Deep Learning for Bone Fracture Detection: A Comprehensive Review of Models, Datasets, and Diagnostic Performance

Galiveeti Poornima[1], Manju Bargavi Ranjeth[2]

[1] Lincoln University, Malaysia and S-Vyasa University, Bangalore, India; [2] Department of Computer Science and IT, Jain (Deemed-to-be University), Bangalore

pdf.Galiveeti@lincoln.edu.my

Abstract: Bone fractures are a prevalent kind of musculoskeletal trauma globally, and their prompt and precise diagnosis is essential for the prevention of long-term complications. Traditional diagnostic techniques are effective but rely on professional radiologists and are susceptible to human error, especially when a fracture is minor or concealed. In the era of artificial intelligence (AI), deep learning (DL) has emerged as a transformative approach for automated and enhanced fracture identification across various imaging modalities, including X-ray, computed tomography (CT), and magnetic resonance imaging (MRI). A comprehensive analysis of the latest deep learning methods employed in bone fracture detection is provided. We categorize earlier models, such as ConvNets, attention-based models, and hybrid methodologies, and evaluate their diagnostic effectiveness using performance metrics including sensitivity, specificity, and area under the receiver operating characteristic curve. Publicly accessible datasets, along with their characteristics and constraints, are provided to demonstrate the issues associated with data annotation, imbalance, and generalization. Furthermore, we examine interpretability methodologies, lightweight models for real-time applications, and their connection with clinical workflows. This study identifies existing research deficiencies and proposes new avenues, including federated learning, self-supervised learning, and domain adaptation techniques, to improve model robustness and expand generalization across diverse healthcare contexts. The authors aim for this paper to serve as a valuable reference for researchers and practitioners in the field of artificial intelligence in medical imaging diagnostics, describing the latest advancements.

Keywords: Deep Learning; Bone Fracture Detection; CNN; Radiology AI; Medical Imaging

Introduction

Bone fractures represent a significant global health issue, with trauma, osteoporosis, and sports accounting for millions of instances annually. Radiographic imaging, primarily utilizing X-rays, remains the benchmark for the initial assessment of fractures [1]. Frank's saccular origins and sidewall blebs are the predominant etiologies of subarachnoid hemorrhage; however, subtle fractures, overlapping structures, and interobserver variability among radiologists may lead to diagnostic failures or delayed treatment, particularly in emergency and rural facilities with limited radiological expertise [2].

Artificial intelligence (AI), intensive learning, has recently attained remarkable success in the diagnosis of medical imaging. Deep learning models have shown significant potential in diagnosing bone fractures,

with sensitivity and specificity levels that are sufficient to aid clinical decision-making and mitigate diagnostic errors [3]. These algorithms provide not only more consistent and expedited evaluations but also decision support through heatmaps and elucidated outputs, fostering enhanced trust and integration within healthcare workflows.

Despite the growing body of papers and prototype systems developed for fracture diagnosis via deep learning, a systematic overview of the available literature remains absent. It is essential to consolidate model designs, datasets, appropriate evaluation criteria, interpretability tools, and deployment techniques in clinical settings to advance the advances achieved to date. This study aims to address this deficiency by providing a critical examination of current advancements and the unresolved issues that remain.

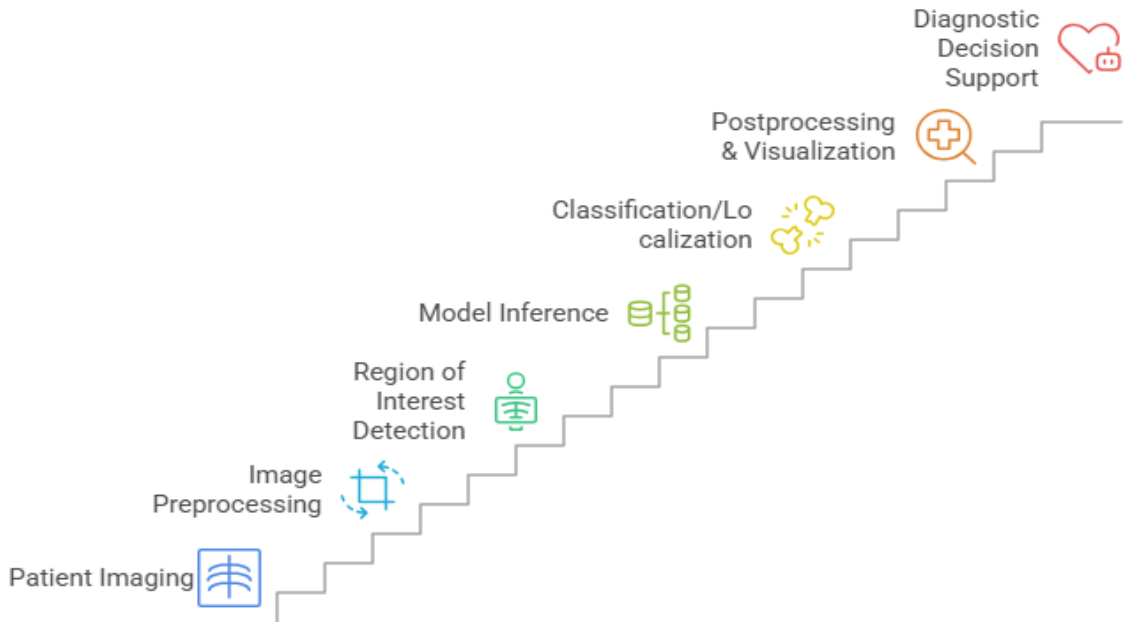

Figure. 1. Deep Learning Workflow for Bone Fracture Detection. The end-to-end pipeline encompasses stages ranging from patient imaging and preprocessing to region-of-interest detection, model inference, classification/localization, postprocessing, and ultimately, diagnostic decision support. Each step represents a critical component in automated fracture analysis systems.

