Towards Transparent AI in Medical Imaging: Fracture Detection in Hand Radiographs with Grad-CAM Insights

Mustafa Juzer Fatehi, Siddharath Malavalli Nagesh, Mandalam Akshit Rao, Stellin John George, J. Angel Arul Jothi[a] and Elakkiya Rajasekar[b]

Department of Computer Science, Birla Institute of Technology and Science, Pilani, Dubai Campus, Dubai International Academic City, Dubai, U.A.E.

Keywords: Deep Learning (DL), YOLO, Faster R-CNN, Explainable Artificial Intelligence (XAI), Grad-CAM, Fracture Detection.

Abstract: Timely and accurate detection of bone fractures in hand radiographs, particularly in fingers and wrists remains a critical challenge in clinical diagnostics due to anatomical complexity and subtle fracture patterns. This study presents an explainable AI framework for automatic fracture detection using a single-shot detection framework-YOLOv5 Medium (YOLOv5m) model, optimized through targeted preprocessing and interpretability techniques. A dedicated preprocessing pipeline is used to enhance fracture visibility and reduce irrelevant noise. This includes key steps like histogram equalization, Gaussian filtering, Laplacian filtering, and intensity normalization. To foster clinical trust and transparency, we integrate Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize regions of interest influencing the model's predictions. Trained on a curated dataset of over 9,000 annotated X-ray images, YOLOv5m achieved outstanding performance, with a mean Average Precision mAP@50 of 95.87% and an inference speed of 690 ms, making it suitable for real-time diagnostic support. This work demonstrates the potential of AI-assisted systems not only to improve fracture diagnosis but also to bridge the trust gap in clinical deployment through transparent decision-making support.

1 INTRODUCTION

Finger and wrist fractures represent a significant portion of the 178 million global bone fractures annually, posing diagnostic challenges due to their subtle and varied presentation in X-ray imaging Wu et al. (2021). Undetected or misdiagnosed fractures can lead to long-term complications, underscoring the need for accurate and timely diagnosis. While X-ray imaging remains the gold standard for fracture detection, manual interpretation is time-intensive and prone to variability among clinicians Valali et al. (2023). This work aims to address these limitations by leveraging artificial intelligence (AI) to improve fracture detection and localization, ultimately reducing diagnostic time, easing the burden on medical staff, and improving patient outcomes.

There are various difficulties in using AI to detect fractures. The scarcity of high-quality annotated data, which is necessary for building reliable models, is one of the main challenges. Furthermore, fractures can happen in a variety of places along the bone and have a wide range of orientations, which makes it challenging to create a universal detection model. Overlapping structures in X-ray images add to the complexity and can make fractures harder to see and make detection less accurate. Moreover, different patients present unique anatomical variations, adding to the difficulty of creating models that can generalize across diverse populations. Addressing these challenges is crucial for developing reliable AI models capable of improving diagnostic accuracy and consistency in clinical settings.

This study explores the use of deep learning (DL) techniques for detecting and localizing bone fractures in X-ray images, with a focus on finger joint and wrist fractures captured from diverse perspectives. By leveraging a dataset of over 9,000 images, significantly larger and more comprehensive than those used in prior studies, we address critical

[a] https://orcid.org/0000-0002-1773-8779
[b] https://orcid.org/0000-0002-2257-0640

Fatehi, M. J., Nagesh, S. M., Rao, M. A., George, S. J., Jothi, J. A. A. and Rajasekar, E.
Towards Transparent AI in Medical Imaging: Fracture Detection in Hand Radiographs with Grad-CAM Insights.
DOI: 10.5220/0013676200004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 39-48

limitations in existing research related to dataset diversity and size. We finetuned the YOLOv5 Medium (YOLOv5m) model. It is a single stage object detection model known for its balance of speed, accuracy, and lightweight architecture-on this dataset and extended the detection pipeline to incorporate explainability via Grad-CAM, modifying the model's final layers to enable attention-based visualization of predicted regions.

Our evaluation examines the model's performance across multiple fracture types within the same anatomical region, assessing both detection accuracy and generalizability. By offering a robust evaluation framework, our research contributes to the development of more reliable, effective and transparent automated diagnostic systems, paving the way for future advancements in fracture detection and medical imaging.

This paper is organized into seven sections. Section 1 provides an overview of the research problem and objectives. Section 2 situates the study within existing literature. The Section 3 outlines the data used, followed by the Methodology in Section 4 which details the proposed approach. System Requirements and Evaluation Metrics are specified in Section 5. The Section 6 discusses the findings, and the Conclusion section highlights key insights and future directions.

