A Clinical Evaluation of Deep Learning-Based Detection of Intertrochanteric Femoral Fractures in X-ray Images

Abstract

Intertrochanteric femoral fractures are a serious condition common among the elderly because they are associated with high morbidity and mortality. Precise identification is crucial for prompt treatment, but it can be quite difficult when interpreting X-rays individually due to anatomical complexity and variability. In this study, deep learning models are used to optimize the binary classification of intertrochanteric femoral fractures within X-ray images. An X-ray dataset of about 1,000 images from hospital PACS was used, including both fracture and non-fracture cases. A convolutional neural network (CNN), comprising ResNet50, ResNet101, InceptionV3, and a custom CNN, was fine-tuned using preprocessing and augmentation to manage clinical variability. The models' performance was evaluated based on accuracy, precision, and sensitivity, with ResNet101 achieving the highest results (accuracy: 0.8000, precision: 0.8179). Statistical validation was conducted using confidence intervals (90%, 95%, and 99%) CI for fracture and non-fracture detection . The improved diagnostic accuracy has significant clinical implications, enabling faster diagnoses and better outcomes in time-sensitive environments. This study demonstrates the promising potential of deep learning in addressing diagnostic challenges in the identification of intertrochanteric femoral fractures, supporting its integration into clinical practice.

Keywords: intertrochanteric femoral fractures; deep learning; x-ray imaging; fracture detection; convolutional neural network

Introduction

Intertrochanteric femoral fractures occur along the site between the greater and lesser trochanters of the proximal femur, making them a significant orthopedic concern, especially in older adults [1]. These fractures are typically caused by minimal impact injuries, such as falls in individuals with poor bone density due to osteoporosis, though high-impact injuries can also affect younger populations [2]. The intertrochanteric region is crucial for support and movement, as it bears weight and allows for easy mobility. Fractures in this area can destabilize the hip, leading to serious functional loss. Early and accurate diagnosis is essential to determine the appropriate treatment, which may involve surgical procedures like intramedullary nailing or a dynamic hip screw [3] to restore mobility or conservative management to prevent complications. Conventional X-ray diagnosis can be challenging due to the complex anatomy of the proximal femur, overlapping structures, subtle fracture patterns, and potential for misleading or incomplete information [4].

Intertrochanteric femoral fractures are medically significant because of their elevated mortality and morbidity rates among elderly patients due to the presence of other concomitant diseases like heart conditions and diabetes [5]. These fractures are linked with duration of immobilization, development of pressure ulcers, development of thromboembolism, and a drastic drop in the quality of life, with mortality rate reported up to 2030% [5] in the initial year after the accident. The difficulties in the diagnosis of X-rays come at a clinical aspect, and it is caused by the variation in the displacement of fracture, position of the patient, and quality of the X-rays [6], which may obstruct the detection of the fracture lines. Deep learning and especially pretrained convolutional neural networks (CNNs) bring a potential solution to the problem of detecting fractures (fracture vs. non-fracture) due to the benefit of sophisticated feature extraction to find nuanced patterns in X-rays [7]. With the ability to make faster and more accurate diagnoses, deep learning has the potential to make interventions timelier, minimize diagnostic errors, and assist clinicians at high-stress moments, with the overall consequence of subsequent patient outcomes improving in a critical care environment.

Materials and Methods

Dataset

A dataset of anteroposterior X-ray images of the proximal femur of 1,000 patients recorded in a period between 2015 and 2020 in the images Archiving and Communication System (PACS) of the hospital was used in the study. The data set was homogeneous, including 500 images of confirmed cases of intertrochanteric fractures of the femur and 500 pictures of non-fractures, which were confirmed by professional radiologists. Demographic characteristics of patients were heterogeneous (aged between 60 and 90 years, 60% were female, 40% were male) since they were representative of the patients who come seeking treatment in elderly populations after suffering osteoporotic fractures. The dataset was partitioned into 70% training (700 images), 15% validation (150 images), and 15% test (150 images) sets that would allow for producing a robust model and assessing it.

Deep Learning Models

Four binary classifications (fracture vs. non-fracture) convolutional neural network (CNN) architectures were used: ResNet50, ResNet101, InceptionV3, and a custom CNN. ResNet50 and ResNet101 are known to have a deep residual learning framework and thus provide skip connections to overcome vanishing gradient problems so that they can extract the much-needed features of complex X-ray images. Whereas InceptionV3 makes use of multi-scale convolution to preserve a variety of fracture patterns, the user-defined CNN employs three convolutional layers (32, 64, and 128 filter input), max-pooling, and dense layers, and has been designed as a computationally efficient clinical practice. Each model was also trained on ImageNet to initialize the weights to improve feature detection. Clinical variability, including noise and differences in anatomy, was overcome through resizing to 224x224 pixels, normalization of pixel intensity (0-1), and techniques that augmented the input images by randomly rotating them (angling 15 degrees); randomly flipping horizontally, and changing the image brightness (angling 20 degrees).

