|ICAS2025|

Ensemble Deep Learning Framework for Multiclass Classification of Femur and Pelvic Fractures

Abstract: The fractures of the femur and pelvis are life-threatening orthopedic conditions that are common among older individuals, causing severe complications and reducing mobility. This paper introduces a deep learning method for automatically recognizing pelvic and femur fractures in X-ray images. Our approach relies on ensemble deep learning models such as convolutional neural network (CNN) architectures, including ResNet50, InceptionV3, ResNet101, EfficientNetB0, EfficientNetV2, MobileNet and Xception, which classify fractures into five possible categories: non-displaced, incomplete non-displaced, complete non-displaced, partially displaced, and fully displaced fractures. Data, comprising around 1000 X-ray images from various hospitals, were pre-processed and augmented to strengthen the model. The ResNet50 model achieved the highest classification accuracy at 80 percent on the test set and was identified as the best-performing model for distinguishing fracture types. The framework combines modern feature engineering and ensemble learning models to enable early and accurate diagnosis, leading to improved clinical outcomes in treating femur and pelvic fractures, reducing diagnostic errors, and significantly enhancing the diagnostic process.

Keywords: Femur fracture, Pelvis fracture, Deep learning, X-ray imaging, RestNet50, Medical image classification

Author Note

The dataset used in this study was collected from multiple hospitals, including Base Hospital Tellipalai Jaffna, Northern Central Hospital Jaffna, Teaching Hospital Kandy, Anuradhapura Hospital, and Venus Speciality Hospital Pvt Ltd. The authors declare no conflicts of interest related to this work.

Introduction

Fractures are considered discontinuities of the bone (Mutasa et al., 2020), namely that are characterized by breaks or cracks in the continuity of the bone, which is a major medical issue because it might lead to pain(Beyaz et al., 2020), impairment of function, and a long-term problem. They are usually caused by trauma, e.g., by a fall, or by some underlying cause such as osteoporosis that weakens the bone structure (Alzaid et al., 2022). In elderly people, fractures are especially common because of the loss of bone density, which is caused by age, and there are both short-term and life-threatening (Hardalaç et al., 2022) implications of this occurrence. Fractures are best addressed under proper diagnosis, and with the help of newer diagnostic tools, this becomes extremely crucial to facilitate a proper diagnostic process to counter adverse effects that can be related to any fracture.

Femur and pelvic fractures constitute one of the most serious (Hsieh et al., 2023) forms of fractures because of the amount of anatomical importance and the complications that might arise. The femur, the largest and strongest bone of the human body (Yıldız Potter et al., 2024), may get fractured in its neck, shaft, or distal portions (Qi et al., 2020), which generally need surgical repair. These complex rings of bones, involved in providing stability to the spine, as well as connecting with the lower limbs, may result in pelvic fractures (Sharrab et al., 2021), especially during high-energy trauma (Inoue et al., 2022) that interferes with stability and the normal functioning of the internal organs. As compared to other types of fractures, these can be compared to tibial fractures, radial fractures, or vertebral fractures because of their high morbidity (Ukai et al., 2021), the length of recovery, and the need for specific diagnostic imaging to plan the treatment based on their unique characteristics.

This paper uses multiclass classification based on deep learning algorithms to analyze X-rays in analyze, incomplete, complete, partially, and fully displaced femur and pelvic fractures (Wang et al., 2023) to find out how to identify these fractures in diagnosis procedures. Multiclass classification allows separating these kinds of fractures in terms of their radiographic appearance, which plays an important role in the choice of treatment strategies. This will be accomplished by possibly increasing the accuracy of the diagnosis, minimizing the existence of human error, and optimizing clinical processes to provide a better outcome to patients with orthopaedic conditions.

Related Work

The clinical significance of femur fractures and pelvis fractures, especially femoral neck fractures, is in the severe effect the fracture produces on mobility and the eventual occurrence of the complication of avascular necrosis, especially in old age. Such fractures are commonly caused by trauma or bone weakness, which requires proper diagnosis, as well as time, to direct treatment to prevent negative results. With the introduction of deep learning, it has become possible to automate the X-ray analysis to make the diagnosis of fractures more accurate, as well as error-free. In modern history, there have been numerous studies on different deep learning models that can be used to classify femur and pelvic fracture sites, proving much improvement in comparison with the old radiographic techniques.