The study is organized as follows: Section 2 elucidates certain technical and clinical aspects essential for comprehending bone fracture diagnosis and imaging modalities. Section 3 examines deep learning frameworks utilized in medical imaging. Finally, Section 4 delineates the datasets used in fracture detection research. Section 5 addresses benchmark outcomes and assessment standards. Section 6 delineates constraints and unresolved issues. Section 9 provides a conclusion along with insights and prospective directions.

Background

Overview of Bone Fracture Types and Clinical Challenges

Bone fractures are a varied category of injuries characterized by varying causes, forms, and levels of severity. Fractures can be clinically classified into several categories: open (complex), closed, comminuted, greenstick, spiral, compression, avulsion, stress, and pathologic. All require sophisticated reasoning and management. Open fractures present a significant infection risk and necessitate prompt assessment; conversely, clinical examinations may neglect less apparent injuries like stress fractures [4].

Accurate identification of the fracture type is crucial for prompt treatment, precise prognosis, and effective prevention. Diagnosis may be hindered by inter-observer and intra-observer variances, including weariness, low-contrast regions of the image, and structural overlap [5]. These constraints are especially apparent in time-sensitive clinical situations, where swift triage is essential.

Medical Imaging Modalities Used: X-ray, CT, and MRI

Radiology is a crucial determinant for the identification and diagnosis of bone fractures. Among the diverse modalities:

- X-ray remains the most prevalent diagnostic method, providing a rapid, cost-effective, and efficient means of identifying various fracture types. However, it may overlook a minor or hairline fracture, especially in a challenging anatomical region [6].

- CT is an exceptional instrument for complicated and intra-articular fractures, as it offers sectional imaging. It is frequently utilized for orthopedic trauma situations [7].

- MRI is particularly beneficial for diagnosing stress fractures, bone marrow edema, and associated soft tissue injuries, even in the absence of apparent structural disruptions on plain radiographs or CT scans [8].

The clinical presentation, anatomical site, and urgency of diagnosis dictate the chosen method. Understanding these modalities is crucial for developing strong, generalizable AI-based fracture detection models.

Table 1. Classification of Bone Fracture Types and Their Associated Imaging Modalities[1] .

Fracture Type	Common Imaging Modalities	Clinical Notes
Closed Fracture	X-ray	Bone breaks but does not puncture the skin
Open (Compound) Fracture	X-ray, CT	Bone breaks through the skin; high infection risk
Comminuted Fracture	X-ray, CT	Bone shatters into multiple fragments
Greenstick Fracture	X-ray (Common in	Bone bends and cracks, which are common in

[1] https://my.clevelandclinic.org/health/diseases/15241-bone-fractures

	Pediatrics)	children
Spiral Fracture	X-ray, MRI	Twisting force causes a spiral break pattern
Compression Fracture	X-ray, CT, MRI	Collapse of bone, often in the spine/osteoporosis
Avulsion Fracture	X-ray, Ultrasound (Children)	Tendon or ligament pulls off a bone fragment
Stress Fracture	MRI, Bone Scan, X-ray	Small cracks from repetitive stress or overuse
Pathological Fracture	X-ray, MRI, CT (for lesion assessment)	Occurs due to a disease weakening the bone (e.g., tumor)

Why Fracture Detection is Complex?

Despite advancements in technology, detecting fractures remains a formidable challenge, mainly due to the inherent complexities of the issue.

- Visual subtlety: Certain fractures, including hairline and stress fractures, may be inconspicuous and potentially overlooked, particularly in initial-stage pictures or low-resolution scans [9].

- Anatomical Diversity: The skeletal structures exhibit significant variation throughout distinct age groups, body sizes, and populations. Pediatric bones exhibit a distinct fracture pattern (e.g., greenstick) compared to adults [10].

- Imaging Deficiencies: The compromised quality of pictures significantly impedes algorithmic interpretation; factors such as noise, motion blur, and convergence issues diminish image quality, leading to erroneous diagnoses [11].

- Imbalance Issue: In a clinical dataset, the number of patients without fractures significantly exceeds those with fractures, complicating the training of AI models [12].

- Overlapping Anatomy: In X-rays, overlapping bones (e.g., wrist, ribs) may obscure fracture lines, complicating both manual and automatic evaluations [13].

These factors underscore the significance of intelligent, data-driven systems that can achieve expert-level performance while offering application flexibility, scalability, and consistency.

Overview of Deep Learning in Medical Imaging

Evolution from Machine Learning to Deep Learning in Healthcare

Over the past decade, artificial intelligence (AI) has been swiftly incorporated into medical imaging. Initial AI applications in healthcare utilized traditional machine learning methods, such as support vector

machines (SVMs), decision trees, and k-nearest neighbors. These approaches required manual feature extraction, which depended significantly on domain-specific design.

