Pediatric Wrist Fracture Detection in Radiographs Using the YOLOv11n Object Detection Model

Conduct by:

Name: : Ahmed Mohammed Moahmmed Nasser Alghaili

Student ID: : 22523231

PROGRAM STUDI INFORMATIKA – PROGRAM SARJANA

FAKULTAS TEKNOLOGI INDUSTRI

UNIVERSITAS ISLAM INDONESIA

2026

SUPERVISOR ENDORSEMENT PAGE

Pediatric Wrist Fracture Detection in Radiographs Using the YOLOv11n Object Detection Model

THESIS

Conduct By:

Name : Ahmed Mohammed Mohammed Nasser Alghaili

Student ID : 22523231

Yogyakarta, January 09st, 2026

Advisor,

(Izzati Muhimmah, S.T., M.Sc., Ph.D.)

Pediatric Wrist Fracture Detection in Radiographs Using the YOLOv11n Object Detection Model

THESIS

Has been defended in front of the examiners as one of the requirements to obtain a Bachelor of Informatics degree from the Undergraduate Program in Informatics at the Faculty of Industrial Technology, Universitas Islam Indonesia

Yogyakarta, January 09st, 2026

Chair

Izzati Muhimmah, S.T., M.Sc., Ph.D. _____

Examiner 1

Irving Vitra Paputungan, S.T., M.Sc. Ph.D. _____

Examiner 2

Ahmad Fathan Hidayatullah, S.T., M.Cs. Ph.D. _____

Acknowledged by,

Head of Undergraduate Program in Informatics

Faculty of Industrial Technology

Universitas Islam Indonesia

(Ir. Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D.)

AUTHENTICITY STATEMENT

The undersigned:

Name	:	Ahmed Mohammed Mohammed Nasser Alghaili
Student ID	:	22523231

Final project with title:

Pediatric Wrist Fracture Detection in Radiographs Using the YOLOv11n Object Detection Model

Stating that all components and contents in this final project are my own work. If in the future it is proven that some parts of this work are not my own work, the final project submitted as my own work is ready to be withdrawn and ready to bear any risks and consequences.

Thus this statement letter is made, hopefully it can be used properly.

Yogyakarta, January 09st, 2026

(Ahmed Mohammed Mohammed Nasser Alghaili)

DEDICATION

This thesis is dedicated to my parents, my family, and everyone who supported me throughout my academic journey. Their encouragement and prayers have been my strength.

MOTTO

Success is not final; failure is not fatal. It is the courage to continue that counts.

FOREWORD

The foreword is the section used to express gratitude for the completion of the final project report. It may also include:

1. The purpose of writing the report or conducting the research.
2. Non-academic difficulties encountered during the research.
3. Acknowledgements to those who helped or supported the completion of the report or research.
4. The author's expectations regarding the completed research.

The foreword is usually closed with the author's signature.

Yogyakarta, January 09st, 2026

(Ahmed Mohammed Mohammed Nasser Alghaili)

ABSTRACT

Fracture detection in pediatric wrist radiographs is challenging due to incomplete skeletal ossification, small bone structures, and subtle hairline (non-displaced) fractures that can be difficult to identify, while growth-plate (physeal) radiolucency often mimics fracture appearance. This study evaluates YOLOv11n, a lightweight one-stage object detector that incorporates multi-scale feature extraction components (e.g., Spatial Pyramid Pooling–Fast, SPPF), for automated pediatric wrist fracture detection and localization. The model was trained and evaluated on the GRAZPEDWRI-DX benchmark dataset comprising 20,327 pediatric wrist radiographs (14,269 training, 4,048 validation, 2,010 test images) using transfer learning with the Ultralytics training pipeline and default online augmentation strategies. YOLOv11n achieved mAP@50 = 0.936 on the validation set and mAP@50 = 0.940 on the test set, with precision = 0.923 and recall = 0.850 on validation and precision = 0.926 and recall = 0.870 on test. Runtime profiling on an NVIDIA Tesla T4 GPU indicated end-to-end per-image latency below 5 ms, supporting near-real-time clinical decision-support workflows. The mAP@50–95 values (0.564 on validation and 0.552 on test) indicate reduced localization tightness under stricter IoU criteria, consistent with the greater difficulty of precisely localizing subtle fracture regions. Overall, YOLOv11n provides a favorable balance between detection performance and computational efficiency for pediatric wrist fracture detection; however, external multi-institutional validation and targeted strategies (e.g., multi-view fusion and pediatric anatomy-aware modeling) are recommended before clinical deployment to improve sensitivity to subtle fractures and reduce growth-plate-related false positives.

Keywords: Computer Vision, Deep Learning, Fracture Detection, Medical Imaging, Pediatric X-rays

LIST OF TABEL

CHAPTER I
INTRODUCTION

1.1 Background

1.1.1 Global burden of pediatric wrist fractures

Fractures are one of the most frequent injuries in childhood, accounting for a significant portion of pediatric emergency department (ED) visits. Population-based studies suggest that 40–50% of boys and 30–40% of girls experience at least one fracture before reaching adulthood, with the distal forearm and wrist region being the most commonly affected areas, and a substantial proportion of these injuries involve the distal forearm and wrist region (Korup et al., 2022; Südow & Mellstrand Navarro, 2021). Distal forearm and distal radius fractures are consistently reported as the most common fractures in children, with incidence rates in the range of roughly 300–700 per 100,000 person-years, depending on age, sex, and geographic region (Korup et al., 2022; Mamoowala et al., 2019).

Fracture incidence typically peaks in late childhood and early adolescence, particularly among boys, and is strongly associated with falls during play and organized sports activities (Korup et al., 2022; Südow & Mellstrand Navarro, 2021). Given demographic trends and high participation in physical activity, pediatric wrist fractures impose a considerable clinical and economic burden on healthcare systems, requiring radiographic imaging, accurate diagnosis, treatment, and follow-up at scale. These observations highlight the big picture: pediatric wrist fractures are common, recurrent, and clinically significant injuries, making efficient and reliable diagnostic workflows a public health priority.

1.1.2 Diagnostic challenges and missed fractures

Despite standardized radiographic protocols, interpreting pediatric skeletal radiographs presents unique challenges. Children's bones differ significantly from adults due to the presence of open physes (growth plates), incomplete ossification, and smaller bone size, all of which can obscure or mimic fractures in X-ray images (Shelmerdine et al., 2024; Smith et al., 2016). As a result, missed or delayed diagnoses of fractures remain a persistent problem.

Classical work by Wei et al. (2006) reported that approximately 3–4% of extremity fractures were initially missed in emergency radiology, with many errors attributed to perceptual or interpretative mistakes rather than truly invisible lesions. More recent analyses from teaching hospitals similarly document radiologic discrepancies in fracture diagnosis and

emphasize that missed fractures are a leading source of diagnostic error and potential medicolegal risk (Mattijssen-Horstink et al., 2020).

Pediatric-specific investigations show comparable concerns. Smith et al. (2016) reported missed fractures on pediatric ED radiographs, underscoring that even in specialized settings a measurable proportion of fractures remain undetected at first reading. Reviews of pediatric skeletal imaging likewise note both false-negative (missed) and false-positive (overcalled) interpretations, with clinically relevant consequences in many cases (Shelmerdine et al., 2024; Su et al., 2023).

In children, the implications of missed wrist fractures are particularly important. Inadequately managed distal radius fractures can lead to malunion, growth disturbance, and functional limitation, and repeated hospital visits add further burden for families and health systems (Korup et al., 2022; Südow & Mellstrand Navarro, 2021). Consequently, there is a clear main issue: clinicians require better support in reliably detecting pediatric wrist fractures on radiographs, especially in busy ED environments with high workload and time pressure.

1.1.3 Deep learning for fracture detection: advances and limitations

In parallel with these clinical challenges, the last decade has seen rapid progress in deep learning (DL) for medical image analysis. Multiple reviews report that convolutional neural networks (CNNs) achieve high diagnostic accuracy for fracture detection, often approaching or even matching the performance of human readers in experimental settings (Cohen & McInnes, 2022; Kalmet et al., 2020; Su et al., 2023). Kalmet et al. (2020), for example, concluded that deep learning is a reliable tool for fracture diagnosis across several anatomical regions, while Cohen and McInnes (2022) highlighted promising results but also noted methodological limitations in many studies.

More recent evidence focuses directly on wrist fractures and pediatric populations. In a systematic review and meta-analysis, Hansen et al. (2024) found that DL models for wrist fracture detection can perform on par with or better than healthcare experts, achieving high sensitivity and specificity in identifying fractures on radiographs. Franco et al. (2024) evaluated a commercial AI algorithm on both adult and pediatric patients and reported sensitivities around 85–91% and accuracies above 84% for appendicular fractures in children. Ziegner et al. (2025) further demonstrated that implementing an AI-based fracture detection tool in a pediatric ED improved the diagnostic performance of less experienced physicians in real-world practice.

Even so, several limitations and open questions remain. Reviews emphasize that many DL fracture detection models:

- are trained on single-center datasets with limited external validation;
- are developed primarily on adult or mixed-age cohorts; and
- may not fully address the anatomical and developmental characteristics of pediatric bones (Cohen & McInnes, 2022; Su et al., 2023).

Pediatric-specific studies, such as the protocol by Shelmerdine et al. (2024) and the clinical evaluation by Ziegner et al. (2025), underscore the need to better characterize AI performance in children and to understand how these systems integrate into everyday workflows.

1.1.1 YOLO-based detectors and the emergence of YOLOv11

In the field of computer vision, object detection frameworks like the YOLO (You Only Look Once) family are highly effective for fracture detection. YOLO-based systems can simultaneously localize and classify fractures within a single forward pass, enabling near real-time inference, which is particularly valuable for clinical workflows that demand speed without sacrificing accuracy. Several recent works have applied YOLO variants to bone fracture detection in plain radiographs. YOLOv7-based systems, for instance, have been trained on pediatric wrist trauma datasets like GRAZPEDWRI-DX to localize and classify fractures with high precision (Kazi et al., 2025; Zbinden et al., 2021). Other studies have used YOLOv7 or YOLOv10 for general bone fracture detection, showing competitive accuracy and efficient deployment (Khan et al., 2025; Ndoh et al., 2025).

More recently, **YOLOv11 architectures** have been proposed to improve small-object detection and efficiency. Wei et al. (2025) introduced a YOLOv11-based multi-task learning framework for enhanced fracture detection and classification in X-ray images, reporting performance gains compared with earlier detectors. Tariq et al. (2025) further explored YOLOv11-driven pipelines for musculoskeletal fracture detection, emphasizing improved localization and quantification capabilities.

