External Validation of an Artificial Intelligence Model for Paediatric Wrist Fracture Detection

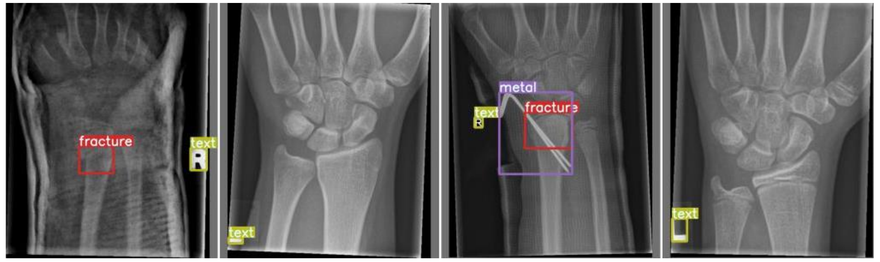
Cato Pauling, Emma Allan, Emily Ashworth, Owen Arthurs, Baris Kanber, Susan Shelmerdine

NHS Great Ormond Street Hospital for Children NHS Foundation Trust

UCL

Background

- >10% of fractures in children go undiagnosed by non-specialised Health Care Practitioners
- Very few commercially available Artificial Intelligence (AI) tools to diagnose fractures in children
- Limited paediatric data available for training AI



Aims

1. **Achieve current state-of-the-art performance with an open-source dataset and model**
2. **Apply the model to a novel, multi-centric, dataset of a British paediatric population**

Methods & Materials

DATASETS:

Training Dataset: GRAZPEDWRI-DX	External Dataset: FRACTURE-WRIST
Cleaned + pre-processed	Real-world, unprocessed data
Paediatric wrist trauma radiographs	Paediatric wrist trauma radiographs
Single centre	2 London paediatric trauma centres
20,327 images	865 images
10,643 radiographic examinations	403 radiographic examinations
6,091 patients	265 patients
Mean age 10.9 years	Mean age 10.4 ± 3.5 [s.d.] years
Age range 0.2 to 19 years	Age range 14 days to 17.2 years

- Ground truth in FRACTURE-WRIST established by consensus opinion of two radiologists

MODEL:

- YOLOv7 - Configuration and hyperparameter selection based on existing literature and experimental optimisations

EXTERNAL VALIDATION:

1. Open-source dataset split into training, validation and test subsets
2. Model trained and tested on the “internal” GRAZPEDWRI-DX dataset
3. Model tested on entire “external” FRACTURE-WRIST dataset

Results

Dataset	Class	Precision	Recall	mAP@.5	mAP@.5:.95:
GRAZPEDWRI-DX	all	0.578	0.654	0.599	0.387
	fracture	**0.879**	**0.89**	**0.919**	**0.515**
FRACTURE-WRIST	all	0.789	0.495	0.46	0.32
	fracture	**0.725**	**0.564**	**0.524**	**0.172**

- The sensitivity (recall) of the model was 89.0% on the internal test data
- When evaluating on the external data, the sensitivity decreased to 56.4% (-32.6%)



Interpretations

1. **Performance does not always transfer between datasets**
2. **Real-world images with significant padding, combined orientations, and different scanners and imaging protocols impact performance**