2 RELATED WORK

In recent years, the application of machine learning (ML) and DL techniques have significantly advanced the field of automated bone fracture detection (Ahmed and Hawezi, 2023). Zhang et al. (2021) proposed a traditional ML pipeline involving grayscale conversion, Gaussian filtering, adaptive histogram equalization, Canny edge detection, and Gray-Level Co-occurrence Matrix (GLCM)-based feature extraction, with classification using models like Support Vector Machine (SVM) achieving up to 92% accuracy. Addressing annotation ambiguity, a point-based annotation with "Window Loss," was introduced achieving an Area Under the Receiver Operating Characteristic curve (AUROC) of 0.983 and Free-Response Receiver Operating Characteristic (FROC) of 89.6%, outperforming standard detectors.

Building on traditional ML, several studies demonstrated the superior capability of DL models in capturing complex patterns. Karanam et al. (2021) emphasized the effectiveness of Convolutional Neural Networks (CNN) for hierarchical feature learning, especially in large datasets. Ghosh et al. (2024) further improved accuracy (97%) by applying anatomical feature enhancement before feeding the images into CNNs. Lee et al. (2020) proposed a meta-learning-based encoder-decoder using GoogLeNet, utilizing shared latent representation for improved classification across modalities.

Hybrid and transfer learning strategies have also shown significant promise. Khatik and Kadam (2022) and Warin et al. (2023) explored the use of pretrained models such as ResNet and Faster R-CNN, integrating transfer learning and data augmentation to enhance performance. Meena and Roy (2022) demonstrated the integration of real-time DL models like ResNet, VGGNet, and U-Net, achieving high accuracy for wrist and hip fractures while highlighting challenges such as class imbalance and rare case detection.

Fracture localization has become increasingly important. Ma (2021) proposed a two-stage framework combining Faster R-CNN with a Crack-Sensitive CNN (CrackNet) for detecting and classifying specific bone regions. Similar detection-refinement pipelines were explored by Abbas et al. (2020) and Su et al. (2023), reporting mAP scores around 60%.

One-shot detectors such as the YOLO family have gained substantial traction for their speed and efficiency. Zou and Arshad (2024) demonstrated YOLO's effectiveness over two-stage detectors. Ju and Cai (2023) showcased YOLOv8's performance, achieving a mAP of 0.638 using multi-scale feature fusion.Morita et al. (2024) confirmed YOLOv8's superiority over SSD after extended training. The YOLOv7-ATT model by Zou and Arshad (2024), with an attention mechanism, achieved 86.2% mAP on the FracAtlas dataset by focusing on subtle fracture-specific cues. Moon et al. (2022) used YOLOX-S for nasal bone fractures, achieving 100% sensitivity and 69.8% precision, thereby easing diagnostic burden for specialists.

Beyond YOLO, other architectures have been tested. AFFNet, as proposed by Nguyen et al. (2024), improved upon ResNet-50 while integrating activation maps to visualize important regions. While RetinaNet lagged behind with approximately 76% accuracy, Yadav et al. (2022) introduced SFNet—using multi-scale fusion and edge detection—to achieve 99.12% accuracy, 100% precision, and 98% recall, outperforming U-Net, YOLOv4, and R-CNN. In addition, Beyraghi et al. (2023) explored microwave imaging as a novel, radiation-free method for fracture detection using S-parameter data and deep neural networks, achieving high classification accuracy and low regression error.

Parallel to the advancements in detection architectures, the role of XAI has grown critical in ensuring

model transparency and trust. Borys et al. (2023) categorized saliency-based XAI methods such as Grad-CAM, LIME, SHAP, and Occlusion into perturbation-based and backpropagation-based techniques, outlining their respective strengths and limitations in medical image analysis. They also addressed the practical challenges in deploying XAI, emphasizing variability in heatmap-based visual attributions across methods. Their study called for a standardized, multi-dimensional evaluation framework to assess XAI reliability and alignment with clinical decision-making, reinforcing the synergy between accurate detection and explainable outputs. Volkov and Averkin (2023) further examined XAI's domain-specific applications, noting Grad-CAM's success in radiology, dermatology, and histopathology, particularly in detecting COVID-19 pneumonia, brain tumors, and skin lesions. They advocated for clinician-centered design and proposed integrating XAI with fuzzy logic to enhance diagnostic support in real-world clinical workflows.

3 DATASET DESCRIPTION

The Bone Fracture Detection Dataset (Phanan, 2024) used in this work contains about 9,585 X-ray images of finger and wrist fractures, encompassing diverse orientations and imaging conditions, including top-down and lateral views. The dataset is split into training (70%), validation (20%), and testing (10%) subsets.

Images of fractured bones are annotated using bounding boxes to highlight the fracture areas as observed in Figure 1. These annotations follow the YOLO format, which includes the class label and normalized coordinates of the bounding boxes (center, width, and height) relative to the image size. The images, provided as 640 $\times$ 640 pixel JPEGs, are readily compatible with standard computer vision frameworks and preprocessing pipelines. The dataset also includes images featuring multiple fractures within the same frame, enabling the detection of cases with more than one fracture simultaneously. Focused specifically on finger and wrist fractures, this dataset offers a rich collection of clinically relevant images that enable rigorous evaluation and comparison of automated fracture detection approaches.

