Pattern Recognition Letters 190 (2025) 1–7

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

FracNet: An end-to-end deep learning framework for bone fracture detection

Haider A. Alwzwazy, Laith Alzubaidi *, Zehui Zhao, Yuantong Gu

School of Mechanical, Medical, and Process Engineering, Brisbane, 4000, QLD, Australia

ARTICLE INFO

Editor: Jiwen Lu

Keywords:
Deep learning
Fracture detection
Feature fusion
Attention mechanisms
Medical imaging

ABSTRACT

Fracture detection in medical imaging is crucial for accurate diagnosis and treatment planning in orthopaedic care. Traditional deep learning (DL) models often struggle with small, complex, and varying fracture datasets, leading to unreliable results. We propose FracNet, an end-to-end DL framework specifically designed for bone fracture detection using self-supervised pretraining, feature fusion, attention mechanisms, feature selection, and advanced visualisation tools. FracNet achieves a detection accuracy of 100% on three datasets, consistently outperforming existing methods in terms of accuracy and reliability. Furthermore, FracNet improves decision transparency by providing clear explanations of its predictions, making it a valuable tool for clinicians. FracNet provides high adaptability to new datasets with minimal training requirements. Although its primary focus is fracture detection, FracNet is scalable to various other medical imaging applications.

1. Introduction

Fracture detection is crucial in medical imaging for accurate diagnosis and treatment planning in orthopaedic care. Medical imaging techniques such as X-ray, computed tomography (CT), and magnetic resonance imaging (MRI) have significantly improved the ability to detect fractures. However, challenges such as noise, anatomical variations, and the complexity of fracture patterns still exist [1]. Deep learning (DL) has become a key technology in medical imaging, leading to significant advances in the diagnosis and treatment of various conditions [2]. DL algorithms use large datasets to accurately identify complex patterns and anomalies within medical images. This capability has been especially revolutionary in fracture detection, as DL models have shown performance levels comparable to or better than expert radiologists, significantly improving diagnostic precision and efficiency [3]. Despite significant achievements, many challenges in using DL for fracture detection affect its reliability and widespread implementation. A major problem is the lack of annotated training data and the difficulty of generalising features [4]. Moreover, the challenge of explainability is critical, especially when focussing on the region of interest (ROI). DL models are often criticised for being "black boxes," making it difficult to understand their decision-making process [5]. In the case of fracture detection, it is crucial for these models not only to make accurate predictions but also to provide clear visual explanations that highlight the specific areas of an image where fractures are detected. Furthermore, incorporating new data into pre-trained models presents a significant challenge. Medical imaging data are constantly evolving due to the development of new imaging techniques and modalities. Pre-trained models may not perform optimally on these new types of data unless they are regularly retrained or fine-tuned, which can be resource-intensive and time-consuming. To address these complex challenges, we have developed a new framework that uses self-supervised pre-trained (SSP) [6], feature fusion, attention mechanisms, feature selection, and advanced visualisation tools. SSP can address the problem of limited annotated data by allowing models to learn valuable representations from large volumes of unlabelled data [6]. Attention mechanisms help models focus on the most relevant parts of the input data, which is especially helpful in medical imaging where the region of interest can be small or subtle [7]. By dynamically highlighting important features and suppressing irrelevant ones, attention mechanisms enhance the interpretability and accuracy of predictions. Feature fusion involves integrating information from multiple models or modalities to create a more comprehensive representation of the data [8]. This approach can improve the robustness and accuracy of the model by combining complementary features that may not be as effective individually. For example, in fracture detection, integrating features from different DL models can provide a more complete understanding of the fracture region, leading to better diagnostic performance [9]. Feature selection identifies and keeps the most important features extracted from the DL models, which are crucial for accurate fracture detection. This process improves model performance by focusing on key data, reducing overfitting, and speeding up training time. By incorporating these advanced methods, our proposed framework, FracNet, aims to create a more robust, reliable, and clinically applicable solution to identify fractures in medical imaging.

* Corresponding author.
E-mail address: l.alzubaidi@qut.edu.au (L. Alzubaidi).

https://doi.org/10.1016/j.patrec.2025.01.034
Received 18 July 2024; Received in revised form 2 December 2024; Accepted 31 January 2025
Available online 8 February 2025