The emergence of deep learning (DL), particularly following the success of convolutional neural networks (CNNs) in the ImageNet competition, marked the beginning of a new era in machine learning. Deep models can proficiently acquire structured hierarchical features from raw input data, which are ideally suited for the intricate visual patterns present in medical imaging [14]. Furthermore, regarding the diagnosis of fractures, deep learning software can reliably identify minor indicators, such as hairline fractures or uneven cortical margins, which are challenging to detect using conventional approaches [15].

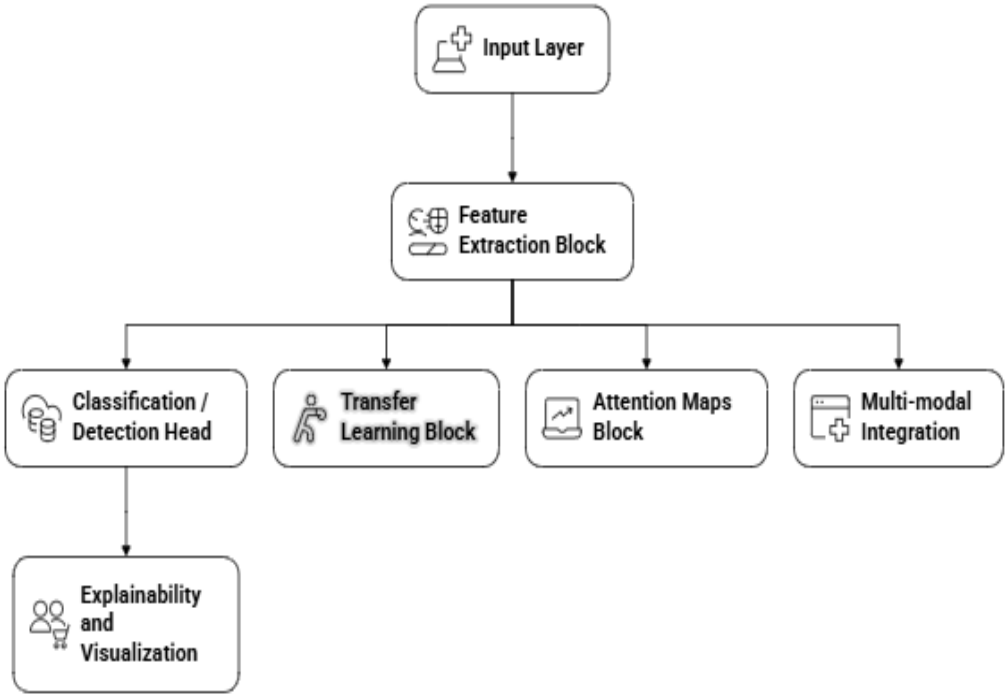

Figure. 2. Modular Deep Learning Architecture for Fracture Detection. This schematic illustrates the significant components of deep learning frameworks used in bone fracture analysis, including input layers, feature extraction, classification/detection heads, attention modules, transfer learning blocks, and multimodal integration. The output is refined through the use of explainability and visualization tools to enhance clinical interpretability.

Core Deep Learning Architectures

Convolutional Neural Networks (CNNs) underpin the majority of computer vision tasks, including the identification of bone fractures, which is pertinent to our research. Convolutional Neural Networks (CNNs) are highly adept at identifying spatial patterns, such as analyzing X-ray, CT, and MRI images.

Prominent topologies include VGGNet[2], ResNet[3], DenseNet[4], and EfficientNet[5], each exhibiting distinct trade-offs in terms of depth, computational expense, and performance.

Recurrent Neural Networks (RNNs) are primarily recognized for their application in sequential data rather than medical imaging; nonetheless, they can be utilized for multiple time-point MRIs and CT scans. Specific versions, such as LSTM networks[6], are occasionally employed in multi-frame analysis or dynamic imaging contexts.

Transformers and their vision-specific adaptation, ViT[7], are recent advancements in the domain of medical image analysis. Transformers, in contrast to CNNs that utilize local filters, employ a self-attention mechanism to capture global context. This is particularly advantageous for detecting fractures in many anatomical regions and for identifying minor deformations worldwide. Vision Transformers (ViTs) have gained prominence in radiology due to their intrinsic scalability and superior performance compared to Convolutional Neural Networks (CNNs) in large-data contexts.

Transfer Learning and Its Relevance in Radiology

Transfer learning has been a critical factor in the application of deep learning models in healthcare. Due to the scarcity of extensive annotated medical datasets, a model pre-trained on a general image dataset, such as ImageNet, is frequently adapted for specialized medical applications [16]. Transfer learning offers the benefit of minimizing training duration and enhancing performance on domain-specific tasks, especially when labeled radiographs are limited.

Transfer learning facilitates the reutilization of standard backbones, previously trained on millions of natural images, for fracture recognition, allowing the last layers to be adaptively retrained to differentiate between normal and broken bones. Static augmentation methods, including freezing lower layers, employing varying learning rates, and implementing multi-service learning, are commonly utilized.