4 METHODOLOGY

The methodology for bone fracture detection in this paper is structured into five key stages: data collection, preprocessing, model implementation, training, evaluation, and testing. These stages are designed to comprehensively address the challenges of accurate fracture detection, from processing the dataset to assessing the performance of fine-tuned models.

Figure 1: X ray images with ground truth boxes.

The core of this study centers on the implementation and optimization of the YOLOv5m model for accurate detection and localization of finger bone fractures. Leveraging pretrained Common Objects in Context (COCO) weights, the model was fine-tuned on a specialized dataset of annotated hand X-rays to adapt to the nuanced patterns of bone injuries. The YOLOv5m architecture was selected for its balance of speed, accuracy, and efficiency, making it ideal for real-time clinical integration. To contextualize its performance, a comparative evaluation with other detection frameworks including single-stage variants like YOLOv8 and YOLOv11, as well as the two-stage Faster R-CNN was conducted. These comparisons, while secondary, provided insights into the trade-offs between speed, precision, and model complexity. An overview of the end-to-end workflow, including data preparation, model training, and evaluation, is illustrated in Figure 2.

The following subsections provide a detailed discussion of each stage, outlining the specific techniques and strategies employed.

4.1 Data Preprocessing

The images were preprocessed to transform them into a format optimized for object detection models, making it easier to identify fractures accurately. The preprocessing pipeline focused on enhancing subtle features, such as minute fractures and overlapping struc-

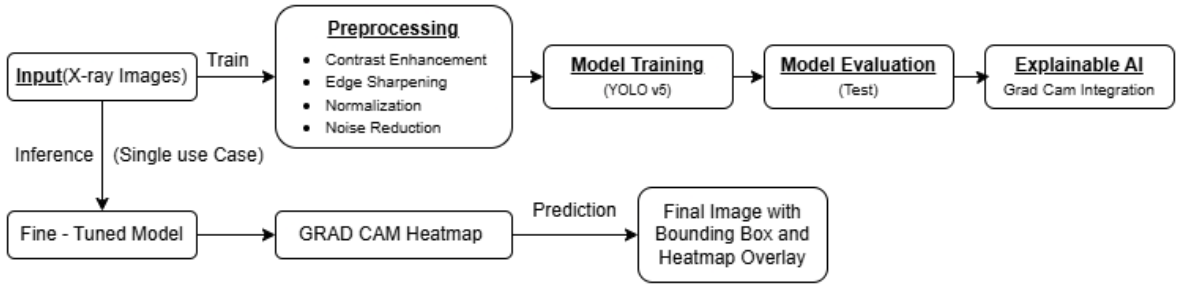

Figure 2: Flowchart Illustrating the Training Process and Single-Image Inference Outcomes

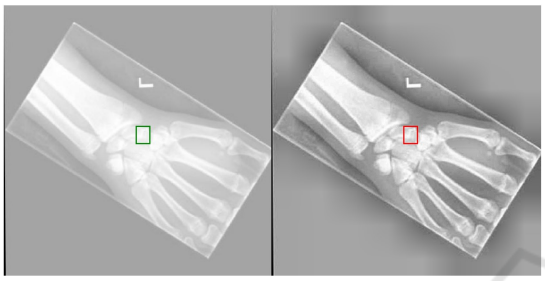

Figure 3: Comparison of Before (left) and After (right) preprocessing.

tures, which often pose challenges for accurate detection.

The preprocessing steps included Contrast Enhancement, Noise Reduction, Edge Sharpening, and Image Normalization. Contrast enhancement was achieved using histogram equalization, which improved the visibility of bone structures by increasing the dynamic range of pixel intensity values. This technique enhanced the contrast-to-noise ratio (CNR), enabling clearer differentiation between fracture lines and surrounding bone. Noise reduction was performed using a Gaussian filter to suppress high-frequency noise introduced by imaging equipment or environmental factors. This step preserved essential spatial details critical for identifying fine bone structures, such as trabecular patterns, while improving overall image clarity. To address inherent blurriness caused by the finite size of X-ray focal spots, edge sharpening was implemented using a Laplacian filter. This step enhanced bone boundaries and fracture lines, enabling the detection model to focus on critical features for accurate identification. Finally, image normalization was applied to standardize pixel intensity values to a range of [0, 1], ensuring consistent input to the detection models and reducing variability across images.

As observed in Figure 3, focusing on preprocessing techniques tailored to the requirements of bone fracture detection ensured that the input data was optimized for object detection, providing a solid foundation for training and evaluating advanced models.

4.2 Object Detection Model and Its Applicability

To effectively detect fractures in hand X-ray images, this study employs YOLOv5m. Building upon a robust preprocessing pipeline that enhances contrast, reduces noise, and sharpens critical edges, YOLOv5m [28] was selected due to its proven ability to perform well in tasks involving small and irregularly shaped objects: characteristics common in bone fractures.

