Uncertainty for Proximal Femur Fractures Classification

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Deep Learning methods over the past years provided high-performance solutions for the medical applications. Yet, robustness and quality control is still required for clinical applicability. In this work, the uncertainty of proximal femur fracture classification,was modeled. We introduce a reliability measure to our predictive model using the Monte Carlo Dropout approach. We performed an extensive quantitative and qualitative analysis to validate the results. We further exposed the results to expert physicians in order to get feedback on the model's performance and uncertainty measures. Results demonstrate a positive correlation between the miss-classification of the model's prediction and high uncertainty scores. Additionally, the uncertainty measures are mimicking the actual radiologist's uncertainty for challenging examples reflected on intra- and inter- experts variability.

Keywords: Deep Learning, Uncertainty, Quality Control, Radiology, Proximal femur fractures.

1. Introduction

In recent years, Deep Learning was excelling at automatically recognizing complex patterns. That being the case, innovative methods are achieving state-of-the-art results in the diagnosis procedure (Rajpurkar et al., 2018; Bien et al., 2018). It has been demonstrated that when emergency medicine clinicians are provided with the assistance of automated AI models, their ability to accurately detect fractures is significantly improved (Lindsey et al., 2018). Proximal femur fractures are among the most common fractures worldwide. Early diagnosis and treatment are essential for the patient's outcome and even survival (Schroeder et al., 2022). This mostly depend on the examiners' experience (Plant et al., 2015). CAD systems could help improve the accuracy of diagnosis, reduce medical error and support time and cost-efficient treatment in future medicine (Gale et al., 2017). Still there is a need for some quality indicator to evaluate the applicability of such a system into clinical routine by evaluating not only the overall performance of the model but also the reliability of each single diagnosis. The goal of this study was to provide a reliable score of uncertainty using Monte Carlo dropout layers, the recent approximations of Bayesian Neural Networks (Kendall and Gal, 2017), as a quality control for automated proximal femur fracture classification using the AO-fracture classification system.

1.1. Methods

Given 1347 X-ray images and their corresponding labels $y \in C$, where C denotes one of the three subsets according to the different classification scenarios, $C \in \{C_1, C_2, C_3\}$; $C_1 \subset \{Fracture, Normal\}$ for the fracture detection scenario, $C_2 \subset \{A, B, Normal\}$ for the

© 2023 CC-BY 4.0, Anonymous.

3 classes scenario, and $C_3 \subset \{A1, A2, A3, B1, B2, B3\}$ for the 6 classes scenario, our aim is to build a femur fracture classification model that assigns each image a class label and uncertainty scores. The database was collected from 672 patients by the trauma surgery department of XXX. The ground truth was created by a participation of three different experts. All of the images were confirmed by one senior radiologist. For the in-depth analysis, further annotations of the testing test were taken from three independent experts, where each expert provided two independent reads in different occasions as different shifts and lighting conditions. As for the qualitative assessment of the uncertainty score, an independent radiologist re-evaluated 30% of the test dataset.

1.2. Results

ResNet model was adopted from (Jiménez-Sánchez et al., 2018), where the MCDO was introduced only at the last dense layer, treating the rest as a deterministic network. As for the stochastic DenseNet121 model, the dropout layers were introduced at each convolutional layer and in the transition blocks, where the same hyper-parameters adopted from (Huang et al., 2017). 5-fold cross validation were conducted for the DenseNet models. To evaluate the performance of our model, we compute the confusion matrix, *F1*-score and the macro average *F1*-score for each classification scenario to account for class imbalance.

The focus of the study was not to outperform the SOTA, but rather capturing the uncertainty while achieving comparable performance. The experiments were designed in a way to analyse the performance of MCDO under 1) Stochastic vs. Deterministic networks, and 2) different loss functions; namely Cross Entropy (CE), and Weighted Cross Entropy (WCE) , and Focal Loss (FL), and 3) different network architectures. Models introduced with MCDO layers achieved comparable F1-score performance to their own baseline models.

In an attempt to analyse and compare the performance of MCDO with different settings, namely, deterministic model (ResNet) and Stochastic DenseNet against the individual readings of the three experts and the majority consensus. We visualise the Receiver operating characteristic (ROC) curves 2. Overall, the performance of our stochastic DenseNet models performed similarly compared to the expert's readings.