However, existing YOLO-based fracture detection research is still dominated by adult or mixed-age cohorts, and only a subset explicitly target pediatric wrist radiographs. Pediatric work using YOLOv7 on GRAZPEDWRI-DX demonstrates feasibility but leaves open questions about the benefits of newer architectures and lightweight variants, especially in the context of pediatric wrist fracture detection (Hansen et al., 2024; Kazi et al., 2025; Su et al., 2023).

Given these developments, this thesis investigates the application of a lightweight YOLOv11n-based object detection model for pediatric wrist fracture detection. A detailed review of existing approaches and the specific research gaps addressed are presented in Chapter 2.

1.2 Problem Statement

Based on the background the core problems addressed in this thesis can be stated as follows:

1. Diagnostic difficulty despite high prevalence Pediatric wrist fractures are highly prevalent and clinically important, yet radiographic interpretation in children is challenging, contributing to missed and delayed diagnoses in emergency and radiology settings.
2. Insufficient pediatric-specific deep learning solutions Existing DL fracture detection models are often trained on adult or mixed-age cohorts and do not explicitly account for pediatric wrist anatomy and developmental features, leading to uncertainty about their suitability for this domain.
3. Unclear benefit of lightweight YOLOv11n models Although YOLOv11 variants are designed for efficient small-object detection, there is limited empirical evidence on how a lightweight YOLOv11n model performs for pediatric wrist fracture detection and whether it offers advantages over earlier YOLO versions or alternative architectures.
4. Lack of comprehensive evaluation and error analysis Many studies emphasize global accuracy but do not provide detailed object-detection metrics and systematic error analysis in pediatric populations, making it difficult to assess clinical risk and integration potential.

1.3 Research Questions

To address the above problems, this thesis is guided by the following research questions (RQs):

- RQ1: How accurately can a YOLOv11n-based model detect fractures in pediatric wrist radiographs when evaluated using standard object-detection metrics such as mAP@50, mAP@50–95, precision, recall, and F1-score?
- RQ2: How does the performance of YOLOv11n compare with selected baseline models or earlier YOLO variants (if implemented) on the same pediatric wrist fracture dataset?
- RQ3: Which anatomical regions and imaging characteristics in pediatric wrist radiographs (e.g., proximity to growth plates, subtle or non-displaced fracture appearance, low contrast, motion blur, and overlapping anatomy) are most frequently associated with false-negative and false-positive detections by the YOLOv11n model?
- RQ4: To what extent is a lightweight YOLOv11n model—optimized for small-object detection and efficient inference—feasible for potential integration into pediatric radiology workflows in terms of diagnostic performance and computational requirements?

1.4 Scope of the Study

To ensure a focused and feasible investigation, the scope of this thesis is defined as follows:

1. Imaging modality and anatomical region: The study is restricted to plain radiographic (X-ray) images of the pediatric wrist region, primarily involving fractures of the distal radius and adjacent structures. Other anatomical regions (e.g., ankle, elbow) and imaging modalities (CT, MRI, ultrasound) are not considered.
2. Population: Only pediatric patients within a defined age range (as specified in the main study) are included. Adult radiographs are excluded from model training and testing.
3. Task definition: The primary task is object detection of fractures (localization and classification) in wrist radiographs. The thesis does not include fracture severity grading, treatment recommendation, prognosis, or long-term outcome prediction.
4. Modeling approach: The core focus is on YOLOv11n as the main detection architecture. Other model families (e.g., transformers, RF-DETR) may be discussed in the literature review but are not implemented unless explicitly stated in later chapters.
5. Clinical deployment aspects: The study is conducted in a research environment. Regulatory, medico-legal, user-interface design, and prospective clinical trial aspects lie outside the formal scope of this thesis, though they are acknowledged in the discussion.

1.5 Research Objectives

The overall goal of this thesis is to develop and evaluate a YOLOv11n-based deep learning approach for detecting fractures in pediatric wrist radiographs. This goal is translated into the following specific objectives:

1. To design and implement a YOLOv11n-based object detection pipeline for pediatric wrist fracture detection, including data preprocessing, augmentation strategies, and model configuration tailored to small bone structures.
2. To train and quantitatively evaluate the YOLOv11n model on a curated pediatric wrist X-ray dataset using standard object-detection metrics (mAP@50, mAP@50–95, precision, recall, F1-score) and clinically relevant measures such as fracture-level sensitivity.
3. To compare (if baselines are available) YOLOv11n against selected baseline models or earlier YOLO variants, assessing potential improvements in detection accuracy, robustness, and computational efficiency.
4. To compare (if baselines are available) YOLOv11n against selected baseline models or earlier YOLO variants, assessing potential improvements in detection accuracy, robustness, and computational efficiency.

5. To conduct a structured error analysis of YOLOv11n predictions, identifying recurrent patterns of false negatives and false positives across fracture types, anatomical locations, and image qualities, and interpreting these findings in relation to pediatric clinical practice.

1.6 Research Contributions

This thesis is expected to contribute to the literature and practice in the following ways:

1. Pediatric-oriented detection framework Development of a YOLOv11n-based fracture detection pipeline specifically adapted to pediatric wrist radiographs, taking into account growth plate visibility and small bone structures.
2. Empirical evidence for YOLOv11n in pediatric fracture detection: Provision of a systematic quantitative evaluation of YOLOv11n for pediatric wrist fracture detection, including detailed object-detection metrics and error analyses, thereby extending current evidence that is dominated by adult or mixed-age cohorts.
3. Insight into error patterns and feasibility: A structured characterization of model failure modes and a discussion of computational requirements, offering practical insight into the feasibility of deploying lightweight YOLOv11n models in pediatric radiology departments.
4. Reproducible methodology for future work: Documentation of dataset preparation, model training, and evaluation protocols, serving as a reproducible foundation for subsequent work, including multi-center validation and integration into clinical decision-support systems.

1.7 General Research Methodology

To achieve the research objectives, this thesis follows a general research methodology commonly used in applied medical imaging and deep learning studies. Although the implementation in this work employs the YOLOv11n model, the methodology is presented in a general form and can be adapted to other detection architectures.

1. Problem Formulation and Literature Review

The research begins by defining the clinical problem of pediatric wrist fractures and the role of radiographic imaging in diagnosis. A structured literature review is conducted covering pediatric fracture characteristics, missed fracture rates, and deep learning–based fracture detection approaches, with particular attention to YOLO-based methods.

2. Data Acquisition and Preprocessing

An annotated dataset of pediatric wrist X-ray images is obtained or utilized in compliance with ethical and privacy regulations. Preprocessing steps such as intensity normalization, image resizing, and optional contrast enhancement techniques (e.g., CLAHE) are applied to improve bone structure visibility. The dataset is then split into training, validation, and test subsets, with bounding-box annotations provided for fracture regions.

3. Model Design and Implementation

The YOLOv11n architecture is configured for pediatric wrist fracture detection by selecting an appropriate input resolution, training hyperparameters, and data augmentation strategies. A complete training pipeline is implemented using a suitable deep learning framework to ensure efficient experimentation and reproducibility.

4. Model Training and Validation

The model is trained on the training dataset while validation metrics are continuously monitored to guide hyperparameter tuning and reduce overfitting. Techniques such as data augmentation, early stopping, and mixed-precision training (where appropriate) are employed to improve generalization performance and computational efficiency.

5. Testing and Performance Evaluation

The final trained model is evaluated on a held-out test dataset using standard object detection metrics, including mAP@50, mAP@50–95, precision, recall, and F1-score. Clinically relevant measures, such as fracture-level sensitivity, are also considered. When baseline or comparison models are available, comparative experiments and, where feasible, statistical analyses are conducted to assess performance differences.

6. Error Analysis and Interpretation

False-positive and false-negative detections are analyzed and grouped based on fracture type, anatomical location, and imaging characteristics. These error patterns are interpreted in the context of pediatric anatomy, radiographic challenges, and existing literature on fracture misdiagnosis and AI-assisted diagnostic systems.

7. Conclusion and Recommendations

The study concludes by summarizing the main findings in relation to the research questions and objectives. Limitations, such as dataset size, single-center data collection, and the absence of external validation, are discussed. Recommendations for future work include multi-center studies, multimodal data integration, and the development of user-friendly clinical deployment interfaces.

Although this methodology is implemented using YOLOv11n in the present work, it is sufficiently general to be extended to alternative object detection architectures and hybrid systems in future research.

CHAPTER II
LITERATURE REVIEW

2.1 Background Research Objectives

Theoretical foundation. This study applies deep learning–based object detection to localize fractures in pediatric wrist radiographs. The core technical concepts include convolutional neural networks (CNNs) for feature extraction, one-stage object detection (YOLO) for joint localization and classification, and IoU- and mAP-based metrics for evaluation. These concepts provide the basis for interpreting the related studies reviewed in this chapter.

2.1.1 Pediatric wrist anatomy and fracture epidemiology

The pediatric wrist comprises the distal radius and ulna, carpal bones, and multiple physes that remain open throughout skeletal growth. Compared with adults, children's bones are characterized by thicker periosteum, greater elasticity, and incomplete ossification, which result in fracture patterns such as buckle (torus), greenstick, and various Salter–Harris physeal injuries that are rarely encountered in adults (Liao & Chong, 2019). These age-specific characteristics influence both the radiographic appearance and clinical management of wrist trauma.

Epidemiological investigations consistently identify distal forearm and distal radius fractures as the most frequent fractures in childhood. Korup et al. (2022) reported more than 4,300 distal forearm fractures over a 7-year period in a Danish pediatric population, with incidence rates in the order of several hundred cases per 100,000 person-years. Similarly, an eight-year review from a UK trauma unit found distal radius fractures to be the single most common pediatric fracture subtype, with incidence peaking in late childhood and early adolescence (Mamoowala et al., 2019). These injuries are predominantly associated with low-

energy falls onto an outstretched hand during everyday activities and organized sports (Liao & Chong, 2019).

Although many distal radius fractures in children are stable and amenable to conservative management, delayed or inadequate diagnosis may lead to malunion, residual deformity, growth disturbance, and functional limitation (Liao & Chong, 2019; Luhmann et al., 2004). Given their high incidence and potential long-term consequences, pediatric distal radius and wrist fractures represent a clinically important and resource-intensive problem, which motivates efforts to improve diagnostic accuracy and efficiency.

2.1.2 Radiographic imaging and diagnostic error

Plain radiography is the first-line imaging modality for suspected wrist trauma in children because it is widely available, relatively inexpensive, and associated with a modest radiation dose. Standard protocols typically include posteroanterior and lateral projections and, where necessary, oblique views to improve visualization of cortical continuity and joint congruity (Liao & Chong, 2019; Nagy et al., 2022).

Interpretation of pediatric wrist radiographs, however, is non-trivial. The presence of open growth plates, secondary ossification centers, and overlapping small bones can mimic or obscure fracture lines, particularly in subtle, nondisplaced, or buckle fractures. Such anatomical and developmental factors increase the cognitive load on readers and may contribute to diagnostic error.