Interestingly, while a range of models including other YOLO variants and two-stage detectors such as Faster R-CNN were briefly explored, YOLOv5m consistently outperformed them in both mean Average Precision (mAP) and inference speed. This unexpected lead in performance is likely attributed to its efficient use of anchor-based detection, optimized feature aggregation through the Spatial Pyramid Pooling (SPP) block, and its strong inductive bias toward learning small object patterns, which aligns well with the fracture detection task.

4.2.1 YOLOv5 Medium: Architecture and Suitability

YOLOv5m utilizes an anchor-based detection mechanism and a Spatial Pyramid Pooling (SPP) module, which allows the model to aggregate spatial information across multiple scales. This design is particularly effective in detecting fractures that vary in size, shape, and intensity. The model's backbone is optimized for extracting deep spatial features, while the head simultaneously predicts bounding boxes and class probabilities, enabling fast and reliable inference.

The anchor boxes were fine-tuned during training to adapt to the dimensions specific to fractures in hand radiographs. YOLOv5m's modular design and efficient convolutional layers allowed it to retain structural nuances in the X-ray images, improving its ability to generalize across varying fracture presentations.

Figure 4: Yolov5m prediction vs Ground Truth.

4.3 Training and Validation

YOLOv5m was trained for 100 epochs using pretrained COCO weights as initialization. This transfer learning approach allowed the model to converge faster while benefiting from prior knowledge of general object features. Training employed a batch size of 16 and a starting learning rate of 0.01 with momentum set at 0.937, adhering to YOLOv5's recommended defaults to ensure training stability and reproducibility.

To improve the model's robustness and generalization capability, several data augmentation techniques were employed. Horizontal flipping simulated anatomical variations in hand orientation, while RandAugment introduced random variations in brightness and contrast to mimic real-world X-ray acquisition inconsistencies. Mosaic augmentation, a technique where four images are combined into one was used during the initial 10 epochs to diversify object context and improve detection under cluttered scenarios. This augmentation was phased out in later epochs to allow for more focused learning on fracture-specific features.Validation was conducted after each epoch using a holdout validation set, evaluating mAP, classification accuracy, and localization precision. YOLOv5m demonstrated consistent performance gains with each augmentation step, converging steadily toward optimal detection capability. Its final performance reached a mAP@50 of 95.87% with an inference time of 690 ms per image, proving its suitability for real-time clinical applications. By tailoring the training process specifically for YOLOv5m and integrating domain-specific preprocessing and augmentation techniques, the model was optimized to detect even the most subtle and overlapping fractures with high confidence and speed.

4.4 Model Testing

The testing phase aimed to evaluate the generalizability and performance of the trained model on unseen data. During testing, the model generated predictions in the form of bounding boxes and confidence scores. These outputs were evaluated against their corresponding ground truth annotations using standard object detection and localization metrics, which are described in detail in Section 5.

Sample outputs were also manually visualized to verify the correctness of the predicted bounding boxes against the ground truth annotations. Figure 4 presents example outputs, demonstrating the overlap between predicted and actual fracture regions, providing a qualitative check of the model's performance.

4.5 Visualizing Model Decisions Using Grad-CAM

To enhance the interpretability of the YOLOv5-based fracture detection pipeline and foster clinician trust, Grad-CAM was integrated into the inference process of the model. Grad-CAM generates class-discriminative heatmaps that visually highlight the regions in an image most influential to a model's decision, offering intuitive insights into its reasoning.

Unlike explainability methods such as LIME, SHAP, or Integrated Gradients which are primarily designed for structured data or classification tasks, Grad-CAM is well-suited for spatial vision tasks like object detection. Its ability to localize important features makes it especially valuable in medical imaging, where understanding the spatial rationale behind predictions is crucial.

To apply Grad-CAM within the YOLOv5 architecture, the inference pipeline was modified to extract gradient information from the final convolutional block before the detection head. Gradients were computed with respect to the confidence scores of high-confidence bounding boxes post non-max suppression. The resulting heatmaps were normalized and superimposed on the original X-ray images, providing visual explanations for the predictions of the model.

This approach allowed to confirm that the model consistently focused on clinically relevant features such as cortical disruptions, fracture lines, or trabecular misalignments while avoiding irrelevant regions like overlapping soft tissues or background noise. These visualizations not only validated true positives

but also provided insight into false negatives, particularly in subtle or ambiguous cases.

By bridging the gap between AI predictions and clinical reasoning, Grad-CAM significantly enhanced the transparency of the system. It empowered clinicians to audit and cross-validate the model's decisions, promoting trust and supporting the safe integration of AI into real-world diagnostic workflows. Overall, this integration reaffirmed YOLOv5's robustness for high-stakes medical imaging and demonstrated Grad-CAM's potential as a valuable diagnostic companion tool.

Figure 5 presents the XAI results, showcasing the key image regions that influenced the model's diagnostic decisions.

