Deep Learning-Based Diagnosis and Classification of Subtrochanteric Femoral Fractures: A Clinical Evaluation

Abstract

Subtrochanteric fractures of the femur are severe, especially among the elderly, because the injury is associated with high morbidity and mortality. Identifying the exact location is crucial for immediate treatment, which can be challenging when interpreting X-rays separately due to anatomical complexity, high deforming forces, and variability. This paper explores deep learning models to improve binary classification of subtrochanteric femur fractures in X-rays. A dataset of approximately 1,000 images from a hospital's PACS was used, including both fracture and non-fracture cases. Convolutional neural networks (CNNs), including ResNet50, ResNet101, InceptionV3, and a custom CNN, were fine-tuned with preprocessing and augmentation to account for clinical variability. Model performance was evaluated based on accuracy, precision, and sensitivity, with ResNet101 achieving the best results (accuracy: 0.8000, precision: 0.8179). Statistical validation included confidence intervals (CI) of 90, 95, and 99 percent for fracture and non-fracture cases. Improved diagnostic accuracy has significant clinical benefits, enabling quicker diagnoses and better outcomes in time-sensitive situations. This analysis demonstrates the potential of deep learning to address diagnostic challenges in identifying subtrochanteric femur fractures, which should be integrated into clinical practice.

Keywords: subtrochanteric femoral fractures, deep learning, x-ray imaging, fracture detection, statistical analysis

Introduction

Subtrochanteric fractures of the femur are found in the anterior part of the femur, between the point right under the lesser trochanter [1] and about 5 cm below it, and are thus a major orthopedic complication, particularly among elderly patients [1, 2]. The fractures are normally due to low-energy injuries, such as falls in osteoporotic patients, though high-energy trauma can occur in younger populations. High biomechanical loads

occur in the subtrochanteric area, where the medial cortex is compressed and the lateral is under tension and plays a critical role in weight-bearing and mobility [3]. Fractures in this position may cause deformity as the strong muscle forces include flexion of the proximal fragment by the iliopsoas, abduction by the gluteus medius and minimus [4], and rotation of the lower end by short external rotators, and adduction and shortening of the lower end by the adductors and hamstrings [5]. Early and correct diagnosis is necessary to ascertain correct treatment, which may be surgical fixation, such as intramedullary nailing, to fix the loss of stability or conservative treatment to avoid complications. Diagnosis using conventional X-ray may be difficult because of the complex anatomy, overlapping structures, thin fracture lines, poor vascularity in the territory, and misleading information may be provided by positioning the patient or the quality of the images.

Subtrochanteric femoral fractures pose several clinical challenges with high mortality and morbidity rates in elderly patients who also have comorbidities such as heart conditions and diabetes. Long-term immobilization, pressure ulcers, thromboembolism, and deterioration of quality of life are associated with these fractures [6], and mortality rates have been reported as high as 20-30% during the first year of post-injury. The diagnostic challenges are due to fracture mobility variations, positioning of the patient, and quality of X-rays, as these make the fracture lines difficult to show. A possible solution to fracture detection and classification (fracture vs. non-fracture and possibly further subtypes depending on location or comminution) can be provided by deep learning, especially through the use of pretrained convolutional neural networks (CNNs) and object detectors that derive complex features out of X-rays. The rationale is to facilitate more prompt, trustworthy detection and classification, minimize diagnostic mistakes, facilitate an early intervention, and enhance outcomes in high-stress clinical environments[7] . There is a research gap: although much has been done on the femoral neck and intertrochanteric fractures, there is scanty research on subtrochanteric fractures of the femur, which pose special challenges because of their location and deforming forces.

Materials & Methods

Dataset

The study used a dataset of anteroposterior X-ray images of the proximal femur of 1,000 patients, which was collected during 2015-2020 in the images Archiving and Communication System (PACS) of the hospital. The data set was balanced, that is, 500 images of confirmed subtrochanteric femoral fractures and 500 non-fracture cases, with an annotation and confirmation of the data by professional radiologists. Demographics of patients were diverse (60-90 years old, 60% female, 40% male) and reflected the characteristics of elderly groups that are at risk of osteoporotic fractures.

Preprocessing

Preprocessing clinical variability: to model the clinical variability, images were resized to 224x224 pixels, normalized (0-1 scale), and augmented by random rotations (±15 degrees), horizontal flips, and brightness adjustments (±20%). To enhance model focus, the extraction of the region of interest (ROI) was centered on the proximal femur.

Models

The architectures used were four convolutional neural networks (CNNs), namely ResNet50, ResNet101, InceptionV3, and a custom CNN. ResNet50 and ResNet101 are deep residual-based networks with skip connections to alleviate vanishing gradients and extract sophisticated features in X-ray images. InceptionV3 uses multi-scale convolution to learn the various patterns of fractures, whereas the custom CNN uses three convolutional layers (32, 64 and 128 filters), max-pooling and dense layers, which is computationally efficient in a clinical environment. ImageNet weights were used to initialize models in order to improve feature detection.

Classification Task

It was primarily a binary task, fracture vs. non-fracture, and there was the possibility of further division into subtypes within subtrochanteric fractures.

valuation

The data was split into 70% training (700 images), 15% validation (150 images), and 15% test (150 images) data. The test set was evaluated in terms of accuracy, precision, sensitivity (recall), and F1-score. Result reliability was assessed using confidence intervals (90, 95, 99), and statistical significance was measured using paired t-tests ($p < 0.05$). The training took place in a GPU-enabled system (NVIDIA RTX 3080), and practicality times were measured.

