ScaphX: Development of an AI-Assisted Tool for Scaphoid Fracture Detection on Wrist Radiographs

Aoife Fox, Mike Woodward, Dijana Vilic, Anil Mistry | Clinical Scientific Computing Team, Guy's and St. Thomas' Hospital

Background / Clinical Problem

Scaphoid fractures are common in young adults, accounting for 60–70% of carpal bone fractures, and can lead to serious complications if not promptly diagnosed. X-rays have limited sensitivity (60–80%), so patients at Guy's and St Thomas' NHS Foundation Trust (GSTT), with negative X-rays are routinely referred for MRI. While this ensures fractures aren't missed, it increases resource use. AI tools could help identify more fractures directly on X-ray, enabling faster diagnosis and reducing the need for MRI. This could improve efficiency, streamline care pathways, and optimise imaging resources.

Project Aims

- To develop an AI-assisted tool (ScaphX) to support the detection of scaphoid fractures on plain radiographs.
- To reduce reliance on MRI by improving the diagnostic yield of X-ray imaging in patients with suspected scaphoid fractures.
- To evaluate the model's impact on diagnostic accuracy, clinician confidence, and imaging utilisation in the emergency department setting.

System Overview

ScaphX is designed as a three-step pipeline :

1. Scaphoid Detection – Automatically localises and crops the scaphoid region from the radiograph.

2. Apparent Fracture Classification – Identifies fractures that are visible on X-ray and typically recognised by experienced clinicians.

3. Occult Fracture Classification – Detects subtler features of fracture that may require confirmatory MRI in current practice.

The system is under development. Threshold tuning is being used to maximise sensitivity, prioritising the detection of true fractures. Future integration into clinical workflows is planned, pending successful evaluation.

Development Timeline

Phase 1 – Initial Model (2023)

- A two-step model (scaphoid detection + binary classifier) was developed using X-rays that were labelled based on MRI data.
- User testing with ED clinicians was conducted but revealed limited performance and reduced clinician confidence.

Phase 2 – Model Redevelopment (2024)

- The dataset was expanded and restructured to train two distinct models:
  - One using MRI-confirmed fractures (to detect occult injuries)
  - One using radiologist-confirmed fractures visible on X-ray (to detect apparent fractures)
- Image quality, labelling rigour, and augmentation techniques were improved.
- Hyperparameter tuning and threshold calibration were introduced to optimise performance.

Current Status (2025)

- A new model trained on MRI-labelled data is under internal validation.
- The next step is to train the third pipeline component on X-ray-visible fractures.

Model Training Pipeline

Two datasets are being used to train separate classifiers:

- **MRI-labelled dataset**: 1,796 radiographs (359 positive, 1,438 negative)
- **X-ray report-labelled dataset**: 8,114 radiographs (1,058 positive, 7,056 negative)

This enables detection of both **occult** (MRI-confirmed) and **apparent** (X-ray-visible) fractures. Scaphoid regions are automatically cropped. Standard X-ray views (PA, oblique, angled) are included. Post-operative images, duplicates, and incomplete studies are excluded.
Augmentations include geometric (rotation, flipping, zoom), photometric (contrast, noise), CutMix and coarse dropout to enhance generalisability.
Models are being trained using the CSC-MLOps framework with PyTorch Lightning.
Hyperparameters are tuned using Optuna.
The best-performing architecture identified is a pretrained **ConvNeXtV2** model. Post-training threshold tuning is conducted to optimise sensitivity, reflecting the clinical aim of reducing missed fractures.

Results

A CNN was trained on the MRI-labelled dataset (n = 1,796) and evaluated on a held-out validation set.

Performance (pre-threshold tuning) (*see Figure 1):*

- **Sensitivity:** 29%
- **Specificity:** 96%
- **False negative rate:** 71%
- **False positive rate:** 3.6%

The model demonstrated high specificity but low sensitivity at the default threshold of 0.5, missing a substantial number of true fractures. This highlights the need for **threshold tuning** to prioritise sensitivity in line with clinical priorities.