Explainability in Deep Learning Models

Deep learning models are recognized for their superior performance; nonetheless, they have been extensively criticized as "black boxes" due to their lack of interpretability. In medical imaging, interpretability is crucial for clinical trust and legal compliance. Various strategies have been developed to meet this requirement:

[2] https://medium.com/@siddheshb008/vgg-net-architecture-explained-71179310050f
[3] https://www.mygreatlearning.com/blog/resnet/
[4] https://medium.com/@alejandro.itoaramendia/densenet-a-complete-guide-84fedef21dcc
[5] https://www.geeksforgeeks.org/computer-vision/efficientnet-architecture/
[6] http://colah.github.io/posts/2015-08-Understanding-LSTMs/
[7] https://www.v7labs.com/blog/vision-transformer-guide

- Grad-CAM (Gradient-weighted Class Activation Mapping)8 is utilized to identify the significant areas of a picture that influenced the model's choice. Frequently employed in the examination of fractures to identify the suspected fracture line.
- SHAP (SHapley Additive exPlanations)9: Evaluates feature significance by assigning the contributions of individual pixels or regions to the model's output.
- Saliency maps10, Integrated Gradients11, and LIME12 are employed to depict model attention and facilitate the comprehension of clinician decisions.

Explainable AI methodologies can enhance user confidence and facilitate model debugging and data curation, ensuring that the model prioritizes clinically pertinent features over artifacts [17].

Table 2. Comparative Analysis of Deep Learning Architectures for Bone Fracture Detection.

Aspect	CNN (Convolutional Neural Networks)	ViT (Vision Transformers)	Hybrid Models (CNN + Transformer)
Feature Learning	Extracts local spatial features via convolutional filters; ResNet and DenseNet show strong baseline performance [18]	Captures global dependencies with self-attention; superior for subtype classification [19]	Combines CNN's local inductive biases with Transformer's global modeling [20] [21]
Input Representation	Processes fixed pixel grids (e.g., 224×224 RGB images) [18]	Converts images into sequences of patches with positional embeddings [19]	Uses CNN-derived features as input to Transformer layers [20] [22]
Context Capture	Limited to local receptive fields; deeper layers expand context [18]	Captures global image context across all positions simultaneously [19]	Fuses local CNN features with global Transformer attention for enriched context [20] [21]
Computational Cost	Low to moderate; MobileNetV2 achieves efficient edge deployment [23]	High; ViTs require significant compute, especially for large X-rays [19]	Moderate to high; fusion modules add inference complexity [20]
Data Requirement	Moderate; transfer learning enables good results with limited data [18]	Very high; requires large-scale pretraining datasets [19] [22]	Moderate to high; generalizes better than ViT alone when trained with self-supervision [20] [22]
Training Time	Fast convergence with transfer learning [18] [23]	Slower training due to heavy attention computations [19]	Medium; depends on CNN–Transformer fusion design [20]

[8] https://medium.com/@divakar1591/grad-cam-gradient-weighted-class-activation-mapping-8e1aeaf96d94
[9] https://www.geeksforgeeks.org/machine-learning/shap-a-comprehensive-guide-to-shapley-additive-explanations/
[10] https://medium.com/@bijil.subhash/explainable-ai-saliency-maps-89098e230100
[11] https://www.tensorflow.org/tutorials/interpretability/integrated_gradients
[12] https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/

Interpretability	High; Grad-CAM and saliency maps widely applied [18] [24]	Moderate; attention maps provide insights but can be noisy [19]	Moderate to high; hybrid attention enhances transparency [20] [21]
Suitability for Small Datasets	Effective with transfer learning and augmentation [18] [23]	Poor unless pretrained on very large datasets [19] [22]	Performs better than ViT alone; self-supervised pretraining improves performance [20] [22]
Real-Time Inference	Feasible; MobileNetV2 achieves rapid edge inference with >94% accuracy [23]	Not optimal; high GPU/TPU demand limits bedside deployment [19]	Balanced; hybrid models optimized for robustness and generalization [20] [25]
Use in Fracture Detection	Widely used: ResNet-50 (97.6%), DenseNet, MobileNetV2 (94.8%) [18] [23]	Emerging but impactful: ViT improves clinician diagnostic accuracy to 97% with CAD support [19]	Strong performance: FracNet achieves near-perfect accuracy across benchmarks [20]; robust under noise [25]

Table 3. Overview of Deep Learning Architectures for Fracture Detection: Key Models, Strengths, and Challenges.

Architecture Type	Key Models	Strengths	Challenges
CNNs	ResNet, DenseNet, EfficientNet [18] [26] [27]	High accuracy, explainability (Grad-CAM) [28] [29]	May miss the global context [30] [31]
ViTs	ViT-B/16, Swin Transformer [32] [33]	Global attention, transferable to varied data [32] [33]	Data-intensive, slower convergence [32]
Hybrid	CNN+RNN [28] [34] , Ensembles [35]	Robust generalization, cross-view learning [34] [35]	Higher complexity and inference time [35]
Lightweight Models	MobileNet, TinyML [26] [30]	Fast, low-power, edge-friendly [26] [30]	Slightly lower accuracy, less expressive [30]

Datasets for Bone Fracture Detection

Public Dataset Descriptions

1. **MURA (Musculoskeletal Radiographs)**[13]
   - **Modality & Scope:** Upper extremity X-ray studies covering shoulder, elbow, forearm, wrist, hand, finger, and humerus—each study contains one or more image views.
   - **Size & Labels:** Approximately 14,863 studies, totaling 40,561 images, with labels as normal or abnormal assigned by radiologists.

[13] https://aimi.stanford.edu/datasets/mura-msk-xrays

- **Test Set Quality:** Includes a radiologist-annotated holdout set to benchmark model and human performance.
- **Challenges:** Potential label inaccuracies—especially in degenerative joint disease—risk of mislabeling; expert visual validation showed sensitivities around 60% and specificity around 82%.