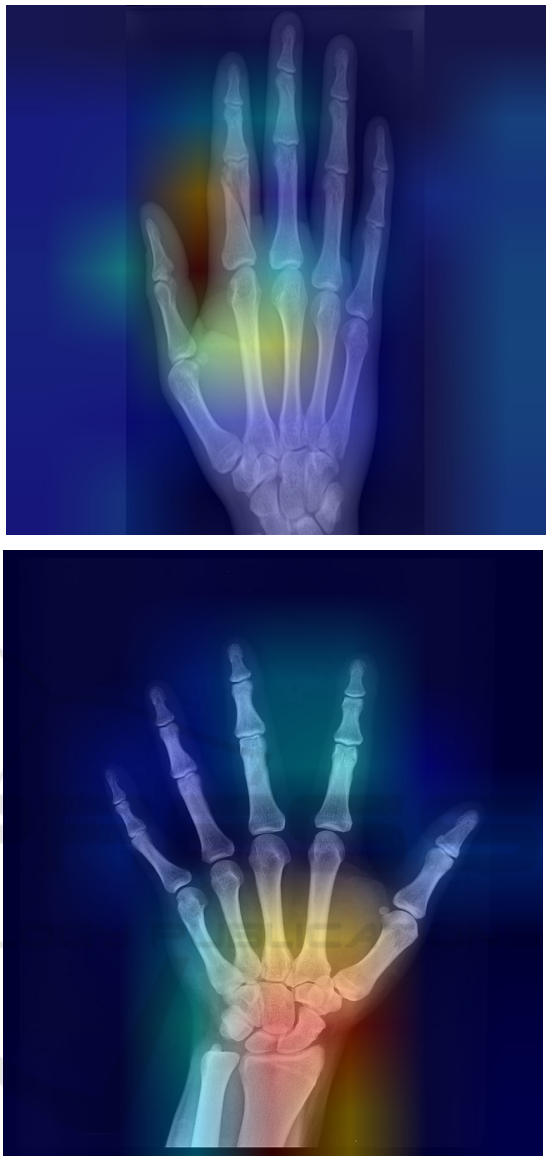

Figure 5: Grad-CAM highlights regions influencing predictions, with red areas showing key focus around the metacarpal (top) and fracture site near the thumb (bottom).

5 IMPLEMENTATION AND EVALUATION METRICS

The training of all models was performed over 100 epochs using an NVIDIA RTX A4500 GPU. The software environment comprised of a Windows operating system, with code development and execution carried out in Visual Studio Code. Key libraries, including PyTorch and Torchvision, were utilized for model implementation and dataset processing. CUDA was employed to accelerate computations on the GPU, while Matplotlib was used for visualizing results and generating performance graphs, facilitating clear and insightful analysis.

For inference evaluation, the trained models were tested on an Intel Core i5 processor. Inference time was used as an additional evaluation metric to assess the real-time applicability of the models. This setup provides a reliable and efficient framework for conducting experiments, by ensuring a stable training and evaluation environment.

The performance of all models was evaluated using standard object detection metrics, including mean Average Precision (*mAP*) and Intersection over Union (IoU). In bone fracture detection, $mAP@50$ measures how well the model identifies fractures when there is at least 50% IoU between the predicted and ground truth bounding boxes. This metric focuses on whether the model can reliably highlight fracture regions, even with some localization error. A high $mAP@50$ means the model is good at finding most fractures. It is given by (1)

$$mAP@50 = \frac{1}{C}\sum_{c=1}^{C} AP_c(50) \tag{1}$$

whHere, C is the number of object classes, c denote a specific object class, and $AP_c(50)$ represent the Average Precision for the class c at an IoU threshold of 50%.

$mAP@50{:}95$ provides a stricter and more comprehensive evaluation, as it considers detection performance across multiple IoU thresholds (from 50% to 95% overlap). This is calculated by (2)

$$mAP@50:95 = \frac{1}{T \cdot C}\sum_{t=1}^{T}\sum_{c=1}^{C} AP_c(t) \tag{2}$$

where T denote the number of IoU thresholds (e.g., [0.50, 0.55, ..., 0.95] in steps of 0.05), and $AP_c(t)$ represents the Average Precision for the class c at the IoU threshold t. A higher $mAP@50{:}95$ indicates that the

model not only identifies the fractures but also accurately localizes their boundaries with minimal error.

IoU measures the accuracy of the bounding box localization. It evaluates how much of the predicted bounding box overlaps with the actual fracture area and is given by (3). A higher IoU score means the model is capturing the fracture's shape and size more accurately. This is vital for ensuring that subtle or small fractures are not missed or mislocalized. The IoU accuracy in this work has been computed at 50% intersection minimum threshold to maintain object detection quality while increasing tolerance for minor localization errors.

$$\text{IoU} = \frac{Area \quad of \quad Overlap}{Area \quad of \quad Union} \tag{3}$$

The number of parameters across the models were analyzed to assess the computational complexity and efficiency. This evaluation provides insights into the trade-offs between model size and performance. By comparing parameter counts, it is easy to determine which models offer a balance between accuracy and resource requirements.