Training optimisation:

Advanced augmentation techniques such as **CutMix** were used to improve generalisability. This method combines portions of two images and blends their labels (e.g. label: [0.95, 0.05]), helping the model learn from partial or ambiguous features. (*see Figure 2)*

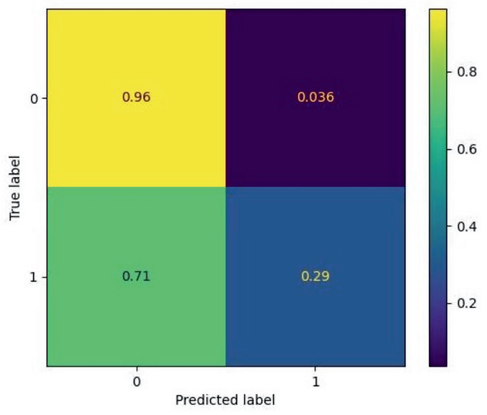
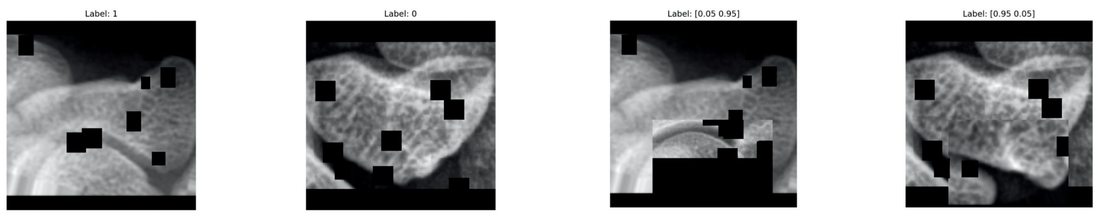


Figure 1 (left): Confusion matrix showing model performance on MRI-labelled data prior to threshold adjustment.

Figure 2 (below): Example of pre-CutMix (left) and post-Cutmix (right) augmentation



Next Steps

- **Train the third pipeline component** to detect *apparent* fractures using X-rays labelled by radiology report, expanding the model's ability to identify clearly visible injuries.
- **Reassess overall model performance** across the three-stage pipeline, with an emphasis on improving sensitivity and maintaining interpretability through heatmaps.
- **Conduct prospective testing** at Guy's and St Thomas' NHS Foundation Trust using real-world patient data, comparing model output against clinical decisions and MRI outcomes.
- **Refine model generalisability and fairness** by evaluating performance across subgroups and ensuring consistent output across demographics.
- **Explore integration pathways** for potential deployment into clinical workflows via PACS and AI deployment engine, following usability testing and further validation.

Discussion

- ScaphX demonstrates the feasibility of developing a locally tailored AI tool for scaphoid fracture detection on plain radiographs. By separating model training into distinct pipelines for occult and apparent fractures, this project aims to improve diagnostic performance while addressing the limitations of existing tools.
- Early results show strong specificity but suboptimal sensitivity at default thresholds, reinforcing the importance of threshold tuning in clinical AI applications. The planned addition of a dedicated model for apparent fractures is expected to enhance real-world performance.
- Advanced data augmentation (e.g. CutMix) and rigorous preprocessing have strengthened model generalisability, and the use of in-house infrastructure has enabled rapid iteration.
- Ongoing development will focus on prospective evaluation, fairness monitoring, and clinician-in-the-loop validation—critical steps toward responsible deployment in the emergency department setting.

References

1. Kraus M, et al. *Artificial intelligence for X-ray scaphoid fracture detection: a systematic review*. *Eur Radiol*. 2024;34(7):4341–51. 2. Yoon AP, et al. *Development and validation of a CNN model for scaphoid fractures*. *JAMA Netw Open*. 2021;4(5):e216096. 3. Bäcker HC, et al. *Systematic review of diagnosis of clinically suspected scaphoid fractures*. *J Wrist Surg*. 2020;9(1):81–9. 4. Kuo RYL, et al. *AI in fracture detection: a systematic review and meta-analysis*. *Radiology*. 2022;304(1):50–62. 5. Miller R, et al. *AI in acute scaphoid fracture detection: a critical review*. *J Hand Surg Eur Vol*. 2025; DOI: 10.1177/17531934241312896.