2. **RSNA Pediatric Bone Age Dataset**[14]
   - **Modality:** Pediatric hand radiographs.
   - **Size & Annotations:** ~14,236 images labeled with skeletal bone age and sex; split into training, validation, and test sets.
   - **Focus:** Bone age estimation—less directly applicable to fracture detection but valuable for transfer learning or related radiological tasks.

3. **FracAtlas**[15]
   - **Modality:** X-ray images spanning hand, leg, hip, and shoulder regions, including various view orientations (frontal, lateral, oblique).
   - **Size & Annotations:**
     - ✓ 4,083 images in total
     - ✓ 717 abnormal scans comprising 922 fracture instances
     - ✓ Rich annotations: classification labels, bounding boxes, and segmentation masks (COCO, Pascal VOC, YOLO formats).
   - **Coverage:** Includes multi-view scans, hardware presence (e.g., fixations), and mixes of body parts.
   - **Challenges:** Potential for human annotation error despite multi-stage expert verification; regional bias (collected from specific hospitals in Bangladesh) may affect generalizability

4. **Other Notable Datasets**
   - **Custom Multi-modal Dataset (HBFMID)**[16]**:** Contains 641 images labeled across 10 fracture classes using X-ray, CT, and MRI; suffers from overfitting due to small size, despite high reported precision (95%) and recall.
   - **Large Multi-region X-ray Dataset**[17] **(~9,463 images):** Includes fractures and normals across multiple anatomical areas; used in transfer learning frameworks, achieving high accuracy (e.g., >99%).
   - **Pediatric Wrist Dataset (GRAZPEDWRI-DX)**[18]**:** Over 10,600 studies involving wrist fractures and other annotations; useful for pediatric applications, though not fracture-centric.

Challenges Across Datasets

- **Imbalance:** Many datasets contain significantly fewer fractured samples compared to normal ones (e.g., MURA, FracAtlas).
- **Annotation Quality:** Labels may lack granularity (e.g., only classification in MURA) or risk inaccuracy despite expert review (FracAtlas).

[14] https://datasetninja.com/rsna-bone-age
[15] https://www.nature.com/articles/s41597-023-02432-4
[16] https://www.kaggle.com/datasets/orvile/human-bone-fractures-image-dataset-hbfmid/code
[17] https://www.kaggle.com/datasets/bmadushanirodrigo/fracture-multi-region-x-ray-data
[18] https://www.kaggle.com/datasets/jasonroggy/grazpedwri-dx

- **Generality vs. Specificity**: Rich annotations (FracAtlas) may lack scale or regional diversity; large general datasets (MURA) may not support detection/localization.
- **Limited Size:** Custom and multi-modal datasets are often too small for robust deep learning training.
- **Heterogeneous Modalities:** Combining X-ray with CT/MRI (in custom datasets) introduces complexity in model development and alignment.

Summary Table: Key Dataset Comparison

Table 4. *Summary of Public Datasets for Bone Fracture Detection: Modalities, Annotations, and Dataset Characteristics.*

Dataset	Modality	Size & Split	Annotations	Strengths
MURA	Upper extremity X-ray	~14,863 studies → 40,561 images	Normal / Abnormal (binary)	Huge scale, radiologist-labeled
FracAtlas	X-rays (hand, leg, hip)	4,083 images; 717 fractured scans	Classification, boxes, masks (COCO, etc.)	Rich task diversity; localization/segmentation support
HBFMID (Custom Multi-modal)	X-ray, CT, MRI	641 images	Multi-class fracture labels	Multi-class capability, multi-modal
Large Multi-region X-ray	X-ray	~9,463 images	Class labels (fracture/no-fracture)	Broad anatomical coverage, transfer learning
GRAZPEDWRI-DX (Pediatric wrist)	X-ray	20,327 images	Multiple pediatric labels	Pediatricly diverse, annotation-rich

Performance Metrics and Benchmarks

Common Evaluation Metrics

To objectively assess deep learning models for bone fracture detection, multiple quantitative metrics are employed. These metrics capture various aspects of classification performance, from general accuracy to clinically relevant sensitivity and specificity.

- **Accuracy**:
  Proportion of correct predictions (fracture vs. non-fracture). While easy to interpret, it can be misleading in imbalanced datasets.
- **Sensitivity (Recall / True Positive Rate)**:
  Ability of the model to correctly detect fractures. High sensitivity is crucial in medical diagnostics to avoid missing actual cases.
- **Specificity (True Negative Rate)**:
  Measures the model's ability to identify normal (non-fractured) cases correctly. Essential to reduce false positives that can lead to overdiagnosis.
- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve)**:

Represents the trade-off between sensitivity and specificity across thresholds. A higher AUC-ROC indicates better overall discriminative ability.

- **F1-Score**:
  Harmonic mean of precision and recall. Especially useful for imbalanced data where both false positives and false negatives are costly.

Cross-validation techniques

To improve the generalizability of results and reduce variance across subsets, **cross-validation** is widely adopted:

- **k-Fold Cross-Validation**: The Dataset is split into *k* subsets. Each subset is used for validation once, while the remaining k-1 subsets are used for training. Typical values: *k = 5 or 10*.
- **Stratified Sampling**: Ensures that fracture/non-fracture proportions remain consistent across folds.
- **Leave-One-Hospital-Out (LOHO)**: Particularly useful in clinical datasets to validate generalization across institutions.