6 RESULTS AND DISCUSSION

The model testing results as seen in Table 1 provide important insights into the performance of various object detection models for bone fracture detection. Among all the models, YOLOv5m emerged as the best performer, achieving the highest *mAP*50 (95.87%), *mAP*50:95 (61.70%), and IOU50 (78.12%) while maintaining a reasonable inference speed of 690 ms. This demonstrates that YOLOv5m effectively balances accuracy, computational efficiency, and generalization, making it the most suitable model for this application. Its performance suggests that its architecture with spatial pyramid pooling strategy is well-aligned with the dataset's characteristics. As a result of multi-scale feature representation, it enables the model to capture both fine-grained details and broader contextual information, resulting in precise detection and localization of fractures without excessive computational overhead.

A clear trend also observed across the medium-sized models (YOLOv5m, YOLOv8 Medium (YOLOv8m), and YOLOv11 Medium (YOLOv11m)) is their consistent superiority in accuracy compared to both their smaller and larger counterparts. The performance trend across epochs is visualized in Figure 6, which shows how *mAP*@50 evolves during training. Figure 7 further illustrates the trend by comparing model performance (*mAP*@50:95) at different epochs. The results reinforces that YOLOv5m

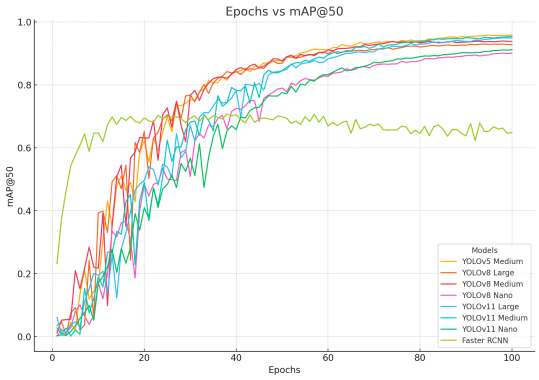

Figure 6: Epochs vs *mAP*@50 for all models.

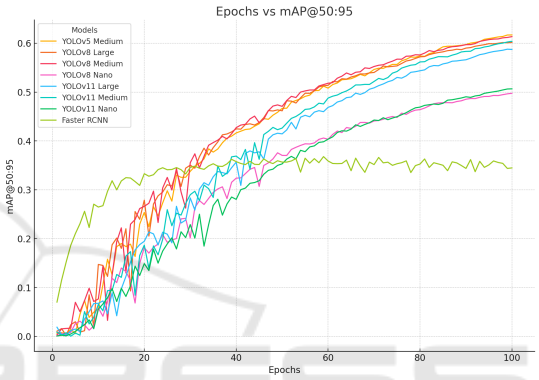

Figure 7: Epochs vs *mAP*@50:95 for all models.

not only converges faster, but also achieves the best trade-off between accuracy and speed. Medium-sized models have enough parameters to capture complex features in the dataset without the risk of overfitting, which is often observed in larger models. On the other hand, smaller models, such as YOLOv8 Nano (YOLOv8n) and YOLOv11 Nano (YOLOv11n), excel in inference speed (166 ms and 180 ms, respectively) but compromise significantly on accuracy, particularly at higher IoU thresholds. This makes nano models well-suited for applications reliant on central server processing, where speed is prioritized, with a modest tradeoff in precision and accuracy.

Interestingly, the larger models, including YOLOv8 Large and YOLOv11 Large, underperform compared to the medium models. Their lower *mAP*@50:95 values indicate that these models may suffer from overfitting due to their higher parameter counts (43.7M and 25.34M, respectively). Larger models typically require more extensive training data and longer training times to generalize effectively, which might not have been sufficiently addressed in this study. Moreover, their higher inference times (1,451 ms and 998 ms) further reduce their practicality for time-sensitive applications.

Table 1: Model Performance Comparison.

Model	*mAP*@50 (%)	*mAP*@50:95 (%)	IOU50 (%)	Inference Speed (ms)	Num Parameters
YOLOv5 Medium	95.87	61.70	78.12	690 ms	21.2M
YOLOv8 Nano	90.07	49.75	67.87	166 ms	3.2M
YOLOv8 Medium	93.37	61.35	77.06	712 ms	25.9M
YOLOv8 Large	92.76	60.16	76.51	1,451 ms	43.7M
YOLOv11 Nano	91.18	50.66	68.59	180 ms	2.62M
YOLOv11 Medium	95.40	60.39	76.51	768 ms	20.09M
YOLOv11 Large	94.91	58.73	75.54	998 ms	25.34M
Faster R-CNN	70.32	36.17	17.51	3,151 ms	41.2M

Faster R-CNN, despite being a well-known two-stage object detection model, performs poorly across all metrics, with an *mAP*@50 of 70.32% and a particularly low IOU50 of 17.51%. Its inference time of 3,151 ms is significantly slower than all YOLO models, highlighting its computational inefficiency for this task. The architecture of Faster R-CNN likely struggles to adapt to the dataset's requirements, as it relies on generating region proposals in the first stage, which can be less effective for subtle or small features like bone fractures. Additionally, its large parameter count (41.2M) increases the risk of overfitting, especially if the training data is not diverse or large enough. The relationship between *mAP* values and the number of epochs is illustrated in Figures 6 and 7, offering further insights into the learning patterns and supporting similar conclusions.