The study (Inoue et al., 2022) used a database of

Table **(1)**| Summary of Techniques and Limitations in Related Studies

Study	Techniques Used	Limitations
(Inoue et al., 2022)	Customized residual network, DRRs, GANs, data augmentation	Limited dataset diversity, challenges with complex fractures
(Killeen & DeMeo, 1999)	DAFDNet (DCNN with Gabor filter, Squeeze-and-Excitation ghost convolution)	Reduced performance with subtle fractures due to image noise
(Kukar et al., 1996)	DR-FDS (Multi-domain Fracture Classification Network, Faster R-CNN)	Low sensitivity for subtle fractures
Moon et al., 2022	Faster R-CNN with ResNet-50 and FPN, data augmentation	Lower accuracy for complex fractures due to imbalanced data
Mu et al., 2021	Vision Transformer (ViT) with deformable attention framework	Limited by availability of paired AP and lateral views

1063 anteroposterior radiographs of the hip and split them into Garden I/II, Garden III/IV, and normal groups and reached a three-category AUC of 0.96 and 86 percent accuracy with a personalized residual network. The augmentation of the data with digitally reconstructed radiographs (DRR) and generative adversarial network (GAN) increased the AUC to 0.91 and demonstrated the significance of training data. Nevertheless, the model had difficulties with classifying complex fractures accurately because of the diversity of the data sets. In another research (Killeen & DeMeo, 1999), a deep convolutional neural network (DCNN) refers to DAFDNet with Gabor filters and Squeeze-and-Excitation ghost convolution that affords an accuracy of 94.8 percent in characterizing non-displaced feminine neck breaks. Its drawbacks involved poor results in images with small differences in fractures caused by image noise, and the complicated anatomy of bones.

The DR-FDS model (Kukar et al., 1996), built upon the Multi-domain Fracture Classification Network and Faster R-CNN, was used in analyzing the multicentric pelvic radiographs and reflected the increased performance of clinicians in minimal fracture detection, but is poor in low-sensitivity detection of subtle fractures. Similarly, a work (Moon et al., 2022) on 2,333 femoral X-ray radiographs using a Faster R-CNN model with a ResNet-50 and the Feature Pyramid Network (FPN) returned a mAP of 68.8, compared with junior surgeons who can diagnose complex fractures (C1, C3), but not do well due to a non-balanced data set. These experiments demonstrate the possibility of deep learning in facilitating the process of fracture detection and also note the difficulties of data imbalance and the intricacies of the subtle fracture patterns.

Non-displaced femoral neck fractures on 1,250 paired hip radiographs reached a binary accuracy of 95.8 and an AUC of 0.988 with a vision transformer (ViT) model (Mu et al., 2021). It had its validated generalizability that was externally derived, but performance was restricted to paired availability views. All of them together prove that deep learning models offer substantial potential to enhance the localization and classification of femur and pelvic fractures compared to the traditional techniques and bypassing inexperienced clinicians, but these methods are not free of limitations where data imbalance, image noise, and difficult patterns remain. Table 1 shows that Existing studies about the Femur and pelvic fracure of multiclass claddisifcation.

Materials and Methods

The paper uses a multiclass classification approach to categorize femur and pelvis fractures from X-ray images into five groups: non-displaced, incomplete non-displaced, complete non-displaced, partially displaced, and fully displaced. The training dataset includes 1,000 X-ray images collected from various hospitals, such as Base Hospital Tellapalas Jaffna (450 images), Northern Central Hospital Jaffna (800 images), Teaching Hospital Kandy (450 images), Anuradhapura Hospital (450 images), and Venus Speciality Hospital Pvt Ltd (300 images). Image preprocessing involved resizing them to a standard size (224x224x3) and applying random flips and rotations to improve model robustness. A custom classification layer was added to pre-trained deep learning models by replacing the final layer with one featuring a dense layer with 512 nodes, ReLU activation, and a dropout rate of 0.5 to prevent overfitting, followed by a softmax layer to output probabilities for the five classes. The dataset was split into 80 percent for training, 10 percent for validation, and 10 percent

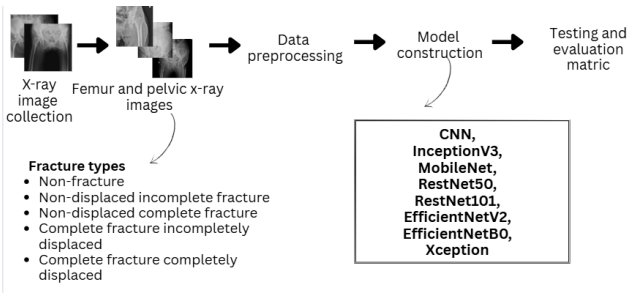

Figure **(1)**| High-level architecture

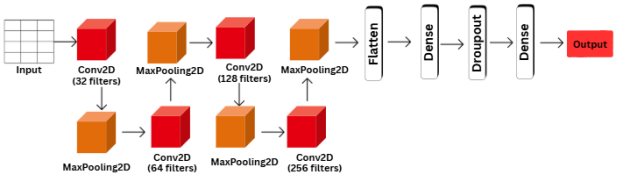

Figure **(2)**| The deep learning models architecture used in femur and pelvic fracture

for testing, enabling assessment of the model's ability to detect and classify femur and pelvic fractures accurately for proper diagnosis and treatment planning. Figure 1 shows that high-level architecture of femur and pelvic fracture classification. Table 2, 3 show that the collection of data from Sri Lankan hospitals and the number of fracture types and number of images for each types.