Clinical vs. model-level validation

While standard model-level metrics provide quantitative insights, **clinical validation** focuses on model performance in real-world workflows:

- **Model-Level Validation**:
  - ✓ Based on holdout test sets or public benchmark splits (e.g., MURA validation set).
  - ✓ Useful for A/B comparison of architectures and hyperparameter tuning.
- **Clinical Validation**:
  - ✓ Assesses model utility in aiding radiologists in triage, second opinions, or real-time reporting.
  - ✓ Involves **reader studies**, where human and AI diagnoses are compared.
  - ✓ May consider **time-to-diagnosis**, **confidence scores**, and **radiologist-AI agreement**.

Performance Benchmark Table

Below is a comparative summary of reported metrics from recent studies using public datasets like MURA, FracAtlas, and RSNA Bone Age:

Table 5. *Performance Metrics of Deep Learning Models for Bone Fracture Detection.*

Model / Study	Dataset	Accuracy (%)	AUC-ROC	Sensitivity (%)	Specificity (%)	Notes
ResNet-50 + Transfer Learning [18] [36]	MURA	85.3	0.929	87.1	83.6	Comparable to radiologists in upper-limb fractures

EfficientNet-B4 [37]	FracAtlas	90.1	0.947	91.5	88.8	Best results on segmentation-backed samples
ViT-B/16 + Grad-CAM [19] [37]	FracAtlas	87.5	0.934	89.2	85	Shows strong generalization on multi-view inputs
CNN + LSTM Hybrid [28] [34]	Custom X-ray	83.6	0.912	85	82	Useful for sequential/multi-view scenarios
MobileNetV2 (Edge Deployment) [23]	MURA	81.2	0.884	82.6	80.1	Fast inference (<50 ms), lightweight model
Ensemble (ResNet + ViT) [20] [37]	FracAtlas	92.3	0.956	94	90.4	High performance, but increased computational cost
FracNet (CNN + Transformer Fusion) [20] [33]	Multi-benchmark datasets	~98–100	>0.95	High (reported >95)	High (reported >94)	End-to-end framework; robust generalization across fracture types
RibFrac Challenge (Instance Segmentation) [29]	CT (RibFrac)	89.0	-	91.0	87.0	Specialized for rib fractures; focuses on lesion-level segmentation

Key Insights

- **Trade-offs exist** between sensitivity (necessary for diagnosis) and specificity (essential for avoiding over-treatment).
- **Transfer learning significantly improves model performance**, especially on smaller datasets.
- **Ensembles and hybrid models** yield better AUC and sensitivity but may be unsuitable for real-time inference without optimization.
- **Clinical reader studies** remain the gold standard for validating AI in practice and should complement benchmark metrics.

Limitations and Challenges

Despite significant progress in employing deep learning for bone fracture detection, several obstacles remain. These constraints stem from highly technical challenges, including data and model generalizability, as well as broader ones of clinical integration, explainability, and regulatory approval.

Data Scarcity and Annotation Bias

One of the most pressing challenges is the **limited availability of high-quality annotated datasets** [38]. Unlike natural image datasets, medical datasets:

- They are smaller due to **privacy concerns, institutional gatekeeping**, and **limited sharing infrastructure**.
- Often lack **fine-grained annotations** such as exact fracture boundaries or labels for fracture type (e.g., comminuted, spiral).
- Rely on **expert radiologists** for labeling, making the annotation process time-consuming and expensive.

Moreover, **annotation bias** can arise due to inter-rater variability, subjective interpretations of subtle fractures, or inconsistent labeling protocols across institutions. These biases can severely impact model generalization and diagnostic reliability.

Generalization Across Hospitals and Modalities

Deep learning models frequently suffer from **domain shift**, where a model trained on one dataset performs poorly on another due to:

- **Differences in imaging hardware** (e.g., brand/model of X-ray machine),
- **Variations in patient demographics**, and
- **Protocol discrepancies** across hospitals or countries.

For example, a model trained on adult wrist X-rays in a tertiary hospital may fail to generalize to pediatric cases or to radiographs from rural clinics [39]. Similarly, models trained on 2D X-rays may not adapt well to CT or MRI inputs without significant retraining.

Addressing generalization requires robust techniques such as:

- **Domain adaptation**
- **Data augmentation**
- **Federated learning across institutions**

However, these approaches are still in the early stages of clinical deployment.

Explainability and Black-Box Concerns

The **"black-box" nature** of many deep learning models remains a barrier to clinical trust and adoption. Radiologists and clinicians demand transparency in how and why a model makes a particular diagnosis [40]. Yet, many high-performing models do not provide a human-interpretable rationale for their decisions.

Although methods like **Grad-CAM**, **SHAP**, and **saliency maps** are promising, they:

- Often provide **noisy or misleading attention maps**,
- May not align with clinical reasoning,
- And lack standardized evaluation metrics.

Without robust and clinically meaningful explainability, deployment of AI tools for fracture detection risks skepticism, underuse, or misuse in clinical workflows.

Regulatory and Ethical Issues

Deploying deep learning models in clinical settings raises a host of **regulatory, ethical, and legal concerns** [41]:

- **Validation and certification**: AI models require regulatory approval (e.g., FDA, CE) before clinical use, necessitating rigorous multi-center validation.
- **Bias and fairness**: Models trained on data from specific regions or demographics may underperform in underrepresented populations, reinforcing healthcare disparities.
- **Accountability**: In the event of a misdiagnosis, it is unclear whether responsibility lies with the model developer, the hospital, or the radiologist.
- **Patient privacy**: Sharing medical images for AI research must comply with data protection laws (e.g., HIPAA, GDPR), often leading to delays or reduced dataset availability.