In summary, the results emphasize the importance of selecting a model that aligns with the specific requirements of the application. While YOLOv5m proves to be the most effective for bone fracture detection due to its balance of accuracy and speed, smaller models like YOLOv8n offer exceptional speed at the cost of precision, and larger models require more extensive optimization to perform well. Faster R-CNN, meanwhile, demonstrates significant limitations for this specific task, underlining the need for efficient, single-stage architectures like YOLO when dealing with datasets of this nature.

Furthermore, the incorporation of XAI through Grad-CAM significantly enhanced the transparency of our fracture detection pipeline by visually highlighting regions that influenced model predictions. Figure 8 presents the combined output of YOLOv5m predictions and Grad-CAM visualizations, clearly demonstrating that the highlighted regions align well with the annotated fracture areas, thereby validating the model's interpretability and attention to clinically relevant features. Grad-CAM was applied to the final convolutional layers of the YOLO-based models, enabling the identification of class-discriminative regions that overlapped meaningfully with predicted bounding boxes. This helped not only in detecting fractures but also in localizing them accurately by highlighting potential cracks or fractures. To quantitatively assess the alignment between the YOLO-generated bounding boxes and the Grad-CAM heatmaps, we employed the Pointing Game Accuracy metric. This metric evaluates whether the maximum activation point from the Grad-CAM heatmap falls within the ground truth bounding box; a successful hit indicates agreement between the model's attention and the annotated region. Our results showed a high pointing game accuracy above 85 percent, demonstrating that the regions the model focused on for decision-making were clinically relevant. As illustrated in Figure 9, the overlap between the Grad-CAM heatmaps and the predicted bounding boxes is clearly visible and aligns well with the actual fracture sites. We further verified this behavior across multiple images, consistently observing similar localization quality, which confirms the robustness and reliability of our XAI integration. This synergy between detection and interpretability not only validates the model's performance but also reinforces its suitability for real-

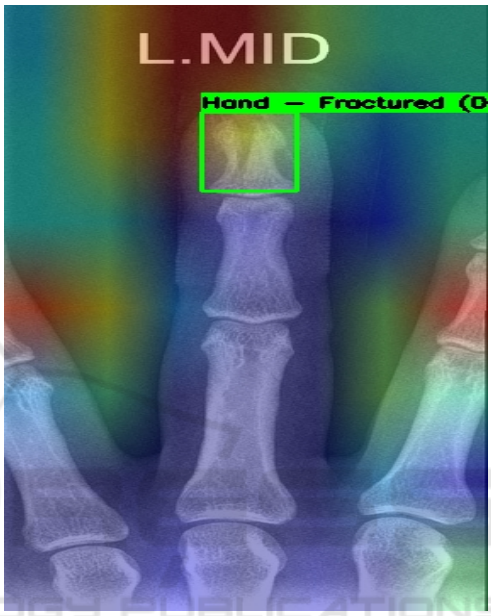

Figure 8: Combined output of YOLOv5m and Grad-CAM visualizations.

Figure 9: Grad-CAM visualization doesn't align with the micro fractures.

world clinical deployment.

In some instances, the Grad-CAM heatmaps did not align with the YOLO bounding boxes. Upon investigation, two primary factors were identified. First, many of the fractures were microfractures, which the model was unable to detect. Second, shadows in the X-rays or overlapping bone structures obscured the fracture sites, preventing the model from performing effectively. This finding calls for future work to develop enhanced imaging preprocessing techniques and more robust model architectures that can reliably detect microfractures and compensate for shadows or overlapping anatomical features.

7 CONCLUSION

This study evaluates object detection models for detecting finger fractures in X-ray images. By preprocessing images to enhance anatomical details and training several state-of-the-art models, we assessed their performance in fracture detection and localization.

Single-stage detectors, especially those using spatial pyramid pooling for feature aggregation, consistently outperformed two-stage models. Notably, YOLOv5 surpassed newer models like YOLOv8 and YOLOv11, indicating that its architecture may be better suited for the specific features present in finger fractures. Lightweight models like YOLO-Nano also performed well despite having fewer parameters, suggesting that smaller models can be effective when applied to narrowly defined tasks.

In contrast, Faster R-CNN, typically strong in general object detection, underperformed in this task. Its lower generalizability in this context reinforces the need for architecture-specific tuning when dealing with medical imagery. Our YOLOv5 based approach achieved a detection accuracy of 95.8%, slightly higher than the 95.1% reported by RoboFlow's fracture model on comparable datasets. These results are significantly higher than the sub 70% accuracy commonly observed in broader fracture datasets, underscoring the benefits of task-specific optimization.