2. Related work: Musculoskeletal fracture image classification

Various factors influence the incidence of bone fractures, including sex, age, physiology, biology, and access to treatment and prevention programmes [10]. Bone fracture impairments increase morbidity and mortality with age [11]. Osteoporosis [12] and bone trauma [13] are major causes of fractures. Several studies have assessed fracture image classification using CNNs. Chung et al. [14] used shoulder images and a ResNet model, achieving 96% accuracy, higher than the human experts. Rajpurkar et al. [15] evaluated a CNN model on the MURA dataset, outperforming radiologists. Uysal et al. [16] tested various CNN architectures on MURA shoulder images, with an ensemble model achieving the best kappa score of 69.4%. Huynh et al. [17] proposed a novel CNN design for humerus images from the MURA dataset, achieving a validation accuracy of 68.4%. Kitamura et al. [18] evaluated CNN models on ankle images, with an ensemble that achieved 81% accuracy. Kandel et al. [19] used TL with ImageNet weights for fracture detection, showing that pre-trained weights significantly improve performance. Recent studies have focused on feature fusion and ensemble techniques. Alammar et al. [8] introduced a TL approach for medical images, achieving superior performance in humerus and wrist classification tasks. Alzubaidi et al. [9] applied similar techniques to the forearm, achieving 90.7% accuracy. Alzubaidi et al. [20] developed a fusion framework for shoulder abnormalities, achieving 99.2% accuracy and outperforming expert surgeons. Abdelazim and Fouad [21] proposed an AI-based system to automate root fracture detection in periapical radiographs using a voting mechanism that integrates five pretrained models: VGG16, VGG19, ResNet50, DenseNet121, and DenseNet169. DenseNet121 and DenseNet169 outperformed the others in sensitivity and specificity, effectively addressing issues such as overfitting and noise. Although VGG16 showed good data alignment, it was complex and ResNet50 had biases. The ensemble voting system improved reliability, highlighting the potential to assist less experienced dentists and enhance diagnostic workflows. Previous methods for detecting fractures encounter three main challenges. First, the lack of labelled datasets hinders the training of DL models, as obtaining annotated medical imaging data is difficult due to privacy concerns and the expertise required for accurate labelling. Second, these methods often struggle to generalise across diverse datasets, given that fractures can occur in various anatomical locations with unique imaging characteristics. Finally, many approaches have difficulty detecting multiple fractures in a single image, as they need precise localisation and segmentation to differentiate between normal bone structures and types of fracture. To overcome these challenges, techniques such as SSP, attention mechanisms, and feature fusion should be used to improve the accuracy and generalisability of fracture detection models that are employed in the proposed framework, FracNet.

3. Materials and method

3.1. Datasets

We used three datasets labelled “Fractured” and “Not Fractured.” The first dataset, FracAtlas, is a musculoskeletal bone fracture data set with annotations for DL tasks such as classification, localisation, and segmentation, containing X-ray images annotated in COCO, VGG, YOLO, and Pascal VOC formats [22]. The second data set is a collection of X-ray images designed to train a DL model specifically designed for fracture detection [23]. The third dataset, Wrist Fracture — X-rays, consists of X-ray images for detecting tiny fractures in the wrist [24]. In total, these three datasets comprise 10,763 images, which were divided into training and testing sets. Furthermore, we collected more than 50,000 X-ray images of different body parts for self-supervised pre-training, including the MURA dataset [15]. This collected dataset shares several features similar to those of the target datasets, enhancing the model’s ability to generalise across different types of fractures.

Table 1
Selected CNN models details.

–	Xception	Mobile	Efficient
Depth	71	53	82
Parameter Memory	88 MB	14 MB	20 MB
Parameters (Millions)	22.9	3.5	5.3
Image Input Size	299 × 299	224 × 224	224 × 224

3.2. The proposed framework: FracNet

We introduce a new framework called FracNet (Fig. 1), which integrates key components (from modified base models to final predictions) to make an accurate decision for fracture detection and addresses the generalisation issue across three different datasets.

We have chosen three DL models, including Xception, MobileNetv2, and EfficientNet, as base models, as listed in Table 1. These models were selected for their excellent performance on ImageNet, lightweight nature, and various structural designs. To enhance these models, we have added two attention mechanisms. The first is a self-attention layer that helps capture long-range dependencies and improve feature representation by allowing the model to focus on essential regions of the image. The second is the addition of a Squeeze-and-Excitation (SE) block, which adaptively recalibrates channel-wise feature responses by explicitly modelling the interdependencies between channels, thus boosting the model’s sensitivity to informative features.

These models are first trained using a self-supervised pre-training (SSP) approach on a large number of unlabelled X-ray images within the same domain [6]. More than 50,000 X-ray images are used in the SSP process. SSP is used to create pre-trained models with images similar to the target dataset. This approach helps to address the differences between features extracted from the target dataset and those from ImageNet, such as greyscale versus colour features. As a result, this process enables the models to adapt and specialise in the unique features of X-ray images. The newly pre-trained models are first fine-tuned with 70% of the frozen layers, training them specifically on the target datasets. Each of the three models is independently trained on each dataset. Additionally, we used Gradient-weighted Class Activation Mapping (Grad-CAM) to visualise the individual models’ decision-making processes. This technique highlights important regions in the X-ray images, providing insight into the areas on which the models focus for fracture detection. After training, features are extracted from the three models of each dataset and then fused together using horizontal stacking, also known as concatenation, along the horizontal axis. This fusion technique takes advantage of the complementary strengths of each model, leading to enhanced overall representation and robustness of the features. After preparing the features for each dataset, the features from the three datasets are combined using vertical stacking, which involves concatenating them along the vertical axis. This fusion increases the feature dimension and improves generalisation across different datasets. By integrating diverse feature sets, this approach supports the development of a more comprehensive and generalised model.