In a classic analysis of emergency radiology practice, Wei et al. (2006) reported that a non-negligible proportion of extremity fractures were missed on initial interpretation, with a substantial share of errors attributable to perceptual or interpretation failures rather than to invisibility of the fractures. A subsequent retrospective study from a Dutch teaching hospital confirmed that radiologic discrepancies in fracture diagnosis remain a significant source of error and potential medicolegal exposure (Mattijssen-Horstink et al., 2020). Pediatric-focused work similarly documents clinically relevant rates of missed fractures on radiographs, including fractures that necessitated changes in management once identified (Smith et al., 2016).

Taken together, these findings indicate that even in contemporary practice fracture detection on pediatric wrist radiographs is challenging, and that a subset of cases may be overlooked. This context provides a clear rationale for investigating computer-assisted methods to support human readers.

The overall goal of this thesis is to develop and evaluate a YOLOv11n-based deep learning approach for detecting fractures in pediatric wrist radiographs. This goal is translated into the following specific objectives:

1. To design and implement a YOLOv11n-based object detection pipeline for pediatric wrist fracture detection, including data preprocessing, augmentation strategies, and model configuration tailored to small bone structures.
2. To train and quantitatively evaluate the YOLOv11n model on a curated pediatric wrist X-ray dataset using standard object-detection metrics (mAP@50, mAP@50–95, precision, recall, F1-score) and clinically relevant measures such as fracture-level sensitivity.
3. To compare (if baselines are available) YOLOv11n against selected baseline models or earlier YOLO variants, assessing potential improvements in detection accuracy, robustness, and computational efficiency.
4. To conduct a structured error analysis of YOLOv11n predictions, identifying recurrent patterns of false negatives and false positives across fracture types, anatomical locations, and image qualities, and interpreting these findings in relation to pediatric clinical practice.

2.2 Deep Learning and Object Detection in Medical Imaging

2.2.1 Convolutional neural networks in radiology

Deep convolutional neural networks (CNNs) have transformed image analysis by enabling hierarchical feature learning directly from raw image data (Krizhevsky et al., 2012). Subsequent architectures such as VGG, ResNet, DenseNet, and EfficientNet further improved representational capacity, regularization, and optimization, and have been widely adopted for medical imaging tasks.

In radiology, CNNs have been applied to lesion detection, disease classification, and organ segmentation across modalities including X-ray, computed tomography (CT), and magnetic resonance imaging (MRI). Narrative and systematic reviews focused on fracture detection show that CNN-based models can achieve high diagnostic performance and, under controlled experimental conditions, often approach or match that of human readers (Kalmet et al., 2020; Su et al., 2023). Nevertheless, these reviews also emphasize that many studies are retrospective, single-center, and methodologically heterogeneous, with limited external validation and variable reference standards.

Early deep learning approaches to fracture detection predominantly treated the task as image-level classification (fracture present vs. absent). While such models can reach high sensitivity, they do not explicitly localize fracture sites and thus provide limited spatial

interpretability. This limitation has motivated increased interest in object detection architectures that jointly perform localization and classification.

2.2.2 Two-stage and one-stage object detectors

Object detection algorithms are generally categorized into two-stage and one-stage paradigms. Two-stage detectors, exemplified by Faster R-CNN, first generate class-agnostic region proposals and then refine and classify them in a second stage (Ren et al., 2015). This design achieves high detection accuracy at the cost of relatively high computational complexity and inference time.

One-stage detectors, in contrast, such as the YOLO (You Only Look Once) family, directly regress bounding boxes and class probabilities from the full image in a single forward pass (Redmon et al., 2016). Subsequent YOLO versions (YOLOv2–YOLOv7) introduced architectural and training refinements, including anchor boxes, feature pyramid networks, improved backbones, and advanced data augmentation, to enhance accuracy while preserving real-time performance (Bochkovskiy et al., 2020; Wang et al., 2022).

In the context of medical imaging, one-stage detectors offer two key advantages:

1. Computational efficiency, which facilitates near real-time integration into clinical workflows; and
2. Joint localization and classification, which yields explicit spatial information about the predicted pathology and can be more easily interpreted by clinicians than pure image-level predictions.

2.2.3 Evaluation metrics for detection models

Object detection performance is typically evaluated using intersection-over-union (IoU)–based metrics. A predicted bounding box is considered a true positive if its IoU with the corresponding ground-truth box exceeds a predefined threshold. Average precision (AP) summarizes the area under the precision–recall curve at a given IoU threshold, and mean average precision (mAP) averages AP over all classes.

Two metrics are particularly relevant in the object detection literature:

- mAP@50, which computes AP at IoU = 0.5; and
- mAP@50–95, which averages AP across multiple IoU thresholds (0.50–0.95 in steps of 0.05), providing a more stringent assessment of localization quality (Lin et al., 2014).

In medical image analysis, these metrics are often complemented by clinically oriented measures such as exam-level sensitivity and specificity (e.g., “any fracture detected in the image”), which more directly reflect the clinical impact of model outputs (Hansen et al., 2024). A rigorous evaluation of fracture detection models should therefore report both detection-specific metrics and clinically interpretable indicators.

2.3 YOLO-Based Architectures for Fracture Detection

2.3.1 Non-YOLO deep learning approaches

Early deep learning approaches to fracture detection primarily formulated the task as image-level classification or segmentation applied to manually defined or automatically cropped regions of interest. Convolutional neural network (CNN) classifiers were widely used to distinguish fractured from non-fractured radiographs, demonstrating that deep models can learn discriminative fracture-related features from X-ray images (Kalmet et al., 2020; Cohen & McInnes, 2022). In parallel, segmentation-based architectures, particularly U-Net and its variants, were applied to delineate fracture regions or cortical disruptions, achieving promising results in controlled experimental settings (Ronneberger et al., 2015; Olczak et al., 2017).

Despite their effectiveness, these early methods exhibit several practical limitations. Many require multiple preprocessing steps such as manual region cropping, bone segmentation, or heuristic filtering, followed by post-processing to refine predictions, which increases system complexity and reduces robustness in real-world settings (Kalmet et al., 2020; Su et al., 2023). Furthermore, classification-only models lack explicit spatial localization of fracture sites, limiting interpretability and clinical usability, while segmentation-based pipelines often incur high computational cost and slower inference times, hindering real-time deployment in emergency or high-throughput radiology workflows (Cohen & McInnes, 2022). These limitations have motivated a shift toward unified object detection architectures, particularly one-stage detectors such as the YOLO family, which perform fracture localization and classification simultaneously within a single end-to-end framework, enabling efficient inference and improved clinical integration.

2.3.2 YOLOv7–YOLOv10 for bone fractures

With the maturation of one-stage detectors, the YOLO family has been increasingly studied for musculoskeletal radiographs:

- YOLOv7 in mixed skeletal regions. Several implementations have demonstrated that YOLOv7 can detect fractures in various bones on X-ray images with high sensitivity and low latency, illustrating its suitability for real-time clinical support, particularly in adult or mixed-age cohorts.
- Pediatric wrist with YOLOv8–YOLOv10. The publication of the GRAZPEDWRI-DX dataset (Nagy et al., 2022), comprising more than 20,000 pediatric wrist trauma radiographs with expert annotations, has provided a benchmark for pediatric wrist AI research. Using this dataset, Ju and Cai (2023) showed that YOLOv8 can attain strong mAP values for pediatric wrist fracture detection, confirming the ability of one-stage detectors to handle small bony structures. Chien and colleagues extended this line of work by evaluating YOLOv9 and attention-augmented YOLOv8 variants, reporting incremental improvements in detection performance. Ahmed and Manaf (2024) systematically explored YOLOv10 variants, demonstrating that careful hyperparameter tuning and data augmentation can improve accuracy while preserving or even enhancing computational efficiency.

Overall, YOLOv7–YOLOv10 represent a strong baseline for fracture detection, and existing evidence suggests that they are well-suited for pediatric wrist imaging when trained on appropriate datasets. However, the literature is still evolving and largely centered on these earlier YOLO versions.

2.3.3 Emergence of YOLOv11 and lightweight variants

YOLOv11 is a recent iteration in the YOLO series, designed to further improve accuracy, especially for small objects, while maintaining or reducing computational cost. Several research groups have begun to investigate YOLOv11 for fracture detection.

Wei et al. (2025) proposed a YOLOv11-based multi-task framework that simultaneously detects fractures and classifies fracture types on X-ray images, reporting improvements over previous detectors. Zhang et al. (2025) introduced an optimized YOLOv11n model that incorporates specialized modules and loss functions to enhance localization stability and robustness, achieving superior performance relative to a baseline YOLOv11n on general fracture detection tasks. Tariq and Choi (2025) evaluated YOLOv11 in the context of wrist fractures and observed improved accuracy and efficiency compared with earlier YOLO versions, alongside more interpretable visualization outputs.

Although these studies indicate that YOLOv11 and its lightweight variants are promising for fracture detection, they predominantly focus on adult or mixed-age cohorts or on generalized skeletal regions. Evidence specifically addressing YOLOv11n in pediatric wrist radiographs remains scarce. This represents a critical gap that the present thesis aims to address.

2.4 Deep Learning for Pediatric Wrist Fracture Detection

2.4.1 Pediatric-specific datasets and models

The GRAZPEDWRI-DX dataset (Nagy et al., 2022) constitutes a pivotal resource for pediatric wrist AI research. It includes de-identified wrist trauma radiographs collected from pediatric patients, with detailed expert annotations for fractures and other findings. This dataset has enabled systematic benchmarking of deep learning models and has been used in several high-impact studies.

Using GRAZPEDWRI-DX, Ju and Cai (2023) applied YOLOv8 and reported strong detection performance, thereby validating the applicability of modern one-stage detectors in pediatric wrist imaging. Subsequent studies by Chien and colleagues evaluated YOLOv9 and attention-enhanced YOLOv8 variants, demonstrating further improvements in detection accuracy. Ahmed and Manaf (2024) investigated YOLOv10 on the same dataset, emphasizing the importance of hyperparameter configuration and data augmentation strategies.

2.4.2 Summary of key pediatric wrist fracture AI studies

Table 2.1 summarizes selected deep learning studies that focus explicitly on pediatric wrist fracture detection. The table highlights the datasets used, main model families, target tasks, and high-level findings. Detailed metric values (e.g., exact mAP, sensitivity, specificity) should be filled according to the original publications and, where relevant, your own experimental results.

