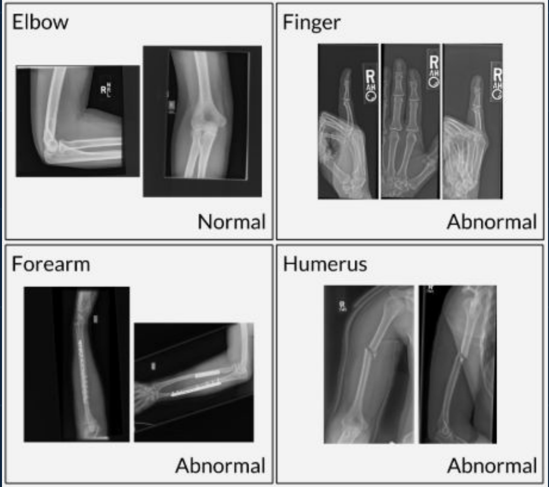
Labeling:

- **Positive (1)** = abnormal finding (fracture, degeneration, etc.)
- **Negative (0)** = normal image

Imbalance: ~58% Negative vs ~42% Positive (varies by study type)

Challenges:

-**High variability** in radiograph quality, orientation, and exposure.
-**Noisy labels:** sometimes based on study-level, not image-level.
-Medical images are visually subtle — anomalies are hard to detect.



Binary classification task:

Is this radiograph **abnormal (1)** or **normal (0)** ?

Other Challenge:

Strong **class imbalance**

10cm

03 Clinical Trials

Model Benchmarking:

Comparison of CNN architectures

MODEL	Accuracy	F1- Score	Notes
ResNet18	91%	0.91	Best Balance & speed
DenseNet121	91%	0.91	Similar to RestNet, more complex
VGG16	82%	0.82	Decent, heavier model
EfficientNet	73%	0.71	Biased to negatives
MobileNet	36%	0.24	Poor anomaly detection

10cm

04 Treatment Plan

Technical framing:

Understanding the Case & the FULL DATA

Version	Main Changes	F1-score	Accuracy
Baseline	Only fc training, no tuning	0.78	78%
V5	Full model trained	0.59	63%
V6	Added class weights + LR scheduler	0.77	77%
V7	Added early stopping	0.79	79%
V8	Better LR tuning + regularization	0.80	80%
V9	+Augmentations & freeze strategy	0.80	80%

No class weighting

Consistent improvement Through fine-tuning and balancing

10cm

05 Surgical Procedure

Architecture:

What the best model looks like

1.Architecture Modification

```
resnet18.fc = nn.Linear(resnet18.fc.in_features, 2)
```

Replaced the fully connected layer to output 2 classes: normal vs. abnormal

2.Advanced Fine-Tuning: Partial Unfreezing

```
for name, child in
resnet18.named_children():
    if name in ['layer3', 'layer4', 'fc']:
        for param in child.parameters():
            param.requires_grad = True
    else:
        for param in child.parameters():
            param.requires_grad = False
```

Unlocked layer3, layer4, and fc for fine-tuning.
Earlier layers remained frozen to preserve ImageNet features.

3. Weighted Loss for Imbalance

```
criterion = nn.CrossEntropyLoss(weight=class_weights_tensor)
```

Applied inverse-frequency class weights to penalize false negatives more

10cm

05 Surgical Procedure

Architecture:

What the best model looks like

4.Regularization with Scheduler and Early Stopping

```
scheduler = ReduceLROnPlateau(optimizer, mode='min', factor=0.5, patience=5)
```

Reduced learning rate automatically if validation loss plateaued

```
if epochs_no_improve >= EARLYSTOP_PATIENCE:
    break
```

Stopped training early to prevent overfitting

5.Clinically Safe Augmentations

Simulated real-world variability in radiograph positioning without distorting key features

```
transforms.Compose([
    transforms.RandomRotation(15),
    transforms.ColorJitter(0.2, 0.2),
    transforms.RandomResizedCrop(IMAGE_SIZE, scale=(0.85, 1.0)),
    ...
])
```

10cm

06

What comes after treatment?:

Conclusions and next step

Key Takeaways

- **80% F1-score and accuracy** in final model
- Optimizations that worked: **partial unfreezing, class weighting, LR scheduling, safe augmentations**
- Learned from **imbalanced, noisy, and subtle data**
- AI as an **assistive tool**, not a replacement — great for **screening** use cases

Future Improvements

- **Per-body-part performance** (some bones may be harder to classify)
- **Explainability tools** like Grad-CAM or saliency maps
- More **augmentation testing** (blurring, low contrast...)
- **Deployable prototype** (Flask/Gradio for demo or clinical API)
- Feedback loop from radiologists to **continually retrain the model**

10cm

Thanks!

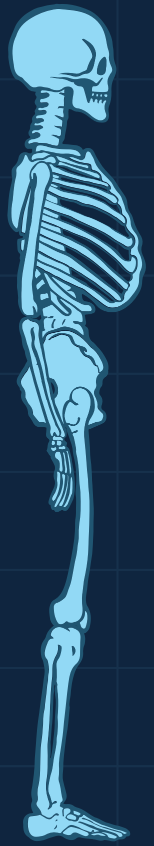
Project available on GitHub

//github.com/alvarosmms/ML_MURA_CLASSIFICATION

Feel free to reach out!

linkedin.com/in/álvaro-sánchez-martín

alvarosanchez91@hotmail.com



10cm