The combined features of the three datasets are then subjected to a feature selection process. To preprocess features and reduce dimensionality, a correlation matrix is computed and features with correlation coefficients exceeding a threshold of 0.95 are removed to avoid redundancy. Outliers are detected and removed to ensure model robustness by calculating z-scores for each feature and setting an outlier threshold of 3 standard deviations, excluding observations that exceed this threshold. A random forest classifier is used to assess the importance of features, configured with 100 trees and enabled for out-of-bag prediction and importance metrics. Features with importance scores below the mean are considered weak and subsequently removed. Principal Component Analysis (PCA) is then applied to reduce dimensionality further, retaining the number of components that explain 95% of the

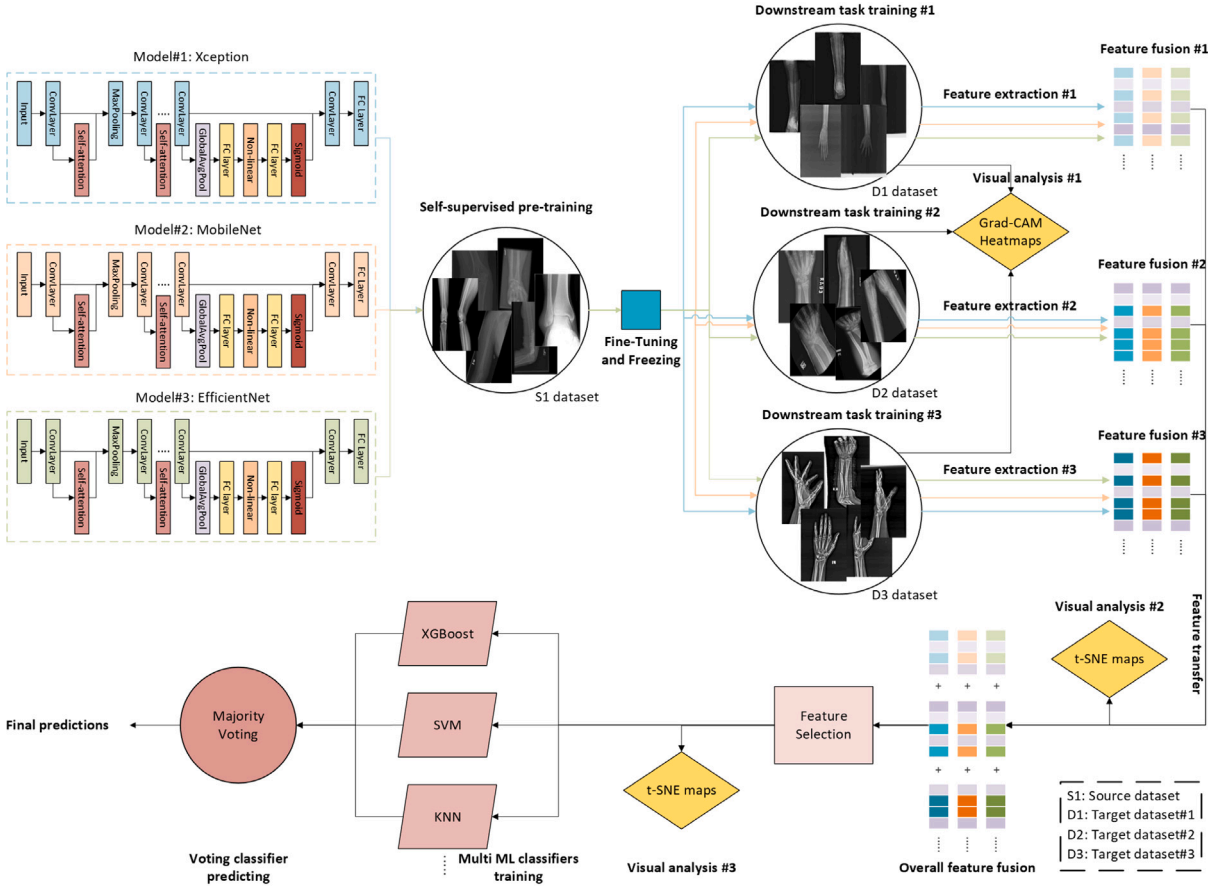

Fig. 1. Workflow of the FracNet framework for bone fracture detection.

variance, thus preserving significant features. t-Distributed Stochastic Neighbour Embedding (t-SNE) is used to visualise reduced features in a two-dimensional space, facilitating the understanding of clustering patterns within the data. The reduced feature set and the corresponding labels are then used to train several machine learning classifiers, ensuring a robust and generalised fracture detection model. The output of machine learning classifiers is combined using a majority voting technique to improve overall performance and reduce the likelihood of misclassification. The datasets used in this study were divided into training, validation, and testing sets. DL models were trained using the Adam optimiser, which combines the benefits of AdaGrad and RMSProp for efficient and robust training. A mini-batch size of 15 was selected to balance gradient stability and computational efficiency. Training was carried out for a maximum of 100 epochs to allow adequate learning while avoiding overfitting. Data were shuffled at the start of each epoch to ensure better generalisation. An initial learning rate of 0.001 was used, with a learning rate decay strategy applied to adjust the learning rate during training based on validation performance improvements.

The experiment was conducted using MATLAB 2024a, leveraging high-performance hardware to ensure efficiency and accuracy. The hardware included a Dell Precision 3680 Tower featuring an Intel® Core™ i9 14th Gen 14900 K processor with 24 cores, 32 threads, and a clock speed ranging from 3.2 GHz to 6.0 GHz, supported by 64 GB DDR5 ECC RAM running at 4400 MT/s for rapid data processing. For GPU acceleration, the system was equipped with an NVIDIA® RTX™ 4500 Ada Generation graphics card with 24 GB GDDR6 memory, ensuring excellent performance for computationally intensive tasks.

4. Results

The results section experimented by evaluating the base models for all datasets, presenting the FracNet result, and finally comparing it with the latest methods.