These factors underscore the need for ethical AI design, transparent validation pipelines, and close collaboration between developers, regulators, and clinicians.

Future Directions

The future of deep learning in fracture diagnosis involves dual advancements in algorithm development and execution. Federated learning techniques enable privacy-preserving collaboration among institutions, allowing AI models to generalize effectively without compromising patient privacy. Unsupervised, self-supervised, and contrastive representation learning techniques can utilize extensive unlabeled imaging data and reduce reliance on manual annotation, thereby mitigating low-data difficulties while achieving enhanced performance.

Clinical fracture detection models must be seamlessly integrated into PACS/RIS systems and provide immediate decision-making assistance without disrupting the radiologist's reading workflow. It is essential to acknowledge the shifts in country, device, and patient population domains and to modify these domains globally. Ultimately, the discipline must advance towards the prospective execution of clinical trials and regulatory validation to guarantee that these AI technologies are safe, interpretable, and effective across diverse clinical environments.

Conclusion

Deep learning has significantly influenced medical imaging and possesses considerable potential for rapid and precise fracture identification. Since the emergence of convolutional neural networks and the development of vision transformers, hybrid architectures have been proposed to analyze radiography

pictures, attaining performance levels comparable to or exceeding those of humans in controlled settings. Despite significant progress — particularly in terms of prediction accuracy, dataset availability, and interpretability — obstacles such as data scarcity, generalization across diverse populations, and integration into clinical practices remain impediments. Surmounting these obstacles through the development of innovative approaches, such as federated learning, self-supervised learning, and domain adaptability, will be crucial for unlocking the full potential of AI in fracture diagnosis. Ultimately, for these technologies to be effectively implemented in practice, it is essential to have not just technical innovation but also clinical validation, regulatory alignment, and strategic integration within the radiology ecosystem.

References

1. B. Kalyani, "Human Fracture Detection Using Machine Learning," *International Journal for Research in Applied Science and Engineering Technology,* 2024.
2. Salimi Ashkezari, Seyedeh Fatemeh, et al., "Differences Between Ruptured Aneurysms With and Without Blebs: Mechanistic Implications," *Cardiovascular Engineering and Technology,* vol. 14, no. 1, p. 92–103, 2022.
3. M. Kutbi, "Artificial intelligence-based applications for bone fracture detection using medical images: A systematic review," *Diagnostics,* vol. 14, p. 1879, 2024.
4. García, Bryam Esteban Coello, et al., "Panoramic review of open fractures, description, epidemiology, assessment, classification, treatment and complications," *EPRA international journal of multidisciplinary research,* vol. 9, no. 11, pp. 309-316, 2023.
5. Mittal, Khushi, et al., "Innovative Fracture Diagnosis: MobileNet CNN Approach for Precise Bone Fracture Detection and Classification," in *International Conference on Intelligent Systems for Cybersecurity (ISCS)*, Gurugram, India, 2024.
6. Shah, Sejal, Rohit M. Thanki, and Anjali Diwan, Artificial Intelligence for Early Detection and Diagnosis of Cervical Cancer, Springer, 2024.
7. Pisoudeh, Karim, Khatere Mokhtari, and Siamak Kazemi, "Recent Advanced in Imaging Technology for Diagnosing Hip Disorders: A Mini-review," *Journal of Research in Orthopedic Science,* vol. 10, no. 4, pp. 183--200, 2023.
8. Blankenbaker, Donna G and Davis, Kirkland W, Diagnostic Imaging: Musculoskeletal Trauma, E-Book: Diagnostic Imaging: Musculoskeletal Trauma, E-Book, Elsevier Health Sciences, 2021.
9. Spies, Amy J., et al., "Case discussions of missed traumatic fractures on computed tomography scans," *SA Journal of Radiology,* vol. 26, no. 1, pp. 1-7, 2022.
10. E. Wright, "Fracture management in the infant, child and young person," *Orthopaedic and Trauma Nursing: An Evidence-based Approach to Musculoskeletal Care,* pp. 349-361, 2023.
11. Ghoben, Marwa Kareem, and Lamia Abed Noor Muhammed, "Exploring the Impact of Image Quality on Convolutional Neural Networks: A Study on Noise, Blur, and Contrast," in *2023 International Conference of Computer Science and Information Technology (ICOSNIKOM)*, Binjia, Indonesia, 2023.
12. M. Kutbi, "Artificial intelligence-based applications for bone fracture detection using medical images: A systematic review," *Diagnostics,* vol. 14, no. 17, p. 1879, 2024.
13. A. Martin, Clark's Essential Guide to Preliminary Clinical Evaluation of Musculoskeletal X-rays, CRC Press, 2025.