Training and Evaluation

The models were fine-tuned using deep learning, with the top layers of DL models unfrozen to adapt to the specific task of intertrochanteric femoral fracture detection. Training was conducted using a binary cross-entropy loss function and the Adam optimizer (learning rate: 1e-4) over 10 epochs, with early stopping based on validation loss to prevent overfitting. Batch size was set to 32 to balance computational efficiency and gradient stability. Performance was evaluated on the test set using metrics including accuracy, precision, sensitivity, and F1-score. Confidence intervals were computed to assess result reliability, and statistical significance was determined using paired t-tests ($p < 0.05$). Training was performed on a GPU-enabled system (NVIDIA RTX 3080) to handle computational demands, with training times recorded for each model to evaluate clinical applicability.

Results and Discussion

ResNet101 turned out to be the most efficient of all the models used, and it has shown the highest testing accuracy (80.00%) and good precision (0.8179), recall (0.8000), and F1-score (0.7690) (Table 1). It also displayed the largest mean confidence in its statistical summary to detect a fracture (mean: 0.8758) which depicts that there were good and consistent predictions. ResNet50 came close behind, achieving a testing accuracy of 77.86 % and maintaining strong statistics regarding back testing key factors (F1-score: 0.7364) backed by a high statistical mean confidence of around 0.8474. InceptionV3 and CNN were at a moderate level with less accuracy in the testing (68.57% and 66.43 % respectively) and lower precision and F1-scores, indicative of reduced reliability. As far as the training times are concerned, it took the ResNet101 the longest (1817.57s), ResNet50 and CNN were the fastest to train, and gave the worst performance. On balance, the best overall model seems to be ResNet101 since the number of layers is large but the model is reliable in terms of its goal.

Table 1. Model Evaluation of Intertrochanteric Femoral Fractures

Model	Training accuracy	Validation accuracy	Testing accuracy	Training time (s)	Precision	Recall	F1
CNN	0.8333	0.6763	0.6643	762.40	0.5530	0.6643	0.5833
InceptionV3	0.7129	0.6763	0.6857	795.97	0.5553	0.6857	0.5906
RestNet50	0.8919	0.7194	0.7786	1237.22	0.7298	0.7785	0.7364
RestNet101	0.8735	0.7338	0.8000	1817.57	0.8179	0.8000	0.7690

The analysis of 1000 intertrochanteric femoral fracture cases (500 Fracture, 500 Non-Fracture) using ResNet50, ResNet101, InceptionV3, and Custom CNN models reveals varying confidence scores, with ResNet101 achieving the highest mean (0.876) for Fracture and Custom CNN the highest (0.656) for Non-Fracture in Tables 2,3,4, and 5. The 90%, 95%, and 99% confidence intervals, detailed in the table, show widening ranges (e.g., ResNet101 Fracture CI from 0.870–0.882 at 90% to 0.866–0.885 at 99%), reflecting increased uncertainty at higher confidence levels. Figure 1 visualizes these sampling distributions with shaded CI regions and mean lines, highlighting ResNet101's strength in fracture detection and Custom CNN's balance across groups, aiding clinical diagnosis from X-ray images.

Table 2. ResNet50 Confidence Interval results

Group	Confidence	Mean	CI Lower	CI Upper
Fracture	90%	0.8474	0.8407	0.8541
	95%	0.8474	0.8394	0.8554
	99%	0.8474	0.8369	0.8579
Non-Fracture	90%	0.6544	0.6437	0.6651
	95%	0.6544	0.6416	0.6671
	99%	0.6544	0.6376	0.6711

Table 3. ResNet101 Confidence Interval results

Group	Confidence	Mean	CI Lower	CI Upper
Fracture	90%	0.8757	0.8695	0.8819
	95%	0.8757	0.8684	0.8831
	99%	0.8757	0.8661	0.8854
Non-Fracture	90%	0.6343	0.6242	0.6444
	95%	0.6343	0.6223	0.6463
	99%	0.6343	0.6185	0.6501

Table 4. InceptionV3 Confidence Interval Results

Group	Confidence	Mean	CI Lower	CI Upper
Fracture	90%	0.8723	0.8653	0.8792
	95%	0.8723	0.8640	0.8806
	99%	0.8723	0.8614	0.8832
Non-Fracture	90%	0.6423	0.6307	0.6539
	95%	0.6423	0.6285	0.6561
	99%	0.6423	0.6242	0.6604