4.1. Base models

The performance of three base models – Xception, MobileNet, and EfficientNet – across three datasets has been listed in Table 2. The metrics include accuracy, specificity, recall, precision, and the F1 score. All models achieved nearly perfect scores on Dataset 1. Xception slightly outperformed the other models. In Dataset 2, all models achieved perfect scores of 100% in all metrics, indicating flawless performance. For Dataset 3, Xception had high performance with an accuracy of 98.8%, while MobileNet and EfficientNet had lower accuracies of 95.4% and 96.5%, respectively. The specificity and precision were consistently perfect across all models and datasets. However, recall varied more, especially in Dataset 3, which decreased for MobileNet and EfficientNet. These results indicate that Xception is the most reliable model across all datasets. The perfect scores in Dataset 2 suggest that this dataset may be less challenging or more consistent with the training data. The variability in Recall in Dataset 3 highlights the importance of dataset-specific characteristics in model performance, which requires further investigation.

Fig. 2 illustrates the predictions generated by three distinct models – Xception, MobileNetV2 and EfficientNet – in a set of diverse images. Each image is labelled with the predicted class (Fractured) and the corresponding confidence value. Although all models consistently predict the correct class, there is a significant difference in their confidence levels. For example, Xception shows confidence values of 77%, 87%, and 100%, MobileNetV2 exhibits 65%, 99%, and 54%, and EfficientNet shows 97%, 69%, and 78%. This variance suggests uncertainty in some predictions, indicating that relying on a single model may not provide the most reliable confidence estimates. To address this, combining

Table 2
Performance metrics of Xception, MobileNet, and EfficientNet across three datasets.

Metrics	Dataset1			Dataset2			Dataset3		
	Xception	Mobile	Efficient	Xception	Mobile	Efficient	Xception	Mobile	Efficient
Accuracy	99.7	99.4	99.4	100	100	100	98.8	95.4	96.5
Specificity	100	100	100	100	100	100	100	100	100
Recall	99.6	99.0	99.0	100	100	100	98.0	92.0	94.0
Precision	100	100	100	100	100	100	100	100	100
$F1_{Score}$	99.8	99.5	99.5	100	100	100	98.9	95.8	96.9

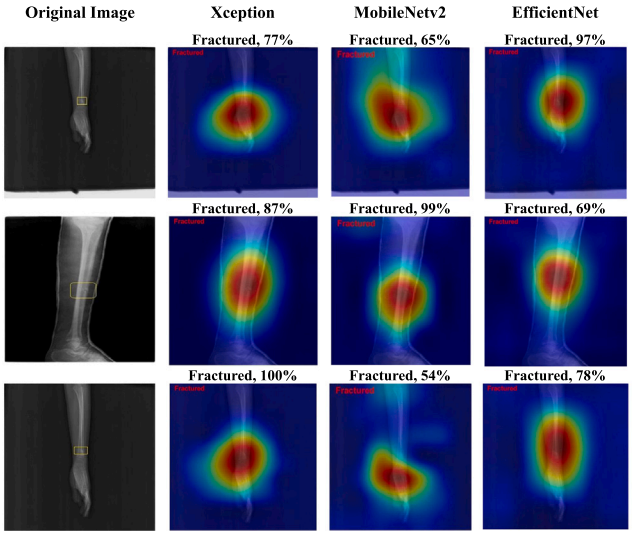

Fig. 2. Sample predictions with heatmap using Grad-CAM by Xception, MobileNetV2, and EfficientNet on Dataset1.

Table 3
Cross datasets evaluation in term of accuracy.

Train → Test	Xception (%)	Mobile (%)	Efficient (%)
D1 → D2	70	69	67
D1 → D3	76	75	73
D2 → D1	74	73	71
D2 → D3	69	68	66
D3 → D1	77	78	75
D3 → D2	68	66	64

the features of all models can improve the overall confidence values, resulting in more reliable fracture detection.

Table 3 presents the results of the cross-data sets for the Xception model, where it was trained on one data set and tested on another to evaluate its generalisability on three data sets. D1 (Dataset1), D2 (Dataset2), and D3 (Dataset3). When Xception was trained on D1 and tested on D2 and D3, it achieved accuracies of 70% and 76%, respectively. When trained in D2 and tested on D1 and D3, the precision was 74% and 69%, respectively. Finally, training on D3 and testing on D1 and D2 yielded accuracy of 77% and 68%, respectively. The results show that although the model performs reasonably well across various datasets, its accuracy typically drops when used on a dataset different from the one it was trained on. This emphasises the domain adaptation challenge and underscores the necessity of feature fusion across all datasets to enhance generalisation.