Table **(2)**| Number of X-ray Images Collected from Various Hospitals

Hospitals	**No. of Images**
Base Hospital Tellipalai, Jaffna	450
Northern Central Hospital, Jaffna	800
Teaching Hospital, Batticaloa	500
Aathura Hospital, Baily Rd, Batticaloa	450
Venus Specialty Hospital Pvt Ltd	300

Table **(3)**| Distribution of X-ray Images by Fracture Type

Fracture Types	**No. of Images**
Non-Fractured	534
Non-displaced incomplete fracture	136
Non-displaced complete fracture	91
Complete fracture incompletely displaced	93
Complete fracture completely displaced	116

Model Construction

The classification models of femur and pelvis fractures are created based on pre-trained deep architectures by exploiting ImageNet-learnt features to classify X-ray images into 5 groups (non-fractured, non-displaced incomplete fracture, non-displaced complete fracture, complete fracture partially displaced, and complete fracture fully displaced). Two variants of pre-trained models, which are in use with weights pre-trained on ImageNet and the parameters include top set to False and exclude the original classification head, are as follows: InceptionV3, ResNet-50, ResNet-101, MobileNet, EfficientNetB0, EfficientNet-V2, and Xception. This allows the models to become efficient in detecting complex features, such as bone edges and fracture patterns, in 224x224x3 X-ray images whilst remaining adaptable enough to perform the task of femur and pelvic fracture classification by means of a custom classification head.

To guarantee a strong extraction of features using small volumes of medical images, the layers in each of the pre-trained models are frozen, creating a solid backbone of hierarchy of features without a lot of retraining. This method is important because of the factors that limit access to medical data because of privacy and also due to class imbalances. Feature extraction is followed by a GlobalAveragePooling2D layer, which shrinks the spatial features by averaging out each feature map into a single value, resulting in a compact feature vector that summarizes the important features of the fracture. This vector is then passed onto the Dense layer of 512 nodes ReLU activation to allow the model to understand complex fracture patterns, and a Dropout layer with a rate of 0.5 that drops out half of neurons in training to eliminate overfitting and improve generalization of unseen data.

The last dense layer has the softmax activation technique to yield a probability across the five classes of fracture to facilitate confident clinical decisions. Its architecture employs commonly used attributes in the base models used as pre-trained models, including residual connections, depthwise separable convolutions, and compound scaling, to make efficient use of both the complicated shapes in the femur and pelvis X-rays and the fracture patterns. Optimization of performance on the femur and pelvic fracture dataset using fine-tuning of these models counteracts limitations in the amount of data and computational resources, with an increased accuracy in the diagnosis. Figure 2 show that femur and pelvic fracture models architecture used.

Testing and evaluation matrix

The analysis of respective models of femur and pelvic fracture classification gives a wholesome approach of how well they can identify and classify X-ray images into the five classes: non-fractured, non-displaced in-

complete fracture, non-displaced complete fracture, complete fracture partially displaced, and complete fracture fully displaced. Accuracy is an assessment of the number of accurate predictions made on all the classes, as it measures the overall performance of the model on the test set. The ratio of the correct predictions of a true positive to the number of all positive predictions is measured as precision, which is critical in the effort to reduce false positives with the fracture classification. Recall (or sensitivity) tests the terms of the model in its capacity to identify all of the veritable cases of fractures with low missed perfusion. As the harmonic mean of both precision and recall, the F1 Score attempts to balance the two metrics to give one estimate of a model's measure and can be especially used when dealing with biases in classes. The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) quantifies a tradeoff curve between different thresholds of true positive and false positive rates, values which are better at implying strong discrimination between fracture and non-fracture cases, and as such, are particularly useful to assess model performance amid the multiclass classification context.

Results

The proposed models, ResNet-101 and ResNet-50, were the highest scoring models in the classification of fractures of the femur and pelvis, superior to custom CNN and other pre-trained models.