Tabel 2.1 Representative deep learning studies on pediatric wrist fracture detection

Study / Year	**Dataset**	**Model / Architecture**	**Task**	**Key Findings (high level)**
Nagy et al., 2022	GRAZPEDWRI-DX	Baseline CNN / classical ML	Dataset description, baseline models	Introduced large pediatric wrist X-ray dataset; provided baseline ML and CNN performance.
Ju & Cai, 2023	GRAZPEDWRI-DX	YOLOv8	Object detection (fracture vs. normal)	Demonstrated strong YOLOv8 performance on pediatric

				wrists; validated suitability of one-stage detectors.
Chien & co-authors, 2024	GRAZPEDWRI-DX	YOLOv9 / attention YOLOv8	Object detection	Reported improved detection by using updated backbones and attention mechanisms.
Ahmed & Manaf, 2024	GRAZPEDWRI-DX	YOLOv10 variants	Object detection	Showed that hyperparameter tuning and augmentation improve accuracy–speed trade-off.
Zech et al., 2023	Single-center cohort	DL-based object detector	Object detection, clinical evaluation	Evaluated deep learning detector in pediatric wrist setting; showed performance comparable to radiologists.
Ziegner et al., 2025	Pediatric ED cohort	Commercial AI / DL detector	Clinical support system	Demonstrated improved diagnostic performance of less experienced physicians with AI assistance.

2.4.3 AI performance relative to human readers

Several investigations have compared AI performance with human readers in pediatric or mixed-age wrist fracture detection. Hansen et al. (2024) conducted a systematic review and meta-analysis and concluded that, on average, deep learning algorithms achieve diagnostic performance comparable to that of healthcare professionals, with some models exceeding the performance of non-specialist readers. Husarek et al. (2024) reported similar conclusions for orthopedic fractures more broadly, while stressing methodological variability and frequent absence of external validation.

In clinical environments, Zech et al. (2023) showed that a DL-based object detector for pediatric wrist fractures attained performance in the range of radiology residents. Ziegner et al. (2025) demonstrated that deploying an AI-based fracture detection system in a pediatric emergency department improved sensitivity of less experienced physicians. Ramadanov et al. (2025) further reported that AI-assisted detection of distal radius fractures could match or exceed certain human raters.

These studies collectively indicate that AI-based fracture detection systems have the potential to function as effective decision-support or triage tools in pediatric wrist imaging, provided that their limitations and error patterns are well understood

2.4.4 Research Gap

Despite substantial progress in deep learning–based fracture detection, the literature reveals several unresolved gaps in the context of pediatric wrist radiographs.

First, although YOLO-family object detectors have demonstrated strong performance in pediatric wrist fracture detection, existing studies predominantly evaluate YOLOv8–YOLOv10 architectures. Evidence regarding the performance of the most recent lightweight YOLOv11n model on pediatric wrist datasets remains limited.

Second, many studies emphasize aggregated performance metrics, such as accuracy or mAP@50, but provide limited structured analysis of false-negative and false-positive detections. From a clinical perspective, understanding which pediatric fracture patterns are most frequently missed or misclassified is essential for safe adoption.

Third, most published studies rely on single-center datasets and lack external validation, leaving uncertainty regarding model robustness and generalizability across institutions and imaging conditions.

Finally, while YOLO-based models are often described as computationally efficient, few studies report detailed inference times, hardware requirements, or practical considerations for workflow integration in pediatric radiology settings.

Together, these gaps indicate a need for systematic evaluation of lightweight YOLOv11n-based fracture detection models in pediatric wrist radiographs, accompanied by detailed performance analysis and clinically oriented interpretation.

2.5 Conceptual Positioning of the Present Study

The literature reviewed above can be summarized along three interrelated dimensions:

1. Clinical dimension: Pediatric wrist fractures are common and can have significant long-term consequences. Radiographic interpretation is difficult, and diagnostic errors are documented. This creates a strong clinical motivation for robust computer-assisted detection systems.
2. Methodological dimension: Deep learning and YOLO-family object detectors have achieved strong results in fracture detection, including in pediatric wrist datasets. However, existing pediatric work is dominated by YOLOv8–YOLOv10, and evidence on the latest lightweight YOLOv11n architecture remains limited.
3. Evaluation dimension: Many studies emphasize headline metrics but provide limited error analysis, external validation, and discussion of workflow integration. A more

granular and clinically oriented evaluation is required to understand model strengths, weaknesses, and potential risks.

Within this landscape, the present thesis is positioned to:

- design and implement a YOLOv11n-based object detection pipeline for pediatric wrist fracture detection;
- evaluate its performance using standard detection metrics and clinically relevant indicators;
- conduct a structured error analysis focusing on pediatric-specific fracture patterns and imaging characteristics.

The next chapter will detail the research methodology, including dataset characteristics, preprocessing techniques, model configuration, training strategy, and evaluation protocol.

CHAPTER III
METHODOLOG
Y

3.1 Introduction

This chapter describes the materials and methods used to develop and evaluate a YOLOv11n-based deep learning pipeline for automated detection of pediatric wrist fractures from radiographs. The section details dataset provenance and ethical compliance, image preprocessing and augmentation strategies, the YOLOv11n network architecture, the end-to-end methodological workflow, training protocol and hyperparameter choices, evaluation metrics, the computational environment, and a brief summary. Wherever applicable, established best practices for object detection in medical imaging were followed and any dataset-specific decisions were justified with respect to preserving subtle fracture morphology and ensuring reproducibility.

3.2 Dataset and Ethical Compliance

3.2.1 Dataset Description

The dataset used in this study is the GRAZPEDWRI-DX dataset, a publicly available dataset designed for pediatric wrist trauma research. It consists of 20,327 pediatric wrist radiographs, including both fractured and non-fractured wrist images. The dataset is available for academic research purposes through the GRAZPEDWRI-DX repository (Nagy et al., 2022).

The images were collected from multiple hospitals, ensuring diversity in radiographic quality, patient positioning, and anatomical variations. Both anteroposterior (AP) and lateral views are included. All images were manually annotated by trained radiologists using bounding boxes to localize fracture regions, providing the supervision required for object detection training. Although the original dataset contains annotations for multiple fracture-related classes, this study focuses on binary fracture detection. Therefore, during dataset preparation, only annotations corresponding to the original fracture class 1 were retained, and all retained annotations were remapped to a single class 0 in YOLO format. All other annotation classes were removed. This preprocessing step converts the task into single-class fracture detection.

After preprocessing, the dataset was organized in Ultralytics YOLO format and split into three non-overlapping subsets at the image level:

- Training set: 14,269 images (70.2%)
- Validation set: 4,048 images (19.9%)
- Test set: 2,010 images (9.9%)

The training set was used to learn model parameters. The validation set was used for model selection (monitoring mAP/precision/recall during training), early stopping, and confidence-threshold selection. The test set was held out until the end and used only for final reporting to provide an unbiased estimate of generalization performance. This 70/20/10 split provides sufficient data for training while preserving independent validation and test partitions.

3.2.2 Class Distribution and Challenges

A common challenge in medical imaging datasets, particularly those involving fracture detection, is class imbalance. In the GRAZPEDWRI-DX dataset, there are more non-fractured cases than fractured cases. This imbalance can potentially lead to a model that favors predicting the majority class. To mitigate potential imbalance effects, data augmentation was applied during training, and performance was monitored using precision–recall–based metrics.

Additionally, the dataset contains subtle fractures, such as non-displaced or hairline fractures, which can be difficult to detect even by human experts. These types of fractures represent a challenge in evaluating the efficacy of deep learning models, making the dataset ideal for assessing fracture detection capabilities.

3.2.3 Ethical Considerations

The dataset is anonymized to ensure patient privacy and complies with ethical standards. It is used solely for academic research, and no direct contact with patients was involved. The

study adheres to the Declaration of Helsinki and complies with data protection laws, ensuring ethical handling of medical data (Nagy et al., 2022).

3.3 Preprocessing and Data Augmentation

3.3.1 Image Preprocessing

Before training, images and annotations were prepared to ensure compatibility with the YOLOv11n detection framework. Image resizing was handled automatically by the Ultralytics YOLO pipeline during training.

All input images were resized to 640 × 640 pixels using letterbox padding, which preserves the original aspect ratio while avoiding geometric distortion of anatomical structures. Pixel intensity normalization was performed internally by the YOLO framework during data loading and training.

Bounding box annotations provided by radiologists were converted to the YOLO label format *(class_id, x_center, y_center, width, height)* with all coordinates normalized to the range [0,1]. As part of dataset preparation, only fracture annotations were retained and remapped to a single class (class ID = 0), while all other annotation classes were removed.

No additional handcrafted image preprocessing steps, such as contrast enhancement or noise filtering, were applied. This design choice ensures that the model learns directly from the original radiographic appearance while relying on the robustness of the YOLOv11n architecture.

3.3.2 Data Augmentation

Data augmentation was applied automatically during training using the built-in augmentation strategies provided by the Ultralytics YOLO framework. These augmentations are performed dynamically at each training iteration, ensuring continuous variability in the training data without permanently altering the original images.

The default YOLO augmentation pipeline includes geometric and photometric transformations such as random scaling, translation, flipping, and mosaic-based composition. These transformations help improve model generalization by exposing the network to variations in object scale, position, and appearance commonly encountered in clinical radiographs.

No custom augmentation parameters were manually specified in this study; all augmentation operations followed the default configuration of the YOLO training framework. This approach maintains reproducibility and reduces the risk of introducing unrealistic anatomical distortions in pediatric wrist radiographs.

3.4 YOLOv11n Architecture

YOLOv11n is a lightweight variant of the YOLO family designed to achieve a balance between detection accuracy and computational efficiency. The nano ("n") configuration has a reduced parameter count and lower computational complexity, making it suitable not only for real-time medical imaging applications but also for deployment in resource-limited environments, such as hospitals or clinics with constrained computational infrastructure, limited GPU availability, or reliance on standard clinical workstations, without compromising detection performance.

3.4.1 Backbone

The backbone is responsible for extracting hierarchical visual features from the input radiographs. It consists of lightweight convolutional layers designed to balance representational power and computational efficiency.

The architecture incorporates Spatial Pyramid Pooling–Fast (SPPF) modules to aggregate multi-scale contextual information while maintaining low computational overhead. Through hierarchical feature extraction, the backbone learns essential anatomical characteristics such as cortical bone continuity, epiphyseal structure, and subtle fracture lines that are critical for pediatric wrist fracture detection.

3.4.2 Neck

The neck aggregates features from multiple scales produced by the backbone and fuses them using a feature pyramid–based structure. This multi-scale fusion enables the model to detect fractures of varying sizes by combining fine-grained spatial details with higher-level semantic information.

3.4.3 Head

The detection head outputs bounding boxes, class probabilities, and confidence scores. YOLOv11n uses decoupled heads, separating classification from localization. This design improves performance in datasets like medical images where object boundaries are subtle and require precise localization.

