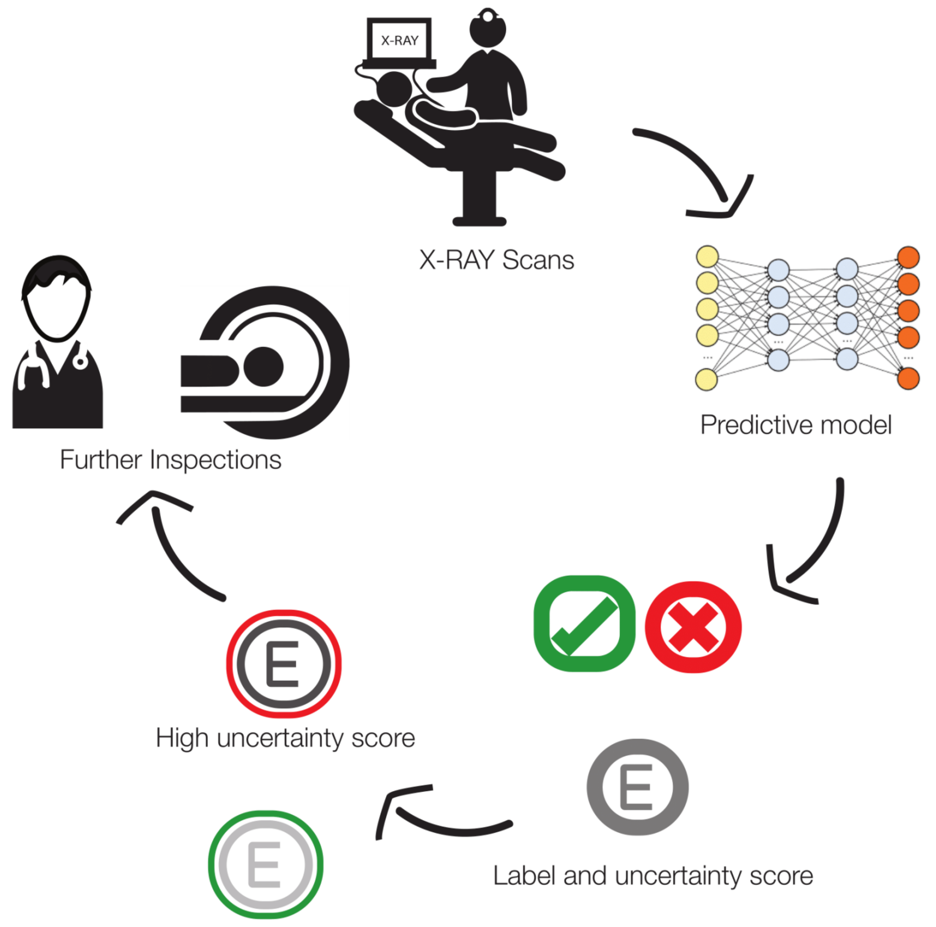
Uncertainty for Proximal Femur Fractures Classification

Mayar Lotfy, Selina Frenner, Marc Beirer, Peter Biberthaler, Shadi Albarqouni

UNIVERSITÄT BONN HELMHOLTZ MUNICH TUM

Motivation

- **Goal:** reliable score of uncertainty as quality control for automated Proximal Femur Fracture classification