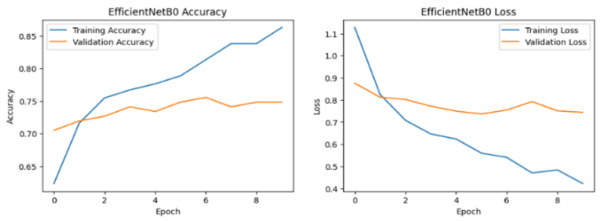

Figure **(3)**| Accuracy and loss of EfficientNetB0

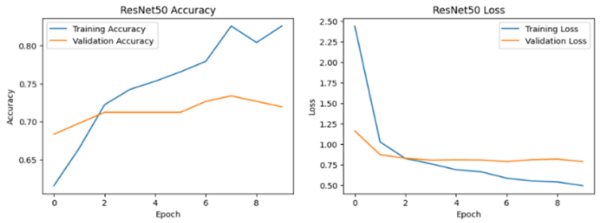

Figure **(4)**| Accuracy and loss graph of RestNet50

ResNet-101 had the best test accuracy with 0.8200 and was more precise (0.8379) and recalled (0.8200), and had an F1 score (0.7890), better generalized and balanced the trade-off between precision and recall, and trained in 1817.57 seconds. Coming behind ResNet-50 in terms of test accuracy by 0.0011, precision by 0.0025, recall by 0.0011, and F1 score by 0.0018, though

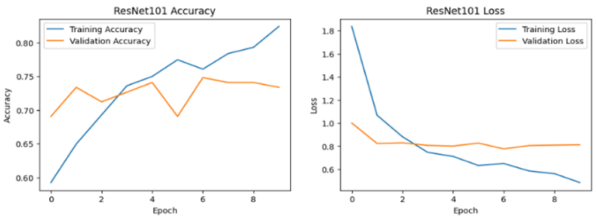

Figure **(5)**| Accuracy and loss graph of RestNet101

it was trained in 1237.22 seconds, is ResNet-50, a popular alternative with a slight drop in accuracy, precision, recall, and F1 score to go with a faster time. The two presented models are based on the residual learning concept to learn detailed fracture features when pre-trained on ImageNet and fine-tuned on the femur and pelvic fracture dataset, so that a robust feature extractor is realized in producing results based on the five classes, including non-fractured, non-displaced incomplete, non-displaced complete, partially displaced, and fully displaced fractures.

The Figures 3,4,5 show of accuracy and interesting loss curves obtained in training deep learning models (ResNet50, ResNet101, and EfficientNetB0), indicating their effectiveness in determining the multiclass classification of femur and pelvic fractures. The most optimal results are demonstrated by ResNet101 and ResNet50, whose validation and testing accuracy curves increase to about 0.82 and 0.80, respectively, and loss divergence is minimal (reaching a constant of 0.5), which is a clear sign of good generalization. Conversely, the Custom CNN has a wide disparity between training and validation accuracy (0.85 vs. 0.70) along with the rise in the loss, implying overfitting. Although being efficient, MobileNet displays a flatter accuracy curve (0.73) along with a greater loss variability, which indicates an accuracy-efficiency trade-off. The burden and the rise of the accuracy (up to 0.77 and 0.75) of EfficientNetV2 and EfficientNetB0 are stable due to controlled losses, which protodeternalizes their effectiveness in fracture classification.

Conclusion

This paper shows that pre-trained deep learning architectures such as InceptionV3, ResNet50, ResNet101, MobileNet, EfficientNetB0, EfficientNetV2, and Xception are very useful in the multiclass classification of femur and pelvic fractures in X-ray images than custom CNN methods because they actively use pre-trained features of large-scale data such as ImageNet. Although ResNet101 required more time to train, 1817.57 seconds, the model achieved the best performance with a test accuracy of 0.8200, a precision of 0.8379, a recall of 0.8200, and an F1-score of 0.7890, which highlights the good generalization ability of the model. ResNet50 offered a trade-off between performance (accuracy = 0.7986) and efficiency (1237.22 seconds) that might be good enough to be deployed in the clinical environ-

Table **(4)**| Performance Comparison of Deep Learning Models

Model	Train Acc.	Val Acc.	Test Acc.	Time (s)	Precision	Recall	F1 Score
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
MobileNet	0.7577	0.6904	0.7071	251.44	0.3134	0.7071	0.6129
EfficientNetB0	0.8750	0.7482	0.7286	404.34	0.6613	0.7286	0.6769
EfficientNetV2	0.8858	0.7338	0.7500	1330.01	0.6678	0.7500	0.6912
Xception	0.7346	0.6906	0.6714	1339.33	0.4838	0.6714	0.5623

ment. Although MobileNet is the fastest with the time to converge as 251.44 s, it produced the minimum accuracy of 0.7271. These findings also emphasize fracture classification advantages with pre-trained models over custom models to make use of larger and diverse datasets to further improve the accuracy of these models and enable clinical decision-making in relation to femur and pelvic fracture treatment.