3.4.4 Training Improvements

The YOLOv11n model was initialized with pretrained weights and fine-tuned on the pediatric wrist radiograph dataset. Training was performed using stochastic gradient descent (SGD) with a batch size of 32 and an input resolution of 640×640 pixels. Mixed-precision training was enabled to improve computational efficiency and reduce memory usage

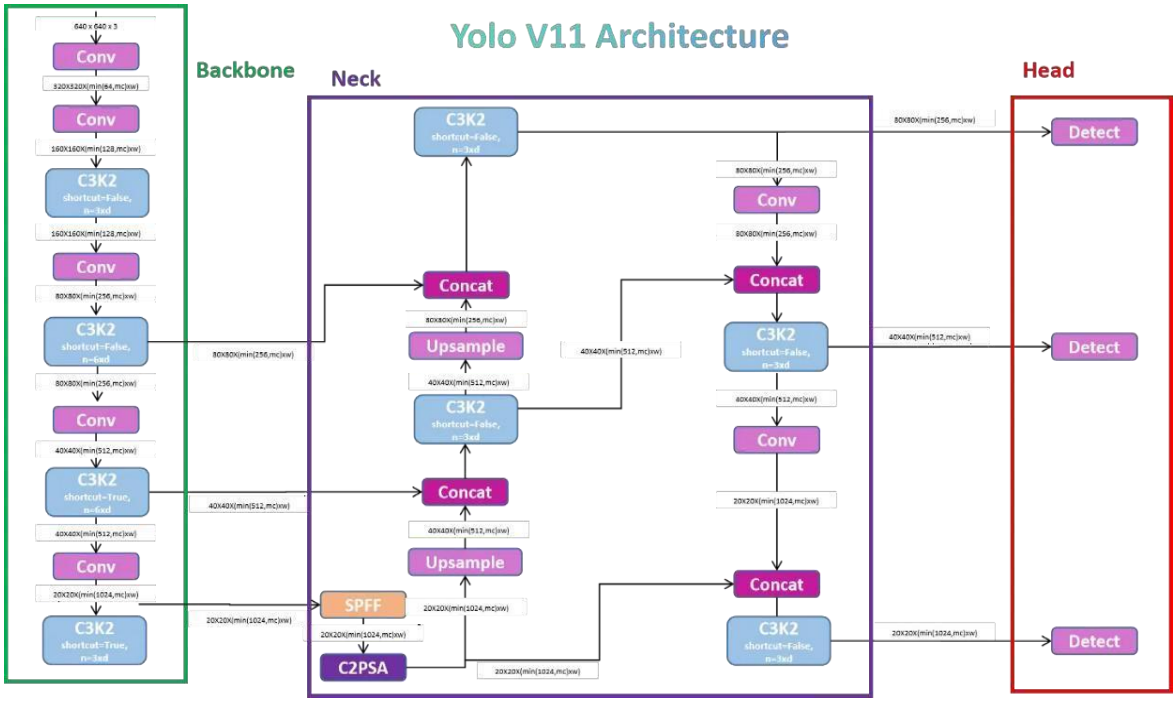

Figure 3.4.4. Architecture of YOLOv11n

Source: https://www.researchgate.net/figure/The-architecture-of-YOLOv11_fig2_390603293.

3.5 Methodology Workflow

The methodology workflow represents the sequential steps from dataset acquisition to model deployment. Each stage is interconnected to ensure the model is trained in a structured and rigorous manner.

The main phases include:

- Dataset Collection: Gathering pediatric wrist X-rays and corresponding fracture annotations.
- Data Cleaning and Preprocessing: Converting images and labels to Ultralytics YOLO format, verifying annotation consistency, and preparing the train/validation/test splits.
- Augmentation: Default data augmentation strategies provided by the YOLO training framework were applied during training.
- Training: Fine-tuning the model on the pediatric dataset.
- Evaluation: Computing metrics such as mAP, precision, recall, and F1-score.
- Validation: Monitoring model performance on the validation set during training and model selection.

- Analysis: Visualizing results, counting false positives/negatives, and measuring localization accuracy.
- Model Export: Saving trained model weights for future inference and potential clinical integration.

Each step is essential in achieving a reliable and clinically applicable fracture detection system.

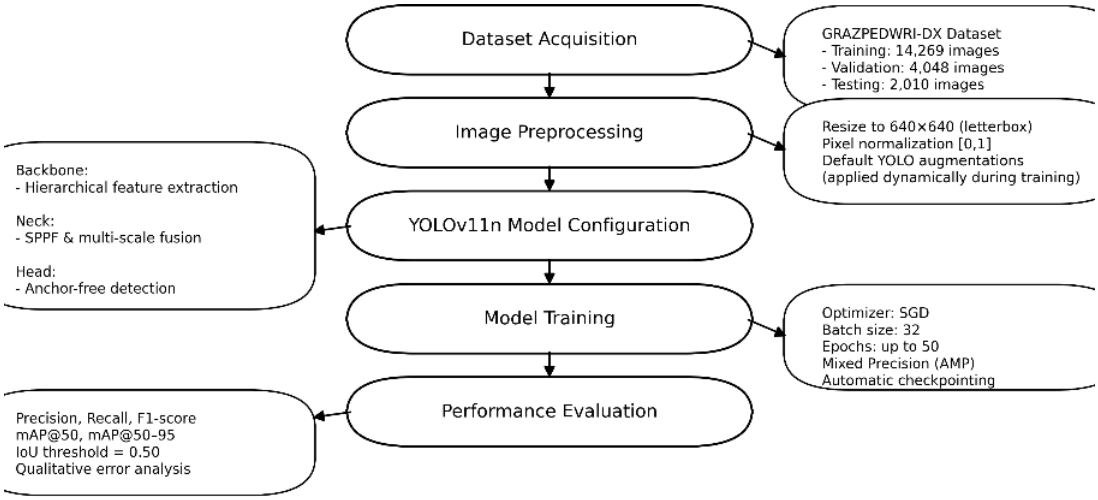

Figure 2.5.1 Methodology Flowchart for pediatric wrist fracture detection using YOLOv11n.

3.6 Training Strategy

3.6.1 Transfer Learning

Model training used transfer learning by initializing from pretrained Ultralytics weights (yolo11n.pt) and fine-tuning on the fracture detection dataset. Training employed Stochastic Gradient Descent (SGD) optimization using Ultralytics default SGD settings (momentum = 0.937 and weight decay = 0.0005) unless explicitly overridden.

Automatic Mixed Precision (AMP) was enabled to improve training throughput and reduce GPU memory usage while maintaining numerical stability.

Early stopping was enabled using patience=10, meaning training terminated if validation performance did not improve for 10 consecutive epochs.

3.6.2 Hyperparameter Configuration

The model was trained using the following configuration (matching the training script):

The hyperparameters were primarily selected to align with the Ultralytics default training configuration for YOLO models, which provides a strong and widely used baseline in object

detection. The input size 640×640 was chosen as a standard YOLO resolution that balances small-object sensitivity with computational efficiency. Batch size 32 was selected to maximize GPU utilization on the Tesla T4 without exceeding memory limits. SGD with lr0 = 0.01, momentum = 0.937, and weight decay = 0.0005 follows Ultralytics defaults and is commonly used for stable convergence in YOLO training. 50 epochs was chosen as an upper bound sufficient for convergence on this dataset, while early stopping (patience = 10) was enabled to prevent overfitting if validation performance stopped improving. AMP was enabled to reduce memory usage and improve training throughput.

Hyperparameter tuning was limited in this study to preserve reproducibility and focus on evaluating YOLOv11n as a baseline detector. No grid search or Bayesian optimization was performed for architecture or optimizer parameters. The only parameter tuned explicitly was the inference confidence threshold, which was selected on the validation set to maximize the F1-score (precision–recall trade-off) prior to final test evaluation.

3.7 Evaluation Metrics

Detection performance was evaluated on the held-out test set using standard object-detection metrics produced by the Ultralytics YOLO validation procedure (val). The reported metrics include precision, recall, F1-score, mAP@0.5, and mAP@0.5:0.95, together with runtime profiling. Counting rules (TP/FP/FN). For each image, the model outputs bounding boxes with confidence scores. Predictions were filtered using a confidence threshold, and duplicate detections were removed using Non-Maximum Suppression (NMS). A prediction was counted as a True Positive (TP) if it overlapped a ground-truth fracture box with IoU ≥ 0.50 and that ground-truth box had not already been matched. Predictions that did not match any ground-truth box under this rule were counted as False Positives (FP). Ground-truth fracture boxes that were not matched by any prediction were counted as False Negatives (FN). This one-to-one matching prevents multiple detections of the same fracture from being counted as multiple true positives.

Metric meanings:

- Precision: the proportion of predicted fractures that are correct (reduces false alarms).
- Recall: the proportion of real fractures that are detected (reduces missed fractures).
- F1-score: a single balanced measure combining precision and recall.
- mAP@0.5: mean average precision at IoU = 0.50 (standard detection benchmark).
- mAP@0.5:0.95: mean average precision averaged across IoU thresholds from 0.50 to 0.95 (stricter localization requirement).

- Confusion outcomes (TP, FP, FN): used for quantitative error analysis.
- Inference time: average per-image runtime split into preprocessing, inference, and postprocessing as reported by Ultralytics.

Threshold reporting. Unless otherwise stated, confidence filtering and NMS settings followed the Ultralytics default configuration. The final confidence threshold used for reporting was selected using the validation set to balance precision and recall.

3.8 Computational Environment

All experiments were conducted on the Kaggle platform using a single NVIDIA Tesla T4 GPU with CUDA acceleration. The training and evaluation pipeline was implemented in Python using PyTorch and the Ultralytics YOLO framework. To ensure reproducibility, the random seed was fixed (seed = 42), and all outputs (checkpoints, logs, and plots) were saved under the specified project directory (fracture_yolo_benchmark/run_seed42). Training was executed on one GPU by setting device="0", with Automatic Mixed Precision enabled (amp=True) to improve throughput and reduce memory usage.

3.9 Evaluation Procedure

After training, the final model was evaluated on a separate held-out test set that was not used during optimization or threshold tuning. The evaluation process included generating predicted bounding boxes, computing detection metrics, analyzing failure cases, and visualizing fracture localization results. Particular attention was given to false negatives, as missed fractures can have significant clinical consequences—especially for faint, small, or anatomically complex cases (e.g., near growth plates). Interpretability analysis (e.g., activation mapping) was not included in the current experimental scope. Future work should incorporate explainability techniques adapted for object detection to highlight image regions that influence model predictions and to strengthen clinical trust for potential adoption.

CHAPTER V
RESULTS

4.1 Overview

This chapter reports the experimental outcomes of applying YOLO11n to pediatric wrist fracture detection using the GRAZPEDWRI-DX dataset. The aim of this chapter is to (1) present the quantitative detection performance achieved on the validation and test partitions, and (2) interpret what these outcomes imply for practical clinical use.