4.2. FracNet model

FracNet is crucial for fracture detection due to its ability to integrate and leverage the strengths of multiple models trained on diverse datasets, thus improving overall detection accuracy and confidence. Fig. 3 presents a test sample from dataset 1, which was evaluated using the Xception model trained separately on each dataset. An orthopaedic

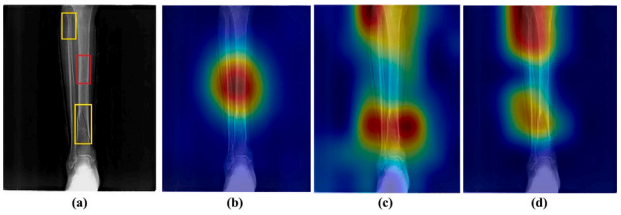

Fig. 3. Test sample from Dataset1 evaluated by Xception trained on three different datasets. (a) annotated by an orthopaedic surgeon, (b) Xception was trained on Dataset1, (c) Xception was trained on Dataset2, and Xception was trained on Dataset3.

surgeon annotated the sample, marking definite fractures with yellow boxes and a potential fracture requiring further investigation with a red box. Interestingly, each model identified different regions of the fractures, with the model trained on the same dataset (Dataset 1) accurately detecting the potential fractures. In contrast, models trained on different datasets identified the confirmed fractures. This variation underscores the importance of feature fusion in FracNet. By combining the outputs of models trained on multiple datasets, FracNet ensures higher accuracy and confidence in fracture detection.

Base models were employed to extract features from each dataset separately. The datasets were fused after feature extraction to create a comprehensive feature set. Feature selection was then applied to this fused dataset to remove redundant or irrelevant features, thereby enhancing the model's performance as shown in Fig. 4. This refined feature set was subsequently evaluated using five machine learning classifiers: XGBoost, SVM, KNN, decision tree, and Naive Bayes. Impressively, all classifiers achieved 100% across all evaluation metrics. Additionally, the output of majority voting also resulted in 100%. Although majority voting did not add value in this context, it may prove useful when incorporating more datasets or testing on unseen datasets, as it can enhance robustness and reliability.

To rigorously evaluate FracNet, we collected 100 X-ray fracture images from the literature and Google search. An expert verified these images. Fig. 5 presents some samples of the unseen test data. FracNet achieved a remarkable accuracy of 100% in all 100 images, accompanied by high confidence values, demonstrating its robust generalisation to unseen data.

4.3. FracNet-related ablation experiments

This section presents ablation experiments conducted to evaluate the performance of FracNet under different training configurations. Three experimental setups were considered to isolate the impact of specific components, including SSP, attention layers, feature fusion, and XGBoost classifier:

- S1: Training CNN models without SSP, using ImageNet.
- S2: Training CNN models with SSP, without attention layers.
- S3: Training CNN models with SSP, attention layers, and feature fusion with three ML classifiers without feature selection.

Table 4 provides a comprehensive breakdown of the accuracy results of ablation experiments conducted on three datasets (Dataset1,

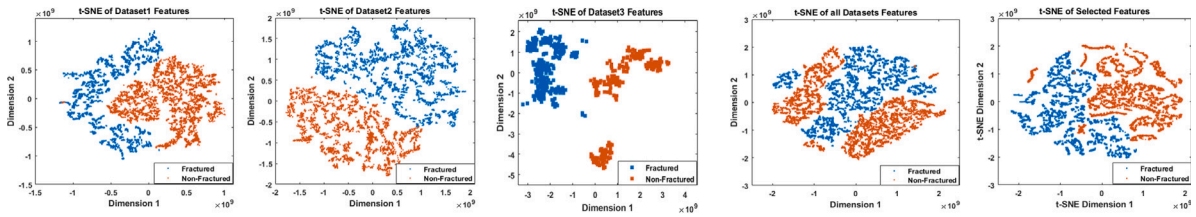

Fig. 4. t-SNE visualisation of extracted features after feature selection and fusion.

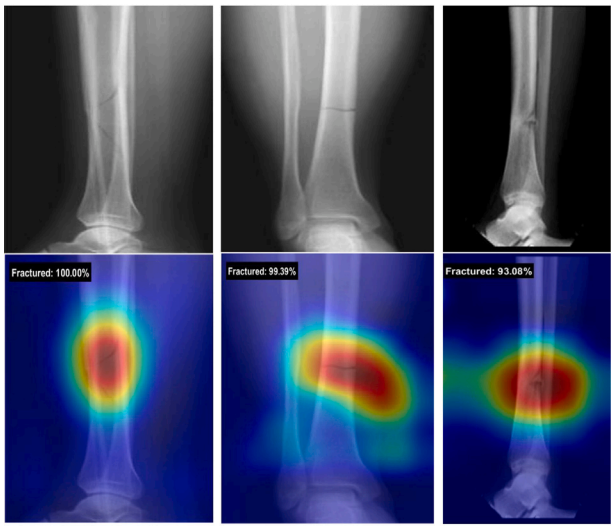

Fig. 5. Sample predictions by FracNet on unseen test data, demonstrating model generalisation.

Dataset2, and Dataset3) using three CNN architectures. Xception, MobileNet, and EfficientNet. The experiments were designed to evaluate the incremental benefits of SSP, attention layers, and feature fusion in training models for the detection of fractures.