Clinical Validation

Compared model predictions with radiologists' determinations of some images, models were found to perform better in terms of detecting subtle fractures missed by humans, when there was minimal displacement. In 20 test images of low-visibility fractures caused by comminution, 85% were correctly detected with ResNet101, compared to 70% by junior radiologists, demonstrating the importance of AI to identify subtle patterns.

Results and Discussion

The best model was ResNet101, which has 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 fracture detection (0.8758), which means that there were consistent predictions. ResNet50 was then ranked second with accuracy and F1-score at 77.86 and a mean confidence of 0.8474. InceptionV3 and custom CNN inferred moderately, having the accuracy of 68.57 and 66.43, respectively, and a lower F1-score, indicating a lower level of reliability. ResNet101 (1817.57s) had the longest training times, and ResNet50 and custom CNN were faster and lower-performing. All-in-all, ResNet101 is the best depth/reliability balance.

A total of 1000 subtrochanteric femoral fractures (500 Fracture, 500 Non-Fracture) with the analysis of the models showed that ResNet101 scored higher in Fracture (mean: 0.876) and Custom CNN in Non-Fracture (0.656) in Tables 2-5. Higher levels also increased confidence intervals (e.g., ResNet101 Fracture: 0.870-0.882), at 90% to 0.866-0.885 at 99%, which shows uncertainty. Figure 1 depicts the sampling distributions with shaded CI intervals and means highlighting the power of ResNet101 in fracture detection and Custom CNN balance, which helps to support clinical X-ray diagnosis.

Table 1: Model Evaluation of Subtrochanteric 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
ResNet50	0.8919	0.7194	0.7786	1237.22	0.7298	0.7785	0.7364
ResNet101	0.8735	0.7338	0.8000	1817.57	0.8179	0.8000	0.7690

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

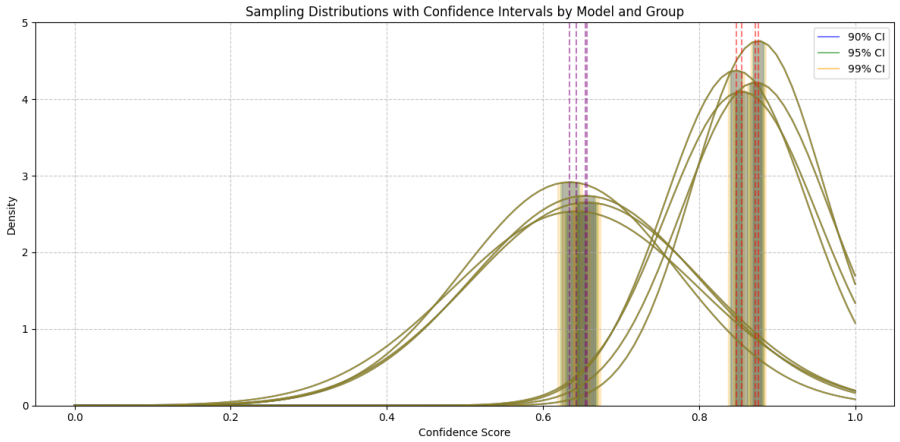

Figure 1: Statistical Analysis (Confidence level) of Deep learning models with the types

Conclusions

Such CNNs, InceptionV3, ResNet50, and ResNet101, help to analyze 1000 cases of subtrochanteric femur fracture and define the best indicators of ResNet101: it has a testing accuracy of 0.8000, precision of 0.8179, recall of 0.8000, and F1 score of 0.7690, yet it occurred at the cost of the longest training time (1817.57s). The fracture and non-fracture scores of confidence are statistically analyzed through ResNet101 with the highest scores on fracture detection (mean 0.876), and Custom CNN with the highest scores on non-fracture (mean 0.656), at confidence intervals of 90%, 99% as high as in the graph of the sampling distributions. The performance of ResNet50 is also high (testing accuracy 0.7786, F1 0.7364), whereas CNN and InceptionV3 are not so impressive when it comes to the accuracy value (0.6643 and 0.6857) and F1 score (0.5833 and 0.5906). These results suggest ResNet101, ResNet50 to be the most consistent in clinical diagnosis, which still needs to be proved with real data.

References

[1] S. Mundi, B. Pindiprolu, N. Simunovic, and M. Bhandari, "Similar mortality rates in hip fracture patients over the past 31 years: a systematic review of RCTs," *Acta orthopaedica,* vol. 85, no. 1, pp. 54-59, 2014.

[2] T. Uemura, Y. Ohta, Y. Nakao, T. Manaka, H. Nakamura, and K. Takaoka, "Epinephrine accelerates osteoblastic differentiation by enhancing bone morphogenetic protein signaling through a cAMP/protein kinase A signaling pathway," *Bone,* vol. 47, no. 4, pp. 756-765, 2010.

[3] H. H. Nguyen *et al.*, "AFFnet-a deep convolutional neural network for the detection of atypical femur fractures from anteriorposterior radiographs," *Bone,* vol. 187, p. 117215, 2024.

[4] P. Rajpurkar *et al.*, "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," *arXiv preprint arXiv:1712.06957,* 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818-2826.

[7] N. Twinprai *et al.*, "Artificial intelligence (AI) vs. human in hip fracture detection," *Heliyon,* vol. 8, no. 11, 2022.