References

Alzaid, A., Wignall, A., Dogramadzi, S., Pandit, H., & Xie, S. (2022). Automatic classification and classification of peri-prosthetic femur fracture. *International Journal of Computer Assisted Radiology and Surgery*, *17*(4), 649–660.

Beyaz, S., Açıcı, K., & Sümer, E. (2020). Femoral neck fracture classification in x-ray images using deep learning and genetic algorithm approaches. *Joint Diseases and Related Surgery*, *31*(2), 175.

Hardalaç, F., Uysal, F., Peker, O., Çiçeklidağ, M., Tolunay, T., Tokgöz, N., Kutbay, U., Demirciler, B., & Mert, F. (2022). Fracture classification in wrist x-ray images using deep learning-based object classification models. *Sensors*, *22*(3), 1285.

Hsieh, S., Chiang, J., Chuang, C., Chen, Y., & Hsu, C. (2023). A computer-assisted diagnostic method for accurate classification of early nondisplaced fractures of the femoral neck. *Biomedicines*, *11*(11), 3100.

Inoue, T., Maki, S., Furuya, T., Mikami, Y., Mizutani, M., Takada, I., Okimatsu, S., Yunde, A., Miura, M., & Shiratani, Y. (2022). Automated fracture screening using an object classification algorithm on whole-body trauma computed tomography. *Scientific Reports*, *12*(1), 16549.

Killeen, K., & DeMeo, J. (1999). Ct classification of serious internal and skeletal injuries in patients with pelvic fractures. *Academic Radiology*, *6*(4), 224–228.

Kukar, M., Kononenko, I., & Silvester, T. (1996). Machine learning in prognosis of the femoral neck fracture recovery. *Artificial Intelligence in Medicine*, *8*(5), 431–451.

Moon, G., Kim, S., Kim, W., Kim, Y., Jeong, Y., & Choi, H.-S. (2022). Computer aided facial bone fracture diagnosis (ca-fbfd) system based on object classification model. *IEEE Access*, *10*, 79061–79070.

Mu, L., Qu, T., Dong, D., Li, X., Pei, Y., Wang, Y., Shi, G., Li, Y., He, F., & Zhang, H. (2021). Fine-tuned deep convolutional networks for the classification of femoral neck fractures on pelvic radiographs: A multicenter dataset validation. *IEEE Access*, *9*, 78495–78503.

Mutasa, S., Varada, S., Goel, A., Wong, T., & Rasiej, M. (2020). Advanced deep learning techniques applied to automated femoral neck fracture classification and classification. *Journal of Digital Imaging*, *33*(5), 1209–1217.

Qi, Y., Zhao, J., Shi, Y., Zuo, G., Zhang, H., Long, Y., Wang, F., & Wang, W. (2020). Ground truth annotated femoral x-ray image dataset and object classification based method for fracture types classification. *IEEE Access*, *8*, 189436–189444.

Sharrab, Y., Al-shboul, S., Alsmira, M., Dwekat, Z., Alsmadi, I., & Al-Khasawneh, A. (2021). Performance comparison of several deep learning-based object classification algorithms utilizing thermal images. *Proceedings of IEEE Conference*, 16–22.

Ukai, K., Rahman, R., Yagi, N., Hayashi, K., Maruo, A., Muratsu, H., & Kobashi, S. (2021). Detecting pelvic fracture on 3d-ct using deep convolutional neural networks with multi-orientated slab images. *Scientific Reports*, *11*(1), 11716.

Wang, L.-X., Zhu, Z.-H., Chen, Q.-C., Jiang, W.-B., Wang, Y.-Z., Sun, N.-K., Hu, B.-S., Rui, G., & Wang, L.-S. (2023). Development and validation of a deep-learning model for the classification of non-displaced femoral neck fractures with anteroposterior and lateral hip radiographs. *Quantitative Imaging in Medicine and Surgery*, *14*(1), 527.

Yıldız Potter, İ., Yeritsyan, D., Mahar, S., Kheir, N., Vaziri, A., Putman, M., Rodriguez, E., Wu, J., Nazarian, A., & Vaziri, A. (2024). Proximal femur fracture classification on plain radiography via feature pyramid networks. *Scientific Reports*, *14*(1), 12046.