The first experimental set-up, S1, serves as a baseline where models were trained without using the SSP technique, relying instead on weights from ImageNet. The accuracy achieved ranged from 75.4% to 82.5%, with Dataset 2 consistently producing the highest results. Both EfficientNet and Xception demonstrated similar performance across the different datasets, while MobileNet slightly lagged behind. This performance disparity emphasises the limitations of conventional transfer learning when applied directly to medical imaging tasks without domain adaptation. The results from S1 highlight the necessity of exploring advanced techniques to bridge the domain gap between ImageNet and task-specific datasets. Introducing SSP in S2 significantly increased accuracy across all models and datasets, with improvements of approximately 12%–17% compared to S1. For example, the accuracy of MobileNet in Dataset2 improved from 82.5% in S1 to 94.7% in S2. This demonstrates that SSP effectively adapts pre-trained models to domain-specific data, allowing for the extraction of more meaningful features. Notably, in S2, MobileNet outperformed both Xception and EfficientNet, achieving 94.6% accuracy in Dataset3. This highlights MobileNet's compatibility with the SSP-enhanced training pipeline, positioning it as a strong candidate for medical imaging tasks. The S3 represents the most advanced setup, combining SSP, attention layers, and feature fusion with three ML classifiers (XGBoost, SVM, and KNN) to achieve exceptional accuracy without feature selection. Across three datasets, XGBoost consistently outperformed with accuracies of 99.6%, 99.5%, and 99.0%, showcasing its robust ability to leverage fused features. SVM followed closely with accuracies of 99.2%, 99.6%, and 98.1%, demonstrating strong generalisation. KNN also performed reliably, achieving 99.2%, 98.9%, and 98.7%. These results highlight the effectiveness of integrating SSP and attention mechanisms in extracting high-quality features, making S3 a powerful and adaptable approach for achieving near-perfect classification in medical imaging tasks.

The final results are presented in the FracNet model results section, incorporating all of FracNet's steps, which achieved 100% accuracy.

4.4. Comparison with SOTA

Table 5 lists the latest SOTA DL methods for fracture classification, highlighting significant trends and findings. Xception achieved high accuracy (98%) for hip fractures using a large dataset of 3123 images [25], demonstrating the effectiveness of large datasets. In contrast, Inceptionv3 showed lower accuracy (86%) on a smaller dataset of 686 hip images [26], indicating the potential limitations of smaller datasets. ResNet demonstrated robustness with an AUC of 89% for knee fractures on a dataset of 6768 images [27]. DenseNet achieved the highest AUC (99%) for proximal femur fractures using 4577 images [28], highlighting DenseNet's efficacy. Hybrid models such as U-Net and ResNet achieved an AUC of 95% for mandible fractures on a large data set of 22,256 images [29], showcasing the benefits of combining models. Feature fusion methods were notable success, with Guan et al. [30] achieving a precision of 88% for femur fractures [30] and Alzubaidi et al. (2024) achieving an accuracy of 99% for shoulder fractures [20], emphasising the strength of feature fusion. FracNet, our proposed method, achieved a perfect accuracy of 100% trained and tested on 10,763 images, demonstrating the superior performance and significance of combining attention mechanisms with feature fusion for improved fracture detection.

5. Conclusion

This study introduces FracNet, an end-to-end DL framework specifically developed to address the challenges of bone fracture detection in medical imaging. FracNet utilises advanced techniques such as self-supervised learning, feature fusion, attention mechanisms, and feature selection to improve diagnostic accuracy, explainability, and generalisation across various datasets. The framework achieved an impressive detection accuracy of 100% on multiple datasets and demonstrated strong performance on unseen test data, significantly surpassing existing methods. The strengths of FracNet lie in its ability to integrate diverse features from multiple models and datasets, ensuring improved reliability and confidence in predictions. Incorporating attention mechanisms and visual explanation tools further improves decision transparency, fostering trust among clinicians. FracNet's scalability to other medical imaging applications also highlights its potential for a broader impact in the healthcare field. Despite these accomplishments, certain limitations require further investigation. Although the model exhibits excellent performance across the tested datasets, its adaptability to different imaging modalities and real-world clinical environments involves validation. Moreover, the computational cost of training and deploying FracNet in resource-constrained settings presents challenges that future research should address. Another area of improvement includes fine-tuning the feature selection process to ensure optimal trade-offs between model complexity and performance. FracNet offers a robust and reliable solution for fracture detection, addressing critical

Table 4
FracNet-related ablation experiments in terms of accuracy.

Metrics	Dataset1			Dataset2			Dataset3		
	Xception	Mobile	Efficient	Xception	Mobile	Efficient	Xception	Mobile	Efficient
S1	79.5	77.3	79.4	80.7	82.5	80.1	77.8	75.4	76.2
S2	92.5	93.6	92.3	94.6	94.7	93.6	91.9	94.6	92.5
	XGBoost	SVM	KNN	XGBoost	SVM	KNN	XGBoost	SVM	KNN
S3	99.6	99.2	99.2	99.5	99.6	98.9	99.0	98.1	98.7

Table 5
Comparison of SOTA deep learning methods for fracture classification.

Reference and Year	Body Part	CNN Model	Number of samples	Best results (%)
[26]	Hip	Inceptionv3	686	Accuracy = 86
[27]	Knee	ResNet	6768	AUC = 89
[31]	Ankle	PCANet	5534	Accuracy = 72
[32]	Vertebra	ResNeXt	1306	Accuracy = 73
[29]	Mandible	U-Net and ResNet	22 256	AUC = 95
[33]	Several body parts	DL algorithm (Rayvolve)	5865	AUC = 92
[30]	femur	ResNeXt	3842	Precision = 88
[34]	Ankle	Inception V3 and Renet-50	2100	Sensitivity = 98
[35]	Foot	ResNet-50	3993	AUC = 96
[36]	Rib	5 DL	2000	Accuracy: 92
[37]	Vertebral	YOLOv4 and ResUNet	3634	Precision = 99, 74, and 94
[38]	Femur	Feature Fusion	1124+4014	AUC = 98
[20]	Shoulder	Feature Fusion	8379+563	Accuracy = 99
FracNet(ours)	Different body parts	Attention & Feature Fusion	10,763	Accuracy = 100

gaps in current DL approaches. This framework paves the way for improved diagnostic workflows in orthopaedic care and beyond by combining innovative methodologies with a focus on clinical applicability. Future research will concentrate on further optimising FracNet for use in diverse clinical environments and exploring its application to other medical imaging tasks.