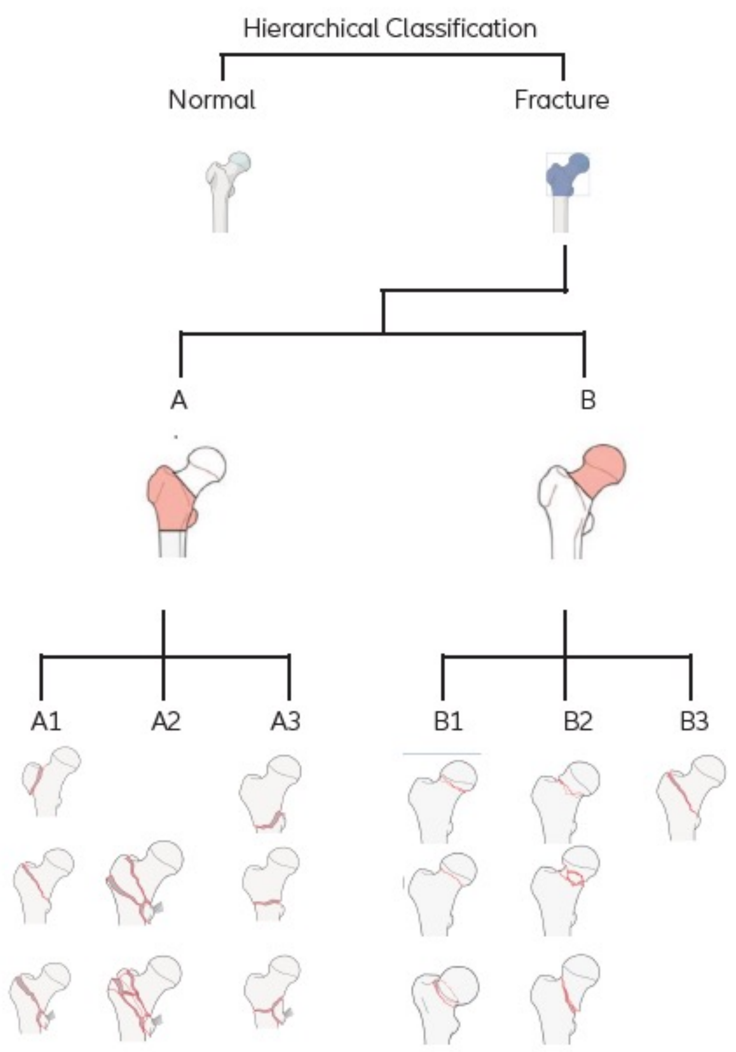
- **CADs**: achieving state-of-the–art results in diagnosis (1,2), improve accuracy of diagnosis (3), reduction of medical error and support time, cost.efficient treatment in future medicine (6)
- **Proximal femur fractures:** high incidence, early diagnosis and treatment are essential for the patient's outcome and survival (4), depends on the examiners' experience (5).

Database

- 672 patients from trauma surgery department of Klinikum Rechts der Isar, Munich, 1347 X-ray images & corresponding labels – using the work of (8) as baseline

Method

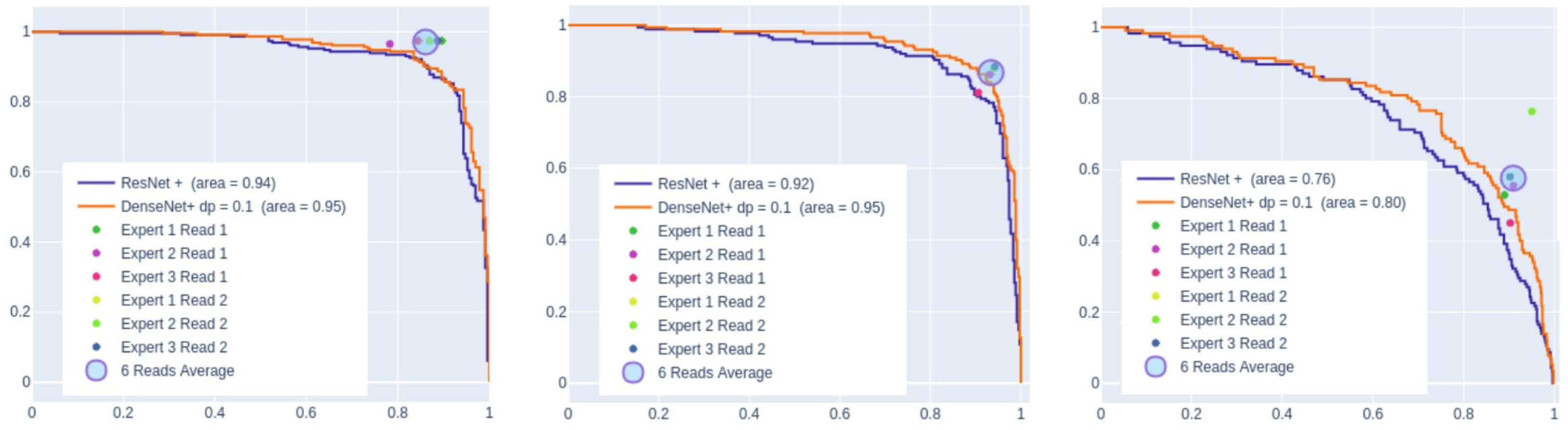
- Three different classification scenarios:
  - –C ∈ {C1, C2, C3}; C1 ⊂ {Fracture, Normal} = fracture detection scenario
  - –C2 ⊂ {A, B, Normal} = 3 classes scenario
  - –C3 ⊂ {A1, A2, A3, B1, B2, B3} = 6 classes scenario