To ensure that reported performance reflects true generalization, the model was evaluated on two unseen partitions:

- Validation set: used during development to monitor performance and support model selection.
- Test set: reserved for final reporting to estimate performance on fully unseen data.

All evaluations were executed under the same computational environment described in Chapter III (Computational Environment), using the Ultralytics built-in validation procedure (val) to compute metrics and runtime profiling.

4.1.1 Deep learning for fracture detection: advances and limitations

Table 4.1.1 summarizes the main detection performance metrics computed on the GRAZPEDWRI-DX validation and test partitions: precision, recall, mAP@50, and mAP@50–95. In addition, the F1-score is reported to provide a single balanced measure of precision and recall.

Tabel 4.1.1 Detection performance on the GRAZPEDWRI-DX validation and test partitions (binary fracture detection).

Split	**Images**	**Instances**	**Precision**	**Recall**	**F1-score**	**mAP@50**	**mAP@50–95**
Validation	4,048	3,604	0.923	0.850	0.885	0.936	0.564
Test	2,010	1,763	0.926	0.870	0.897	0.940	0.552

Note. F1-score is a balanced measure of precision and recall:

$$F1 = 2 \times (precision \times recall) / (precision + recall) \quad (4.0.1.1)$$

Validation F1 ≈ 0.885; Test F1 ≈ 0. 897.

4.1.2 Interpretation

The YOLO11n model demonstrated strong and consistent detection capability across both validation and test partitions. Key observations include:

1. High precision and strong recall (balanced detection).
   Precision was high on both partitions (validation = 0.923; test = 0.926), indicating that most predicted bounding boxes correspond to true fracture findings rather than false alarms. Recall was also strong (validation = 0.850; test = 0.870), showing that the model successfully detects a large proportion of annotated fractures. The resulting F1-scores (validation ≈ 0.885; test ≈ 0.897) confirm a well-balanced tradeoff between false positives and false negatives for this binary detection task.
   Clinically, this balance is important: high precision reduces unnecessary alerts, while high recall reduces missed fractures.
2. Robust detection at IoU = 0.50 (coarse localization performance).
   mAP@50 values were high (validation = 0.936; test = 0.940), demonstrating that the model reliably localizes fracture regions to within the conventional IoU = 0.50 threshold. This level of localization is generally appropriate for triage and decision-support, where the primary goal is to highlight suspicious regions for clinician review rather than provide exact boundary-level delineation.
3. Localization sensitivity under stricter IoU thresholds (tight-box limitation). mAP@50–95 values (validation = 0.564; test = 0.552) are substantially lower than mAP@50, indicating reduced performance when stricter localization accuracy is required. The absolute drops (mAP@50 − mAP@50–95) were 0.372 on validation (0.936 − 0.564) and 0.388 on test (0.940 − 0.552). These correspond to relative reductions of approximately 39.7% (0.372/0.936) and 41.3% (0.388/0.940), respectively.
   This pattern suggests that while the detector typically identifies the correct fracture region, predicted bounding boxes are not always tightly aligned with the annotated boundaries—an important limitation for tasks requiring precise localization (e.g., automated measurement or surgical planning support).
4. Consistency between validation and test results (generalization evidence).
   The close agreement between validation and test metrics (e.g., precision 0.923 vs. 0.926; mAP@50 0.936 vs. 0.940) indicates stable generalization and suggests that the reported performance is not limited to a single split.

4.2 Qualitative detection performance

4.2.1 Representative cases

Qualitative review of test-set outputs corroborated quantitative metrics. Representative examples indicate:

- High spatial overlap between predicted bounding boxes and expert annotations in many displaced or clearly visible fractures (strong cortical discontinuities and fragment separation).
- Consistent detection across a variety of fracture morphologies (transverse, oblique, comminuted) and anatomical locations (distal radius, distal ulna, carpal bones).
- Robustness across projections, with reliable detection in anterior–posterior views and reasonable performance in lateral views for displaced fractures.

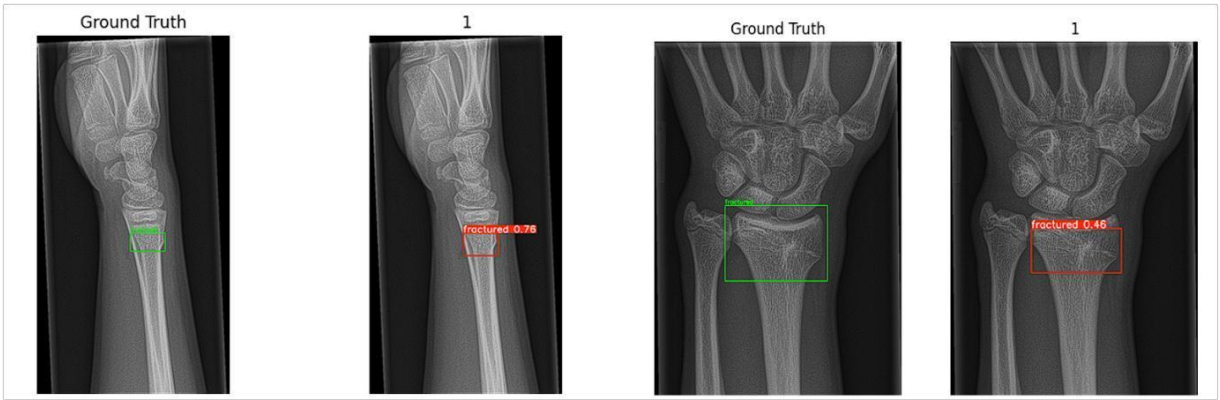

Figure 2.4.1 Representative test image showing YOLOv11n detection (predicted bounding box) closely matching the ground-truth annotation for a wrist fracture.

4.2.2 Common failure modes (qualitative)

Qualitative inspection highlighted several recurrent failure patterns:

- Missed hairline/non-displaced fractures — subtle radiolucent lines with minimal cortical disruption were occasionally undetected, especially in lateral projections where overlapping anatomy can obscure fracture lines. Additionally, fractures near growth plates or in regions with minimal cortical disruption were frequently missed, often due to the presence of normal anatomical features such as incomplete ossification and growth-plate regions. Other common imaging artifacts, such as motion blur, poor image contrast, and low exposure settings, also led to misinterpretation or missed fractures, particularly in challenging pediatric cases.

- False positives near physiologic radiolucency — predicted boxes were sometimes placed on growth-plate (physeal) regions or normal cortical irregularities that mimic fracture appearance in pediatric bone.
- Ambiguous findings — in some cases, imaging artifacts, motion blur, or suboptimal exposure produced appearances that challenged both the model and human readers.

These qualitative observations match the discrepancy seen between high mAP@50 and lower mAP@50–95 scores.

4.3 Training dynamics and convergence behavior

4.3.1 Comparison with related work

Table 4.3 contextualizes the reported test performance against published results on the GRAZPEDWRI-DX dataset. Because studies may use different splits, label definitions (fracture-only vs multi-label), and training/evaluation settings, these values are presented for benchmarking rather than strict head-to-head comparison.

Table 4.3 Comparison with published GRAZPEDWRI-DX results (mAP@0.50)

Study	Task / labels (as reported)	Model	Metric reported	Reported result
This study	Binary fracture-only detection	YOLO11n (Ultralytics)	mAP@0.50	0.940
Nagy et al. (2022)	Fracture object detection baseline (dataset paper)	YOLOv5m	mAP@0.50	0.933
Ju & Cai (2023)	Multi-class detection (fracture + other findings)	YOLOv8	mAP@0.50	0.638
Olczak et al. (2017)	Extremity fracture detection	ResNet-based CNN	Accuracy	0.83
Lindsey et al(2018)	Wrist fracture detection	DenseNet-based CNN	AUC	0.94

4.3.2 Loss and metric trajectories

Training loss components—including classification loss, bounding-box regression loss, and distribution focal loss—declined rapidly during the early epochs, followed by a slower, asymptotic convergence phase after approximately epoch 10. Validation metrics, particularly precision, improved quickly in the initial stages and reached a plateau earlier than recall. This pattern suggests an initial reduction in false positives, followed by a more gradual improvement in sensitivity to subtle fracture features, corresponding to a reduction in false negatives.

An early-stopping mechanism was configured to monitor stagnation of validation mAP@50, but training continued through all 50 epochs, with the best validation performance observed at epoch 40. The checkpoint corresponding to this maximum validation mAP@50 was saved as best.pt and selected as the final model for subsequent evaluation and reporting.

4.3.3 Experimental Setup

The close correspondence between training and validation loss curves, without substantial divergence, indicates that the applied regularization (weight decay), data augmentation strategy, and early-stopping configuration were effective in preventing overfitting on the training partition. The earlier stabilization of precision relative to recall further suggests that additional targeted strategies—such as multi-view image fusion or augmented exposure to hairline fractures—may be required to accelerate recall improvement without compromising precision.

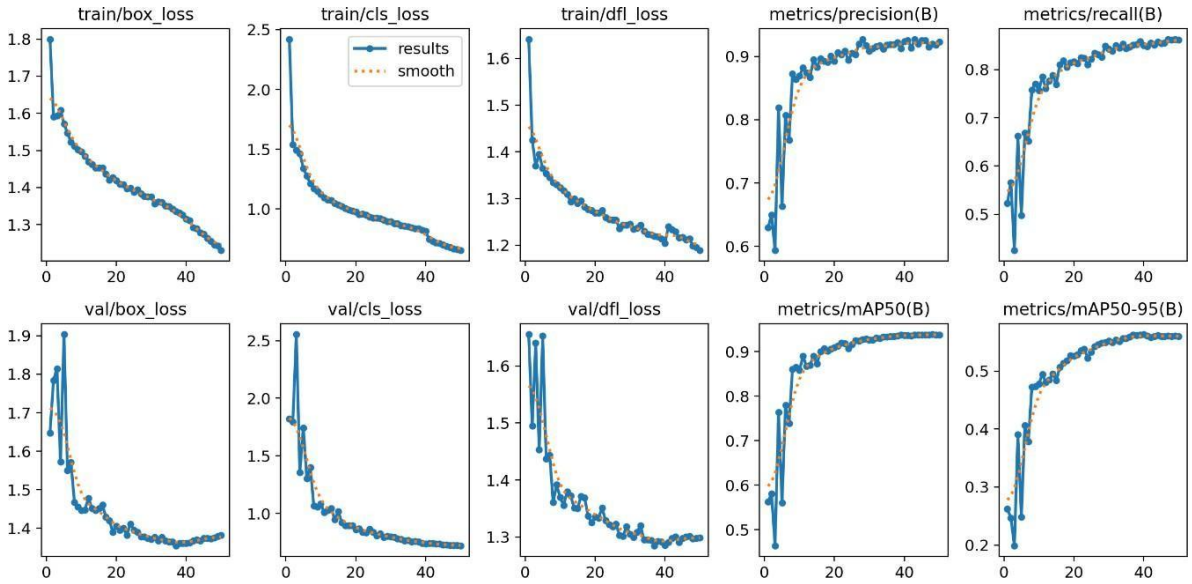

Figure 4.3.2 Training and validation losses, precision, recall, and mAP metrics across epochs, showing stable convergence and effective model generalization.