CRediT authorship contribution statement

Haider A. Alwzwazy: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Laith Alzubaidi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Zehui Zhao:** Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation, Formal analysis. **Yuantong Gu:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Funding

The authors would like to acknowledge the support received through the following funding schemes of the Australian Government: Australian Research Council (ARC) Industrial Transformation Training Centre (ITTC) for Joint Biomechanics under Grant IC190100020.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets are public and cited within the article.

References

[1] J. Jung, J. Dai, B. Liu, Q. Wu, Artificial intelligence in fracture detection with different image modalities and data types: A systematic review and meta-analysis, PLOS Digit. Heal. 3 (1) (2024) e0000438.
[2] A.A. Saihood, M.A. Hasan, M.A. Fadhel, L. Alzubaid, A. Gupta, Y. Gu, et al., Multiside graph neural network-based attention for local co-occurrence features fusion in lung nodule classification, Expert Syst. Appl. 252 (2024) 124149.
[3] B. Guan, J. Yao, G. Zhang, X. Wang, Thigh fracture detection using deep learning method based on new dilated convolutional feature pyramid network, Pattern Recognit. Lett. 125 (2019) 521–526.
[4] Z. Alammar, L. Alzubaidi, J. Zhang, Y. Li, A. Gupta, Y. Gu, Generalisable deep learning framework to overcome catastrophic forgetting, Intell. Syst. Appl. 23 (2024) 200415.
[5] A.S. Albahri, A.M. Duhaim, M.A. Fadhel, A. Alnoor, N.S. Baqer, L. Alzubaidi, O.S. Albahri, A.H. Alamoodi, J. Bai, A. Salhi, et al., A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion, Inf. Fusion 96 (2023) 156–191.
[6] L. Alzubaidi, M.A. Fadhel, F. Hollman, A. Salhi, J. Santamaria, Y. Duan, A. Gupta, K. Cutbush, A. Abbosh, Y. Gu, SSP: self-supervised pertaining technique for classification of shoulder implants in x-ray medical images: a broad experimental study, Artif. Intell. Rev. 57 (10) (2024) 261.
[7] S. Wang, D. Zhu, J. Chen, J. Bi, W. Wang, Deepfake face discrimination based on self-attention mechanism, Pattern Recognit. Lett. 183 (2024) 92–97.
[8] Z. Alammar, L. Alzubaidi, J. Zhang, Y. Li, W. Lafta, Y. Gu, Deep transfer learning with enhanced feature fusion for detection of abnormalities in x-ray images, Cancers 15 (15) (2023) 4007.
[9] L. Alzubaidi, M.A. Fadhel, A. Albahri, A. Salhi, A. Gupta, Y. Gu, Domain adaptation and feature fusion for the detection of abnormalities in X-Ray forearm images, in: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2023, pp. 1–5.
[10] E. Pechlivanidou, I. Antonopoulos, R.E. Margariti, Gender equality challenges in orthopaedic surgery: a systematic review, Int. Orthop. 47 (9) (2023) 2143–2171.

[11] N. Salari, H. Ghasemi, L. Mohammadi, M.H. Behzadi, E. Rabieenia, S. Shohaimi, M. Mohammadi, The global prevalence of osteoporosis in the world: a comprehensive systematic review and meta-analysis, J. Orthop. Surg. Res. 16 (2021) 1–20.

[12] S. Hansen, P. Schwarz, J. Rumessen, A. Linneberg, L.L. Kårhus, Osteoporosis and bone fractures in patients with celiac disease: A nationwide cohort study, Bone 177 (2023) 116913.

[13] I. Thøfner, H.P. Hougen, C. Villa, N. Lynnerup, J.P. Christensen, Pathological characterization of keel bone fractures in laying hens does not support external trauma as the underlying cause, PLoS One 15 (3) (2020) e0229735.

[14] S.W. Chung, S.S. Han, J.W. Lee, K.-S. Oh, N.R. Kim, J.P. Yoon, J.Y. Kim, S.H. Moon, J. Kwon, H.-J. Lee, et al., Automated detection and classification of the proximal humerus fracture by using deep learning algorithm, Acta Orthop. 89 (4) (2018) 468–473.

[15] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R.L. Ball, et al., Mura: Large dataset for abnormality detection in musculoskeletal radiographs, 2017, arXiv preprint arXiv:1712.06957.

[16] F. Uysal, F. Hardalaç, O. Peker, T. Tolunay, N. Tokgöz, Classification of shoulder x-ray images with deep learning ensemble models, Appl. Sci. 11 (6) (2021) 2723.

[17] H.X. Huynh, H.B.T. Nguyen, C.A. Phan, H.T. Nguyen, Abnormality bone detection in X-Ray images using convolutional neural network, in: Context-Aware Systems and Applications, and Nature of Computation and Communication: 9th EAI International Conference, ICCASA 2020, and 6th EAI International Conference, ICTCC 2020, Thai Nguyen, Vietnam, November 26–27, 2020, Proceedings 9, Springer, 2021, pp. 31–43.