14. Fang, Mengjie, et al., "Large models in medical imaging: Advances and prospects," *Chinese Medical Journal,* pp. 10--1097, 2025.
15. Aldhyani, Theyazn, et al., "Diagnosis and detection of bone fracture in radiographic images using deep learning approaches," *Frontiers in Medicine,* vol. 11, pp. 1-15, 2025.
16. Gunturu, Vijaya, et al., "Transfer Learning in Biomedical Image Classification," in *024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, Chennai, India, 2024.
17. Hossain, Md Imran, et al., "Explainable AI for medical data: Current methods, limitations, and future directions," *ACM New York, NY,* vol. 57, no. 6, pp. 1- 46, 2025.
18. Vempati, Gayathri, et al., "Evaluating CNN and Deep Learning Models for Bone Fracture Detection: A Comparative Study of VGG19, ResNet-50, LeNet and DenseNet," in *2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC)*, Davangere, India, 2024.
19. Tanzi, Leonardo, et al., "Vision transformer for femur fracture classification," *Injury,* vol. 53, no. 7, pp. 2625--2634, 2022.
20. Alwzwazy, Haider A., et al., "FracNet: An end-to-end deep learning framework for bone fracture detection," *Pattern Recognition Letters,* vol. 190, pp. 1-7, 2025.
21. Ruhi, Sayeda Sanzida Ferdous, et al., "A novel approach towards the classification of Bone Fracture from Musculoskeletal Radiography images using Attention Based Transfer Learning," in *2024 27th International Conference on Computer and Information Technology (ICCIT)*, 2024.
22. A. Englebert, A.-S. Collin, O. Cornu, and C. De Vleeschouwer, "Self-supervised vision-langage alignment of deep learning representations for bone X-rays analysis," *arXiv preprint arXiv:2405.08932,* 2024.
23. V. Varun, S. K. Natarajan, M. Adithya, P. Nagatejas, M. A., and N. M. Hosahalli, "Efficient CNN-Based Bone Fracture Detection in X-Ray Radiographs with MobileNetV2," in *2024 2nd International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS)*, Manipal, India, 2024.
24. J. Guo, Y. Mu, D. Xue, et al., "Automatic analysis system of calcaneus radiograph: Rotation-invariant landmark detection for calcaneal angle measurement, fracture identification and fracture region segmentation," *Computer Methods and Programs in Biomedicine,* vol. 206, p. 106124, 2021.
25. R. Hoover, N. Elsayed, Z. ElSayed, and C. Li, "Pre-trained Under Noise: A Framework for Robust Bone Fracture Detection in Medical Imaging," *arXiv preprint arXiv:2507.09731,* 2025.
26. S. Chauhan, "Bone Fracture Detection with CNN: A Deep Learning Approach," in *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*, IEEE, 2024.
27. K. Oude Nijhuis et al., "Open-source convolutional neural network to classify distal radial fractures according to the AO/OTA classification on plain radiographs," *Eur. J. Trauma Emerg. Surg.,* vol. 15, no. 1, p. 261, 2025.
28. U. Baid and S. Talbar, "MURA: bone fracture segmentation using a U-net deep learning in X-ray images," in *Techno-Societal 2020: Proceedings of the 3rd International Conference on Advanced Technologies for Societal Applications—Volume 1*, 2021.
29. Y. Yang et al., "Deep rib fracture instance segmentation and classification from ct on the ribfrac challenge," *IEEE Transactions on Medical Imaging,* 2025.
30. M. Kabir, T. J. Tahiti, and T. A. Prome, "A Comparative Study of Certain Convolutional Neural Network Architectures for X-ray Image Analysis in Bone Fracture Detection and Identification," in *2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, 2024.

31. F. Hardalaç, F. Uysal, and O. Peker, "Fracture detection in wrist X-ray images using deep learning-based object detection models," *Sensors,* vol. 22, no. 3, p. Sensors, 2022.
32. T. Meena and S. Roy, "Bone fracture detection using deep supervised learning from radiological images: A paradigm shift," *Diagnostics,* vol. 12, no. 10, p. 2420, 2022.
33. L. Yilmaz et al., "An automated hip fracture detection, classification system on pelvic radiographs and comparison with 35 clinicians," *Scientific Reports,* vol. 15, no. 1, p. 16001, 2025.
34. T. Kodavati and R. P. Kumar, "Detection and Segmentation of Skull Fractures via CNN and U-Net Hybrid Model using Computed Tomography Images," in *2023 Global Conference on Information Technologies and Communications (GCITC)*, 2023.
35. I. Sumon et al., "Automatic Fracture Detection Convolutional Neural Network with Multiple Attention Blocks Using Multi-Region X-Ray Data," *Life,* vol. 15, no. 7, p. 1135, 2025.
36. I. Kandel and M. Castelli, "Improving convolutional neural networks performance for image classification using test time augmentation: a case study using MURA dataset," *Health information science and systems,* vol. 9, no. 1, p. 33, 2021.
37. A. Murphy, V. Venkatesh, A. Sulam, and Y. Yi, "Visual transformers and convolutional neural networks for disease classification on radiographs: a comparison of performance, sample efficiency, and hidden stratification," *Radiology: Artificial Intelligence,* vol. 4, no. 6, p. e220012, 2022.
38. Napravnik, M., Hržić, F., Tschauner, S. *et al.,*"Building RadiologyNET: an unsupervised approach to annotating a large-scale multimodal medical database," *Biodata Mining,* vol. 17, no. 1, 2024.
39. Caliman Sturdza, Olga Adriana, et al., "Deep Learning Network Selection and Optimized Information Fusion for Enhanced COVID-19 Detection: A Literature Review," *Diagnostics,* vol. 15, no. 14, p. 1830, 2025.
40. Muhammad, Dost, and Malika Bendechache, "Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis," *Computational and structural biotechnology journal,* vol. 24, pp. 542-560, 2024.
41. Jha, Debesh, et al., "Ethical framework for responsible foundational models in medical imaging," *Frontiers in Medicine,* vol. 12, p. 1544501, 2025.