4.4 Computational Performance

Computational performance is essential for clinical feasibility. Even an accurate detector may be impractical if it is slow or resource-intensive. Ultralytics reports per-image timing split into preprocessing, inference, and postprocessing. These stages represent:

- Preprocessing: image resizing/normalization and data formatting.
- Inference: forward pass through the neural network (core model computation).
- Postprocessing: confidence filtering and suppression of duplicate predictions (e.g., NMS behavior inside the pipeline).

Validation speed: 0.8 ms preprocess, 2.5 ms inference, 1.0 ms postprocess (≈ 4.3 ms/image).
Test speed: 1.0 ms preprocess, 2.4 ms inference, 1.1 ms postprocess (≈ 4.5 ms/image).

The end-to-end latency is below 5 ms per image on a Tesla T4 GPU. This is compatible with near-real-time workflows where radiographs must be processed immediately after acquisition or during rapid clinical review. The breakdown also shows that postprocessing is a non-trivial portion of runtime, which is typical for object detection systems due to duplicate removal and thresholding.

4.5 Quantitative failure-mode analysis

A failure-mode breakdown was computed using true positives (TP), false positives (FP), and false negatives (FN). This analysis provides insight into the model's detection performance and highlights the relative contribution of correct and incorrect predictions. The totals satisfy the dataset instance relationship:

$$TP + FN = 1{,}763$$

which is consistent with the total number of instances in the test set.

Table 5.2 Confusion-outcome breakdown (test set)

Outcome	Count	Percentage (%)
True Positives (TP)	1,533	81.37
False Positives (FP)	121	6.42
False Negatives (FN)	230	12.21

This distribution is consistent with the reported test metrics:

Precision, which measures the proportion of predicted positives that are correct, is calculated as:

$$\text{Precision} \approx TP/(TP+FP) = 1533/(1533+121) \approx 0.926$$

A high precision indicates that the model generates few false alarms and is reliable in identifying true fractures.

- Recall, which measures the proportion of actual positives that are correctly identified, is calculated as:

$$\text{Recall} \approx TP/(TP+FN) = 1533/(1533+230) \approx 0.870$$

This demonstrates that the model successfully detects the majority of true fractures, though some subtle cases remain challenging.

Interpretation**:** The confusion-outcome analysis confirms that the model achieves a strong balance between precision and recall, reflecting both its ability to accurately detect fractures and its sensitivity to real fracture instances. Understanding the distribution of TP, FP, and FN provides a quantitative foundation for assessing model reliability and identifying areas for future improvement.

4.6 Clinical implications of results

The reported results support the use of YOLO11n as a triage-assist or second-reader tool in pediatric wrist fracture workflows:

- High precision (~0.93): reduces unnecessary alerts and minimizes workflow interruptions caused by excessive false positives.
- Strong recall (~0.87): indicates that most fractures are flagged for clinician review, improving safety in high-throughput settings.
- Sub-5: ms inference enables near-instantaneous processing, allowing integration into time-sensitive pathways such as emergency department imaging review.

However, outputs should be considered assistive, not autonomous: clinical decision-making should remain with qualified clinicians, especially given residual false negatives and localization tightness limitations.

4.7 Limitations evident from results

The following limitations are directly supported by the observed results:

1. Localization precision at strict IoU thresholds. mAP@50–95 remains moderate (~0.55), indicating the model often provides coarsely correct bounding boxes but less frequently produces the tight localization needed for exact measurement or surgical planning.
2. Dataset generalizability. Results were obtained using GRAZPEDWRI-DX only; external validation across other pediatric multi-institutional datasets is required before strong claims about cross-institutional robustness can be made. To enhance generalizability and mitigate domain shift, future studies should consider using a federated learning setup, which would allow model training on multiple pediatric datasets across institutions without sharing sensitive patient data. This approach could also involve incorporating datasets from diverse demographic regions, imaging devices, and radiographic protocols to ensure robustness across various clinical settings.
3. Anatomical confounders. Growth-plate regions and normal pediatric cortical variations produce a substantial portion of false positives, necessitating dedicated modeling strategies to disambiguate these features from true fractures.
4. Quantitative comparisons lacking formal significance testing. This study did not perform a controlled comparison against other YOLO versions (e.g.,

YOLOv5/YOLOv8) under identical training and evaluation conditions; therefore, no superiority claims are made.

Single-image inference profiling. Timing results are measured in single-image mode on a T4 GPU; different deployment environments (CPU only, mobile, or edge devices) will exhibit different latency characteristics and may require further optimization.

4.8 Recommendations for future improvement

Based on the above findings, the following directions are recommended:

- Localization sensitivity at strict IoU thresholds.

  mAP@50–95 (~0.55) is substantially lower than mAP@50 (~0.94), indicating that tight bounding-box agreement is less consistent under strict evaluation criteria.
- Residual missed fractures (false negatives).

  FN = 230 (12.21%) indicates that a clinically meaningful fraction of fractures may be missed, requiring safeguards such as mandatory clinician review and conservative operating thresholds.
- Residual false alarms (false positives).

  FP = 121 (6.42%) indicates that some detections occur in non-fracture regions; threshold calibration and clinician confirmation remain necessary.
- Hardware- and environment-specific runtime.

  Timing results reflect evaluation on a Tesla T4 GPU; latency will differ on CPUs, mobile/edge devices, or other GPUs.

4.9 Discussion

Overall, YOLO11n demonstrated strong fracture detection capability on GRAZPEDWRI-DX, with consistent validation and test performance and near-real-time inference speed on a Tesla T4 GPU. The model achieved mAP@50 = 0.936 (validation) and 0.940 (test), with precision ≈ 0.923–0.926 and recall ≈ 0.850–0.870, indicating stable detection behavior and good generalization from validation to test.

A key observation is the performance gap between mAP@50 and mAP@50–95. This gap indicates that while the detector reliably identifies fractures and places bounding boxes in the correct region, the predicted boxes may not always be tightly aligned with the annotation boundaries. For clinical deployment, this implies strong suitability for screening/triage and prioritization, while more demanding tasks requiring tight geometric localization may benefit

from refinement strategies (e.g., higher-resolution evaluation, two-stage cropping, or complementary segmentation-based refinement).

Finally, the failure-mode distribution (TP 81.37%, FP 6.42%, FN 12.21%) shows that most fractures are detected, but missed cases remain non-trivial. Therefore, the most appropriate clinical role is as a decision-support system that reduces oversight risk and accelerates review, rather than a replacement for expert interpretation.

CHAPTER VI
CONCLUSIONS AND FUTURE WORK

In conclusion, this thesis demonstrates that a lightweight YOLOv11n-based object detector can achieve strong practical performance for pediatric wrist fracture detection on the GRAZPEDWRI-DX dataset. Using a held-out test partition, the final model achieved precision = 0.926, recall = 0.870, and F1-score = 0.897, with robust detection performance at the conventional localization threshold (mAP@50 = 0.940). Under stricter localization requirements, performance decreased (mAP@50–95 = 0.552). This gap is expected because mAP@50 evaluates detections at a single, relatively lenient IoU threshold (IoU $\geq$ 0.50), whereas mAP@50–95 averages AP across multiple IoU thresholds (0.50 to 0.95) and therefore penalizes bounding boxes that are not tightly aligned with the ground-truth annotations. In pediatric wrist radiographs, tight localization is inherently challenging because fractures are often small, subtle, and non-displaced, may appear as faint cortical irregularities, and frequently occur near growth plates or overlapping anatomy; in such cases, even minor deviations in box placement substantially reduce IoU under higher thresholds. Therefore, the results indicate that YOLOv11n reliably identifies and highlights the suspicious region for decision-support and triage use, while precise boundary-level alignment remains less consistent and may benefit from refinement strategies (e.g., higher-resolution training or a second-stage localization module). Runtime profiling on an NVIDIA Tesla T4 GPU showed sub–5 ms per-image end-to-end latency, supporting feasibility for near-real-time decision-support use in emergency and high-throughput radiology settings. Given its nano-scale architecture and low computational overhead, YOLOv11n is also well suited for deployment in resource-constrained environments where hardware capacity may be limited. Overall, these findings support the use of the proposed system as a triage-assist or second-reader tool that highlights suspicious regions for clinician confirmation, rather than an autonomous diagnostic solution, consistent with safety expectations in clinical AI translation (Chen et al., 2021; Mader et al., 2021). The main contributions of this study can be summarized as follows. First, an end-to-end YOLOv11n-based fracture detection pipeline was implemented for pediatric wrist radiographs, including dataset preparation, conversion to Ultralytics YOLO format, and a single-class detection setup via label remapping. Second, the study provides a systematic quantitative evaluation of YOLOv11n on GRAZPEDWRI-DX using standard object-detection metrics (precision, recall, F1-score, mAP@50, and mAP@50–95) together with runtime

profiling to assess computational feasibility. Third, the work includes structured error characterization, identifying clinically meaningful failure patterns related to subtle fracture morphology and pediatric anatomical confounders. Fourth, the results establish evidence that a lightweight detector can achieve a favorable accuracy–efficiency trade-off, supporting potential use in settings with limited computational resources.

These outcomes directly address the research questions posed in Chapter I. RQ1 is answered by the reported detection performance on the held-out test set, showing that YOLOv11n can detect pediatric wrist fractures with high precision and strong recall while maintaining robust mAP@50 performance. RQ2 is addressed by clarifying that a controlled baseline experiment under identical training and evaluation conditions was not conducted in this work; therefore, the thesis does not claim superiority over prior YOLO variants, although the achieved results provide a useful benchmark for comparison with published values on the same dataset. RQ3 is addressed through the qualitative and quantitative error analysis, which indicates that false negatives occur most often in subtle or non-displaced fractures, particularly when fracture appearance is faint or near complex anatomical structures, while false positives are commonly associated with pediatric growth-plate/physeal regions and other normal radiographic variations that can mimic fracture-like patterns. RQ4 is addressed by combining diagnostic performance with computational evidence: the detector's low-latency inference and lightweight design indicate feasibility for integration into pediatric radiology workflows, including resource-limited environments where efficiency and speed are operational constraints.