[18] G. Kitamura, C.Y. Chung, B.E. Moore, Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation, J. Digit. Imaging 32 (2019) 672–677.

[19] I. Kandel, M. Castelli, A. Popovič, Musculoskeletal images classification for detection of fractures using transfer learning, J. Imaging 6 (11) (2020) 127.

[20] L. Alzubaidi, A. Salhi, M. A. Fadhel, J. Bai, F. Hollman, K. Italia, R. Pareyon, A. Albahri, C. Ouyang, J. Santamaría, et al., Trustworthy deep learning framework for the detection of abnormalities in X-ray shoulder images, PLoS One 19 (3) (2024) e0299545.

[21] R. Abdelazim, E.M. Fouad, Artificial intelligent-driven decision-making for automating root fracture detection in periapical radiographs, BDJ Open 10 (1) (2024) 76.

[22] FracAtlas Dataset1, FracAtla Image Dataset, https://www.kaggle.com/datasets/akshayramakrishnan28/fracture-classification-dataset. (Accessed 22 May 2024).

[23] F. Dataset2, Fracture Classification Dataset, https://www.kaggle.com/datasets/devbatrax/fracture-detection-using-x-ray-images. (Accessed 22 May 2024).

[24] W. Fracture, Wrist Fracture Dataset3, https://data.mendeley.com/datasets/xbdsnzr8ct/1. (Accessed 23 May 2024).

[25] Y. Yamada, S. Maki, S. Kishida, H. Nagai, J. Arima, N. Yamakawa, Y. Iijima, Y. Shiko, Y. Kawasaki, T. Kotani, et al., Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs, Acta Orthop. 91 (6) (2020) 699–704.

[26] C. Lee, J. Jang, S. Lee, Y.S. Kim, H.J. Jo, Y. Kim, Classification of femur fracture in pelvic X-ray images using meta-learned deep neural network, Sci. Rep. 10 (1) (2020) 1–12.

[27] A. Lind, E. Akbarian, S. Olsson, H. Nåsell, O. Sköldenberg, A.S. Razavian, M. Gordon, Artificial intelligence for the classification of fractures around the knee in adults according to the 2018 AO/OTA classification system, PLoS One 16 (4) (2021) e0248809.

[28] L. Oakden-Rayner, W. Gale, T.A. Bonham, M.P. Lungren, G. Carneiro, A.P. Bradley, L.J. Palmer, Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study, Lancet Digit. Heal. 4 (5) (2022) e351–e358.

[29] X. Wang, Z. Xu, Y. Tong, L. Xia, B. Jie, P. Ding, H. Bai, Y. Zhang, Y. He, Detection and classification of mandibular fracture on CT scan using deep convolutional neural network, Clin. Oral Investig. (2022) 1–9.

[30] B. Guan, J. Yao, S. Wang, G. Zhang, Y. Zhang, X. Wang, M. Wang, Automatic detection and localization of thighbone fractures in X-ray based on improved deep learning method, Comput. Vis. Image Underst. 216 (2022) 103345.

[31] N.A. Farda, J.-Y. Lai, J.-C. Wang, P.-Y. Lee, J.-W. Liu, I.-H. Hsieh, Sanders classification of calcaneal fractures in CT images with deep learning and differential data augmentation techniques, Injury 52 (3) (2021) 616–624.

[32] H.-Y. Chen, B.W.-Y. Hsu, Y.-K. Yin, F.-H. Lin, T.-H. Yang, R.-S. Yang, C.-K. Lee, V.S. Tseng, Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs, PLoS One 16 (1) (2021) e0245992.

[33] M. Dupuis, L. Delbos, R. Veil, C. Adamsbaum, External validation of a commercially available deep learning algorithm for fracture detection in children, Diagn. Interv. Imaging 103 (3) (2022) 151–159.

[34] S. Ashkani-Esfahani, R.M. Yazdi, R. Bhimani, G.M. Kerkhoffs, M. Maas, C.W. DiGiovanni, B. Lubberts, D. Guss, Detection of ankle fractures using deep learning algorithms, Foot Ankle Surg. (2022).

[35] Y. Wang, Y. Li, G. Lin, Q. Zhang, J. Zhong, Y. Zhang, K. Ma, Y. Zheng, G. Lu, Z. Zhang, Lower-extremity fatigue fracture detection and grading based on deep learning models of radiographs, Eur. Radiol. 33 (1) (2023) 555–565.

[36] S.-T. Huang, L.-R. Liu, H.-W. Chiu, M.-Y. Huang, M.-F. Tsai, Deep convolutional neural network for rib fracture recognition on chest radiographs, Front. Med. 10 (2023).

[37] L.-W. Cheng, H.-H. Chou, Y.-X. Cai, K.-Y. Huang, C.-C. Hsieh, P.-L. Chu, I.-S. Cheng, S.-Y. Hsieh, Automated detection of vertebral fractures from X-ray images: A novel machine learning model and survey of the field, Neurocomputing 566 (2024) 126946.

[38] J. Schilcher, A. Nilsson, O. Andlid, A. Eklund, Fusion of electronic health records and radiographic images for a multimodal deep learning prediction model of atypical femur fractures, Comput. Biol. Med. 168 (2024) 107704.