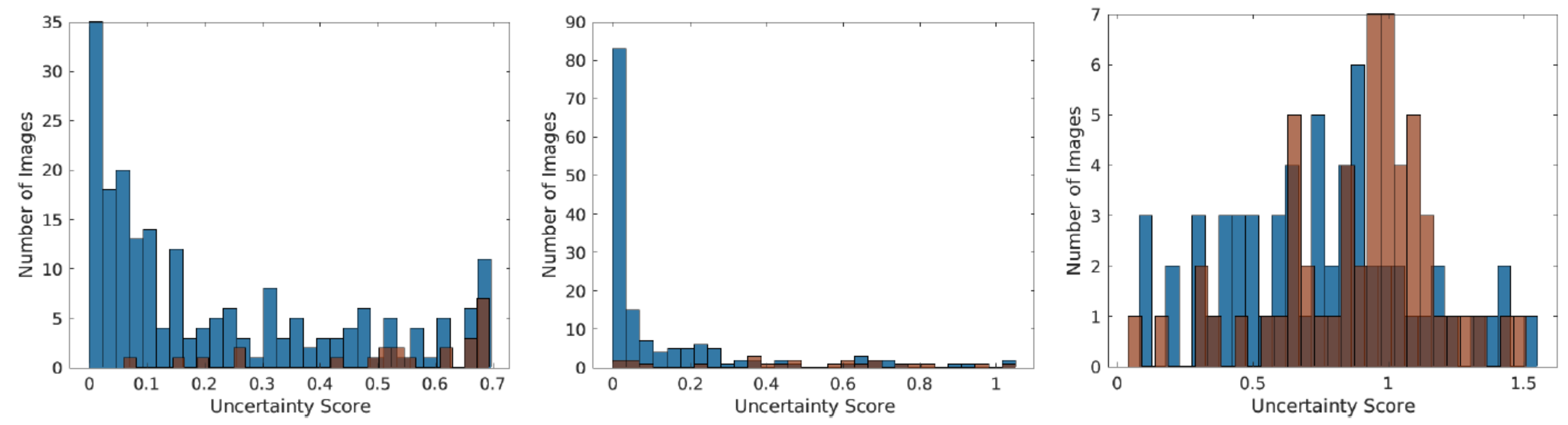
- **Ground truth**: three different experts, confirmed by senior radiologist
- **Monte Carlo dropout**, as an approximation of Bayesian Neural Networks (7) providing a quality control measure.
- ResNet adopted from (8), where the MCDO was introduced only at the last dense layer, treating the rest as a deterministic network.
- Stochastic DenseNet121 model, the dropout layers were introduced at each convolutional layer and in the transition blocks, hyper-parameters adopted from (9). 5-fold cross validation were conducted for the DenseNet models.

Experiments & Results

- Clinical experts vs. CAD system for the 3 classification scenarios



- Uncertainty coherence with miss-classication for DenseNetCE+



(d) Fract. (e) 3 classes (f) 6 classes

- Uncertainty scores on test and validation set

- **Qualitative assessment** - re-evaluation 30% of the test dataset by an independent radiologist & In-depth analysis - further annotations from three independent experts, each with two independent reads in different occasions as different shifts and lighting conditions