Despite these encouraging findings, several limitations must be considered before clinical deployment. The gap between mAP@50 and mAP@50–95 indicates that localization precision under strict IoU thresholds remains moderate, which may limit applicability for tasks requiring tight geometric localization. In addition, residual false negatives and false positives—especially those associated with subtle fractures and growth-plate confounding—underscore that the system should be used as an assistive tool with mandatory clinician oversight. Finally, results were obtained on a single benchmark dataset, and the absence of external multi-institutional testing limits generalizability claims. Concretely, recommended next steps include: (1) external validation across geographically and technologically diverse institutions to quantify generalization and domain shift, with consideration of domain-adaptive or federated approaches where appropriate (Chen et al., 2021; Mader et al., 2021); (2) development of a two-stage refinement pipeline in which YOLOv11n provides coarse localization followed by a

high-resolution refinement module (e.g., segmentation-augmented refinement) to improve tight localization and increase performance under stricter IoU thresholds; (3) investigation of multi-view fusion between AP and lateral projections and incorporation of pediatric anatomical priors to reduce confusion around growth plates and developmental variants; (4) extension beyond binary detection toward richer outputs such as fracture subtype and anatomical site labeling; (5) incorporation of explainability and uncertainty estimation (e.g., activation-based visualization and uncertainty-aware methods) to support safe human–AI interaction and improve clinician trust; (6) prospective user studies to measure impact on workflow efficiency and diagnostic accuracy; (7) engineering a PACS/DICOM-compatible deployment pathway with audit trails and ongoing performance monitoring; and (8) strengthening reproducibility by releasing training code, split indices, and model weights where licensing permits, consistent with best practices in medical imaging research (Shorten & Khoshgoftaar, 2019; Wang et al., 2020). Collectively, these steps are necessary to transition from a strong benchmark result to a responsibly validated clinical decision-support tool that can improve pediatric musculoskeletal care while meeting ethical, legal, and usability requirements (Chen et al., 2021; Mader et al., 2021).

REFERENCE

Cohen, J. F., & McInnes, M. D. F. (2022). Deep learning algorithms to detect fractures: Systematic review shows promising results but many limitations. *Radiology, 304*(1), 63–64.*

Franco, P. N., Maino, C., Mariani, I., et al. (2024). Diagnostic performance of an AI algorithm for the detection of appendicular fractures in pediatric patients. *European Journal of Radiology, 178*, 111637.

Hansen, V., Jensen, J., Weber Kusk, M., Gerke, O., Tromborg, H. B., & Lysdahlgaard, S. (2024). Deep learning performance compared to healthcare experts in detecting wrist fractures from radiographs: A systematic review and meta-analysis. *European Journal of Radiology, 174*, 111399.

Kalmet, P. H. S., Sanduleanu, S., Primakov, S., et al. (2020). Deep learning in fracture detection: A narrative review. *Acta Orthopaedica, 91*(2), 215–220.

Kazi, A., Holzinger, A., Zbinden, A., et al. (2025). YOLOv7-based detection of pediatric wrist fractures in the GRAZPEDWRI-DX dataset. *Preprint*.

Khan, M., Rahman, T., & Ali, S. (2025). Real-time bone fracture detection using YOLOv10: A deep learning approach on X-ray images. *Preprint*.

Korup, L. R., Sørensen, S. W., Søballe, K., & Husted, H. (2022). Children's distal forearm fractures: A population-based epidemiology study of 4,316 fractures. *Bone & Joint Open, 3*(2), 146–154.

Mamoowala, N., Johnson, N. A., & Dias, J. J. (2019). Trends in paediatric distal radius fractures: An eight-year review from a large UK trauma unit. *Annals of the Royal College of Surgeons of England, 101*(4), 297–303.

Mattijssen-Horstink, L., Langeraar, J., Slaar, A., et al. (2020). Radiologic discrepancies in diagnosis of fractures in a Dutch teaching emergency department: A retrospective analysis. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 28*, 47.

Shelmerdine, S. C., Pauling, C., Allan, E., et al. (2024). Artificial intelligence for paediatric fracture detection: A multireader multicase study protocol. *BMJ Open, 14*(12), e084448.

Smith, J. E., Tse, S., Barrowman, N., & Bilal, A. (2016). Missed fractures on radiographs in a pediatric emergency department. *Canadian Journal of Emergency Medicine, 18*(S1), S119.*

Su, Z., Adam, A., Nasrudin, M. F., Ayob, M., & Punganan, G. (2023). Skeletal fracture detection with deep learning: A comprehensive review. *Diagnostics, 13*(20), 3245.

Südow, H., & Mellstrand Navarro, C. (2021). The incidence of distal radius fractures in a Swedish paediatric population: An observational cohort study of 90,970 individual fractures. *BMC Musculoskeletal Disorders, 22*, 564.

Tariq, M., Ahmad, S., & Rahman, A. (2025). YOLOv11-driven deep learning for enhanced bone fracture detection and quantification. *Preprint*.

Mader, S., Sze, W. T., Nwaneri, S., et al. (2021). A comparative study of deep learning-based object detection models for fracture diagnosis in X-ray images. Journal of Medical Imaging and Health Informatics, 11(4), 1045–1052. https://doi.org/10.1166/jmihi.2021.3368

Wei, C.-J., Tsai, W.-C., Tiu, C.-M., et al. (2006). Systematic analysis of missed extremity fractures in emergency radiology. *Acta Radiologica, 47*(7), 710–717.

Wei, W., Huang, Y., Zheng, J., Rao, Y., Wei, Y., Tan, X., & OuYang, H. (2025). YOLOv11-based multi-task learning for enhanced bone fracture detection and classification in X-ray images. *Journal of Radiation Research and Applied Sciences, 18*(1), 101309.

Zbinden, A., Zbinden, R., Weinberg, A., & Kazi, A. (2021). GRAZPEDWRI-DX: A pediatric wrist trauma X-ray dataset with fracture ground truth. *Data in Brief, 39*, 107593.

Ziegner, M., Pape, J., Lacher, M., et al. (2025). Real-life benefit of artificial intelligence–based fracture detection in a pediatric emergency department. *European Radiology, 35*(10), 5881–5890.

Ahmed, A., & Manaf, A. (2024). Pediatric wrist fracture detection in X-rays via YOLOv10 algorithm and dual label assignment system. *arXiv preprint arXiv:2407.15689*.

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Hansen, V., Jensen, J., Weber Kusk, M., Gerke, O., Tromborg, H. B., & Lysdahlgaard, S. (2024). Deep learning performance compared to healthcare experts in detecting wrist

fractures from radiographs: A systematic review and meta-analysis. *European Journal of Radiology, 174*, 111399.

Husarek, J., et al. (2024). Diagnostic accuracy of deep learning in orthopaedic fractures: A systematic review and meta-analysis. *Scientific Reports, 14*, 73058.

Ju, R.-Y., & Cai, W. (2023). Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm. *Scientific Reports, 13*, 20077.

Kalmet, P. H. S., Sanduleanu, S., Primakov, S., et al. (2020). Deep learning in fracture detection: A narrative review. *Acta Orthopaedica, 91*(2), 215–220.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Liao, J. C.-Y., & Chong, A. K. S. (2019). Pediatric hand and wrist fractures. *Clinics in Plastic Surgery, 46*(3), 425–436.

Lin, T.-Y., Maire, M., Belongie, S., et al. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision* (pp. 740–755).

Luhmann, S. J., Schootman, M., Schoenecker, P. L., Dobbs, M. B., & Gordon, J. E. (2004). Complications and outcomes of open pediatric forearm fractures. *Journal of Pediatric Orthopaedics, 24*(1), 1–6.

Mamoowala, N., Johnson, N. A., & Dias, J. J. (2019). Trends in paediatric distal radius fractures: An eight-year review from a large UK trauma unit. *Annals of the Royal College of Surgeons of England, 101*(4), 297–303.

Mattijssen-Horstink, L., Langeraar, J., Slaar, A., et al. (2020). Radiologic discrepancies in diagnosis of fractures in a Dutch teaching emergency department: A retrospective analysis. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 28*, 47.

Nagy, E., Janisch, M., Hržić, F., Sorantin, E., & Tschauner, S. (2022). A pediatric wrist trauma X-ray dataset (GRAZPEDWRI-DX) for machine learning. *Scientific Data, 9*, 222.

Ramadanov, N., et al. (2025). Artificial intelligence-guided distal radius fracture detection on radiographs compared with human raters. *Journal of Orthopaedic Surgery and Research, 20*, 5888.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems* (pp. 91–99).

Smith, J. E., Tse, S., Barrowman, N., & Bilal, A. (2016). Missed fractures on radiographs in a pediatric emergency department. *Canadian Journal of Emergency Medicine, 18*(S1), S119.

Su, Z., Adam, A., Nasrudin, M. F., Ayob, M., & Punganan, G. (2023). Skeletal fracture detection with deep learning: A comprehensive review. *Diagnostics, 13*(20), 3245.

Wei, C.-J., Tsai, W.-C., Tiu, C.-M., et al. (2006). Systematic analysis of missed extremity fractures in emergency radiology. *Acta Radiologica, 47*(7), 710–717.

Wei, W., Huang, Y., Zheng, J., Rao, Y., Wei, Y., Tan, X., & OuYang, H. (2025). YOLOv11-based multi-task learning for enhanced bone fracture detection and classification in X-ray images. *Journal of Radiation Research and Applied Sciences, 18*(1), 101309.

Zech, J. R., Carotenuto, G., Igbinoba, Z., et al. (2023). Detecting pediatric wrist fractures using deep-learning-based object detection. *Pediatric Radiology, 53*(6), 1125–1134.

Zhang, W., et al. (2025). Rehabilitation-driven optimized YOLOv11 model for accurate X-ray fracture detection. *Sensors, 25*(18), 5793.

Ziegner, M., Pape, J., Lacher, M., et al. (2025). Real-life benefit of artificial intelligence-based fracture detection in a pediatric emergency department. *European Radiology, 35*(10), 5881–5890.

Nagy, E., Janisch, M., Hržić, F., Sorantin, E., & Tschauner, S. (2022). A pediatric wrist trauma X-ray dataset (GRAZPEDWRI-DX) for machine learning. Scientific Data, 9, 222. https://doi.org/10.1038/s41597-022-01328-z

Ju, R.-Y., & Cai, W. (2023). Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm. Scientific Reports, 13, 20077. https://doi.org/10.1038/s41598-023-47460-7

Chien, C.-T., Ju, R.-Y., Chou, K.-Y., Xieerke, E., & Chiang, J.-S. (2024). YOLOv8-AM: YOLOv8 based on effective attention mechanisms for pediatric wrist fracture detection (arXiv:2402.09329) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2402.09329

Ju, R.-Y., Chien, C.-T., Lin, C.-M., & Chiang, J.-S. (2024). Global context modeling in YOLOv8 for pediatric wrist fracture detection (arXiv:2407.03163) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.03163