Table 5. CNN Confidence Interval Results

Group	Confidence	Mean	CI Lower	CI Upper
Fracture	90%	0.8544	0.8472	0.8616
	95%	0.8544	0.8459	0.8629
	99%	0.8544	0.8432	0.8656
Non-Fracture	90%	0.6564	0.6453	0.6675
	95%	0.6564	0.6432	0.6696
	99%	0.6564	0.6391	0.6737

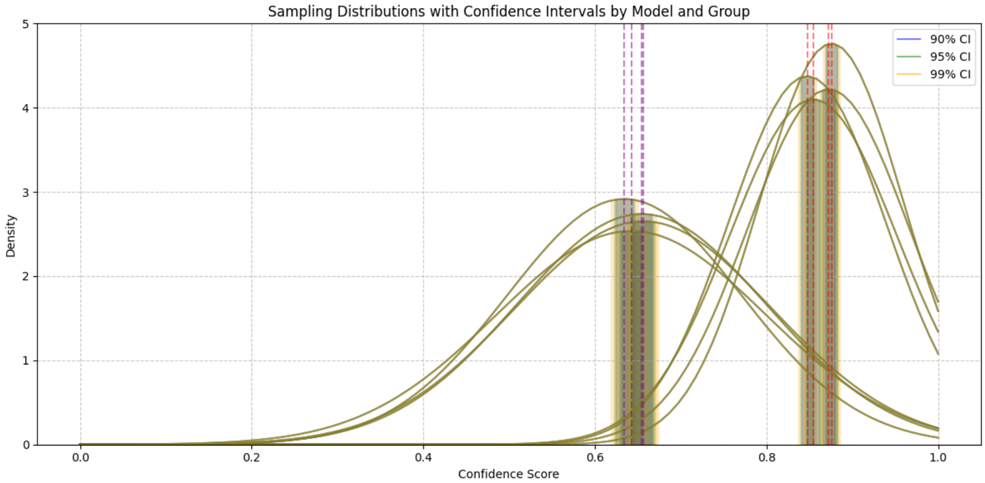

Figure 1. Statistical Analysis (Confidence level) of Deep learning models with the types

Conclusion

Analysis of 1000 cases of intertrochanteric femur fracture with the help of such CNN, InceptionV3, ResNet50 and ResNet101 models, emphasizes the best indicators of ResNet101, the model has a testing accuracy of 0.8000, precision of 0.8179, recall of 0.8000, and F1 score of 0.7690, but it was the longest training time (1817.57s). Confidence scores of fractures and non-fracture are analyzed statistically with ResNet101 having the highest scores on fracture detection (mean 0.876) and Custom CNN having the highest scores on non-fracture (mean 0.656) with confidence intervals of 90% 99% increasing with higher levels, as in the graph of the sampling distributions. ResNet50 also demonstrates a high level of performance (testing accuracy 0.7786, F1 0.7364), whereas CNN and InceptionV3 are not as impressive in terms of the accuracy value (0.6643 and 0.6857) and F1 score (0.5833 and 0.5906). Such findings indicate that ResNet101 and ResNet50 are the most consistent at clinical diagnosis, which remains to be verified on real data.

References

[1] Mundi, S., Pindiprolu, B., Simunovic, N., & Bhandari, M. (2014). Similar mortality rates in hip fracture patients: The influence of age, comorbidity, and time to surgery. *Clinical Orthopaedics and Related Research*, 472(3), 855–861. https://doi.org/10.1007/s11999-013-3217-8

[2] Kannegaard, P. N., van der Mark, S., Eiken, P., & Abrahamsen, B. (2010). Excess mortality in men compared with women following a hip fracture: National cohort study. *Bone*, 47(5), 891–897. https://doi.org/10.1016/j.bone.2010.07.008

[3] Lindsey, R., Daluiski, A., Zhao, S., & Chopra, S. (2018). Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences*, 115(45), 11591–11596. https://doi.org/10.1073/pnas.1806905115

[4] Rajpurkar, P., Irvin, J., Bagul, A., et al. (2017). MURA: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*. https://arxiv.org/abs/1712.06957

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

[6] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[7] Adams, M., Chen, W., Holcdorf, D., McCusker, M. W., & Howe, P. (2019). Computer vs human: Deep learning versus human performance in detecting hip fractures on pelvic radiographs. *Journal of Medical Imaging*, 6(4), 044501. https://doi.org/10.1117/1.JMI.6.4.044501