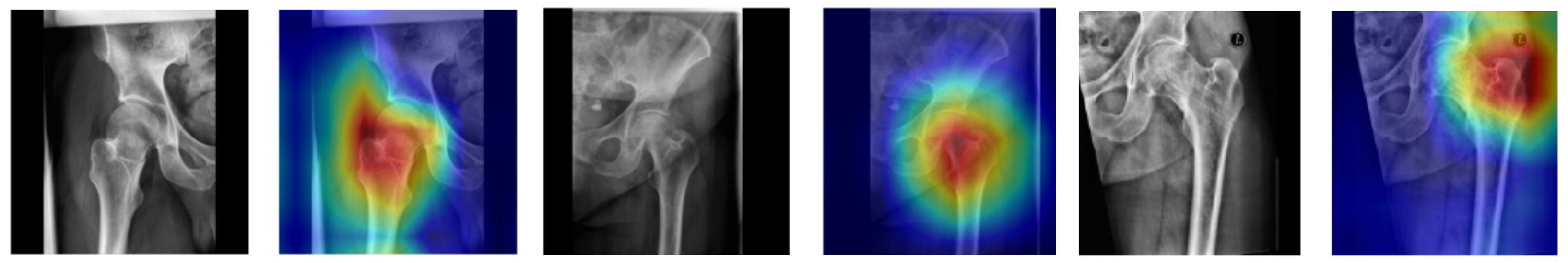


Image is clear			
App.	No	**Frac.**	No
	Exp.1	**Exp.2**	**Exp.3**
Read1	B	B	B
Read2	B	B	B
GT.	A	**Pred.**	B

Low uncertainty-missclassfied → false ground truth

Overlapping soft tissue artefacts as disturbing factor			
App.	Yes	**Frac.**	No
	Exp.1	**Exp.2**	**Exp.3**
Read1	B	N	B
Read2	N	B	B
GT.	N	**Pred.**	B

High uncertainty-missclassfied → high uncertainty among experts

Image is taken after operation Healed fracture with sclerotic transformations after screws removal			
App.	Yes	**Frac.**	Yes
	Exp.1	**Exp.2**	**Exp.3**
Read1	N	B	N
Read2	N	N	N
GT.	B	**Pred.**	B

High uncertainty-correctly classified → false ground truth

- ***Two key outcomes***:

1. Uncertainty score = reliable measure for detecting mistakes in the model performance and a valid robustness quality control.
2. Model's performance is reflected on how well and coherent is the modelling of uncertainty, i.e. ResNet+ vs. DesneNet+ 1

Conclusion

- Coherency between misclassification and uncertainty scores - high uncertainty score means high risk for error in prediction.
- Uncertainty measures mimicking the actual radiologist's uncertainty for challenging and complex examples reflected on intra- and inter-experts variability.
- Possible key element for clinical applicability of CADs

Future work

- Improving robustness of the model
- Extending the work on different datasets/ other parts of the human body

References

(1) Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. PLoS medicine, 15(11):e1002686, 2018.
(2) Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al. Deeplearning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. PLoS medicine, 15(11):e1002699, 2018.
(3) Robert Lindsey, Aaron Daluiski, Sumit Chopra, Alexander Lachapelle, Michael Mozer, Serge Sicular, Douglas Hanel, Michael Gardner, Anurag Gupta, Robert Hotchkiss, et al. Deep neural network improves fracture detection by clinicians. Proceedings of the National Academy of Sciences, 115(45):11591–11596, 2018.
(4) Jeremy D Schroeder, Sean P Turner, and Emily Buck. Hip fractures: Diagnosis and management. American Family Physician, 106(6):675–683, 2022
(5) CE Plant, C Hickson, H Hedley, NR Parsons, and ML Costa. Is it time to revisit the ao classification of fractures of the distal radius? inter-and intra-observer reliability of the ao classification. The bone & joint journal, 97(6):818–823, 2015.
(6) William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P Bradley, and Lyle J Palmer. Detecting hip fractures with radiologist-level performance using deep neural networks. arXiv preprint arXiv:1711.06504, 2017.
(7) Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In Advances in neural information processing systems, pages 5574–5584, 2017.
(8) Amelia Jimenez-Sanchez, Anees Kazi, Shadi Albarqouni, Sonja Kirchhoff, Alexandra Strater, Peter Biberthaler, Diana Mateus, and Nassir Navab. Weakly-supervised localization and classification of proximal femur fractures. arXiv preprint arXiv:1809.10692, 2018.
(9) Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700–4708, 2017.