Further, the coherence between the uncertainty scores and miss-classification was qualitatively measured. Results demonstrated that the miss-classified images mostly occur in the highly uncertain region. This is mostly apparent in cases of fracture detection and 3 classes classification. As for 6 classes classification, it shows a Gaussian distribution for the uncertainty scores with almost no coherence with miss-classification. This scenario in particular is the most challenging out of the previous cases and is aligned with the F1-score reported. This confirms two key outcomes; First, the uncertainty score is a reliable measure for detecting mistakes in the model performance and a valid robustness quality control. Second, the model's performance is reflected on how well and coherent is the modelling of uncertainty, i.e. ResNet+ vs. DesneNet+ 1

Lastly, the coherence between the calculated uncertainty scores of the test images and the uncertainty of radiologists' annotation were validated. In a like manner, the aforementioned scores with the inter- and intra-observer variability of the three independent experts who participated in providing two distinctive reads each were compared. The radiologist was asked to provide comments on the image and the fracture specifying if the classification

is an easy or challenging task, along with stating if the difficulty comes from the fracture complexity (i.e. cognitive) or from the appearance (i.e. perceptual). To this end, we expect a reflection of the high uncertainty on the appearance and the fracture difficulty, also a reflection on the disagreement and the inter-/intra- variability among the three experts.

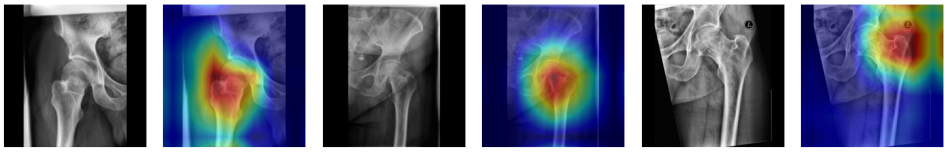

Image is clear			
App.	No	**Frac.**	No
	Exp.1	**Exp.2**	**Exp.3**
Read1	B	B	B
Read2	B	B	B
GT.	A	**Pred.**	B

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

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

Figure 1: Qualitative Assessment: 3-class, showing Low Uncertainty Miss-Classified, High Uncertainty Correctly classified, High Uncertainty Miss-classified, respectively.

1.3. Conclusion

In this work the performance of different networks for classification of proximal femur fractures was compared, adding an uncertainty score for quality control. The coherency between misclassification and uncertainty scores can be demonstrated, with DenseNet stochastic implementation showing the best alignment with misclassified cases. This indicates that a high uncertainty score means a high risk for error in prediction. Additionally, by putting the main focus on qualitative analysis, it can be shown that the uncertainty measures are mimicking the actual radiologist's uncertainty for challenging and complex examples reflected on intra- and inter- experts variability. These findings can be useful not only in scientific research, for example by being used while training CADs to optimise the performance by detecting errors and to identify difficulties, in this case mostly complex classification. Furthermore, it can also be a key element in making the clinical application of deep learning models possible by giving clinical experts a quality measurement for the CADs predictions. Future work should concentrate on improving the robustness of the models and extending the work on different datasets, like other parts of the human body.

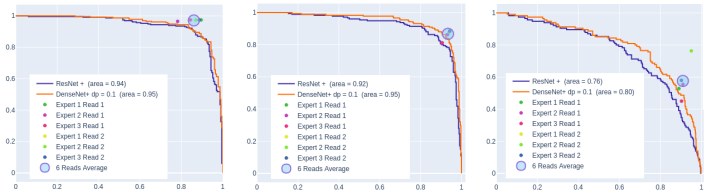

Figure 2: **Clinical experts vs. Our model.** Comparison of different architecture models and the clinical experts for the 2, 3 and 6 classes respectively from left to right.

Acknowledgments

Acknowledgments withheld.

References

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. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. *PLoS medicine*, 15(11):e1002699, 2018.

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.

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.

Amelia Jiménez-Sánchez, Anees Kazi, Shadi Albarqouni, Sonja Kirchhoff, Alexandra Sträter, Peter Biberthaler, Diana Mateus, and Nassir Navab. Weakly-supervised localization and classification of proximal femur fractures. *arXiv preprint arXiv:1809.10692*, 2018.

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.

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.

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.

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.

Jeremy D Schroeder, Sean P Turner, and Emily Buck. Hip fractures: Diagnosis and management. *American Family Physician*, 106(6):675–683, 2022.