To enhance interpretability, we integrated Grad-CAM with the YOLOv5 model, achieving a pointing game accuracy of over 85%. The resulting heatmaps reliably highlighted fracture regions, providing visual insight into model decisions and improving transparency, an essential factor in medical AI applications.

This work demonstrates that carefully tuned object detection models, particularly single-stage detectors with spatial pooling mechanisms, can effectively handle specialized medical tasks. It also highlights the role of lightweight models and explainability tools in building clinically relevant AI systems.

For future work, we propose transitioning from detection to semantic segmentation of fractures. This would allow pixel-level mapping of fracture morphology, offering more detailed characterization, which is critical for surgical planning and outcome prediction. Integration of these models into clinical decision support systems could further streamline workflows and enhance diagnostic precision.

REFERENCES

Abbas, W. et al. (2020). Lower leg bone fracture detection and classification using faster rcnn for x-rays images. *IEEE Xplore*.

Ahmed and Hawezi (2023). Detection of bone fracture based on machine learning techniques. *Measurement Sensors*, 27:100723.

Beyraghi, S. et al. (2023). Microwave bone fracture diagnosis using deep neural network. *Scientific Reports*, 13(1).

Borys, K. et al. (2023). Explainable ai in medical imaging: An overview for clinical practitioners – saliency-based xai approaches. *European Journal of Radiology*, 162:110787.

Ghosh, S. et al. (2024). Automated bone fracture detection in x-ray imaging to improve orthopaedic diagnostics in healthcare. *Procedia Computer Science*, 233:832–840.

Ju, R.-Y. and Cai, W. (2023). Fracture detection in pediatric wrist trauma x-ray images using yolov8 algorithm. *Scientific Reports*, 13(1):20077.

Karanam, S. R. et al. (2021). A systematic review on approach and analysis of bone fracture classification. *Materials Today: Proceedings*.

Khatik, N. I. and Kadam, N. S. (2022). A systematic review of bone fracture detection models using convolutional neural network approach. *Journal of Pharmaceutical Negative Results*, pages 153–158.

Lee, C. et al. (2020). Classification of femur fracture in pelvic x-ray images using meta-learned deep neural network. *Scientific Reports*, 10(1).

Ma, Y. (2021). Bone fracture detection through the two-stage system of crack-sensitive convolutional neural network. *Informatics in Medicine Unlocked*, 22:100452.

Meena, T. and Roy, S. (2022). Bone fracture detection using deep supervised learning from radiological images: A paradigm shift. *Diagnostics*, 12(10):2420.

Moon, G. et al. (2022). Computer aided facial bone fracture diagnosis (ca-fbfd) system based on object detection model. *IEEE Access*, 10:79061–79070.

Morita, D. et al. (2024). Automatic detection of midfacial fractures in facial bone ct images using deep learning-based object detection models. *Journal of Stomatology, Oral and Maxillofacial Surgery/Journal of Stomatology Oral & Maxillofacial Surgery*, pages 101914–101914.

Nguyen, H. H. et al. (2024). Affnet - a deep convolutional neural network for the detection of atypical femur fractures from anteriorposterior radiographs. *Bone*, 187:117215.

Phanan (2024). bonefracturedetection_v1 dataset. Open Source Dataset, Roboflow Universe. Retrieved from https://universe.roboflow.com/phanan/bonefracture detection_v1.

Su, Z. et al. (2023). Skeletal fracture detection with deep learning: A comprehensive review. *Diagnostics*, 13(20):3245.

Valali et al. (2023). Bone fracture detection and classification using deep learning models on x-ray images. *Diagnostics*, 15(3):271.

Volkov, E. N. and Averkin, A. N. (2023). Explainable artificial intelligence in medical image analysis: State of the art and prospects. In *Proc. 2023 XXVI Int. Conf. on Soft Computing and Measurements (SCM), Dubna, Russia*, pages 133–137.

Warin, K., Limprasert, W., Suebnukarn, S., Paipongna, T., Jantana, P., and Vicharueang, S. (2023). Maxillofacial fracture detection and classification in computed tomography images using convolutional neural network-based models. *Scientific Reports*, 13(1).

Wu, A. et al. (2021). Global, regional, and national burden of bone fractures in 204 countries and territories, 1990–2019: a systematic analysis from the global burden of disease study 2019. *The Lancet Healthy Longevity*, 2(9):e580–e592.

Yadav, D. P., Sharma, A., Athithan, S., Bhola, A., Sharma, B., and Dhaou, I. B. (2022). Hybrid sfnet model for bone fracture detection and classification using ml/dl. *Sensors*, 22(15):5823.

Zhang, X. et al. (2021). Window loss for bone fracture detection and localization in x-ray images with point-based annotation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 724–732.

Zou, J. and Arshad, M. R. (2024). Detection of whole body bone fractures based on improved yolov7. *Biomedical Signal Processing and Control*, 91:105995–105995.