Thesis

Automated fracture detection in children using difficult cases from routine pediatric traumatology

submitted by

Daniel Matthias Stütz

in partial fulfillment of the requirements for the degree of

Doktor der gesamten Heilkunde

(Dr. med. univ.)

at the

Medical University of Graz

executed at the

Division of Pediatric Radiology

Department of Radiology

under the supervision of

Sebastian Tschauner MD PhD

and

Michael Janisch, MD

Graz, 30.03.2026

Declaration of Academic Integrity

I hereby confirm that the present diploma thesis is the result of my own independent scholarly work. I also confirm that in all cases, where material from the work of others (in books, articles, essays, dissertations, and on the internet) is acknowledged, quotations and paraphrases are clearly indicated. No material other than that cited in the reference list has been used. I have read and understood the Medical University's regulations and procedures concerning plagiarism.

Furthermore, I hereby declare that if artificial intelligence (AI) tools were used for the generation and/or correction of certain text passages in the creation of this work, such employment was conducted in compliance with ethical principles, academic integrity, and the regulations of my university. Additionally, it was ensured that this usage was transparently disclosed and appropriately attributed.

Graz, 30.03.2026 Daniel Matthias Stütz m.p.

Acknowledgements

Firstly, I would like to express my profound appreciation to my supervisors Sebastian Tschauner MD, PhD and Michael Janisch MD. In particular, I am deeply indebted to Sebastian Tschauner: thank you for your wisdom, guidance, unwavering support, patience and encouragement throughout every single step of this journey. I feel truly privileged to have been under your mentorship.

I am forever grateful to my loving and supportive family. To my mother Judith, my father Thomas, my sister Theresa and my grandmother Gisela and my late grandfather Daniel. Your constant love and encouragement have been the foundation of everything I have achieved. Also, I am immeasurably thankful for my godparents Anna and Robert for their continuous support and for believing in me no matter what.

To my dear friend Christoph - thank you for standing by my side all these years. The long and exhausting nights we shared in the library as well as our travels have become some of the most precious memories of my life. Your friendship means more to me than words can say.

Furthermore, I want to express my gratitude for my dear friends Samuel, Francis, Lukas, Johannes, Florian who have filled this journey with countless nights and days of laughter and joy.

A very special place in my heart belongs to Viktoria, my best friend. Your untiring presence and your light have carried me through difficult days and made the good ones even brighter. Thank you for being in my life.

I am deeply grateful to my dear friends Letizia and Fabian: thank you for your friendship and all the precious moments we shared.

Equally to my dear friends Paul, Andreas, Lorena, Eva, Verena, Hanna and Eva for the countless moments of true happiness. Thank you.

My dear friend David, thank you for your companionship and the joy you have brought into my life.

Finally, to Molly and Timmy – thank you for your cuddles and warmth, you will forever hold a special place in my heart.

Zusammenfassung

Hintergrund: Frakturen im Kindesalter stellen ein erhebliches Gesundheitsproblem dar und gehören zu den häufigsten Gründen für den Besuch in der Notaufnahme. Außerhalb der regulären Dienstzeiten der Kinderradiolog*innen werden Röntgenbilder von der Universitätsklinik für Kinder- und Jugendchirurgie befundet. Am darauffolgenden Arbeitstag werden diese Bilder vom ärztlichen Personal der Universitätsklinik für Kinderradiologie überprüft, wobei es gelegentlich zu Diskrepanzen zwischen den Befunden der beiden Abteilungen kommt. Diese Fälle werden in der Abteilung für Kinderradiologie auf einer internen Liste dokumentiert. Computer-Vision-Modelle haben bemerkenswerte Ergebnisse in der automatisierten Frakturerkennung gezeigt, allerdings kann ihre Leistung je nach Zusammensetzung der zugrundeliegenden Testdatensätze variieren. Wir gehen davon aus, dass die beiden KI-Algorithmen EfficientNet und YOLOv8 mit zunehmender Fallkomplexität eine signifikant geringere diagnostische Gesamtleistung in einem Testdatensatz mit schwierigen Fällen im Vergleich zu einem abgeglichenen Testdatensatz aufzeigen.

Material und Methoden: Es wurden zwei Testdatensätze retrospektiv erstellt: einer bestehend aus Röntgenbildern mit “schwierigen“ Fällen (n=786), bei denen es anfänglich zu besagten Diskrepanzen zwischen den diagnostischen Befunden gekommen ist, und ein Datensatz (n=786), der anhand Körperregion, Alter und Frakturrate angepasst wurde. Klassifizierung (EfficientNet) und Objekterkennung (YOLOv8) wurden anhand gängiger Leistungsmetriken für maschinelles Lernen bewertet: Precision, Recall, F1-score, Accuracy und PR AUC. Die statistische Signifikanz wurde mit dem Wilcoxon Signed Rank Test ermittelt.

Ergebnisse: Die diagnostische Leistung beider Modelle war beim schwierigen Testsatz im Vergleich zum angepassten durchgängig geringer. EfficientNet: mittlere Precision 0,683 (± 0,030) vs. 0,766 (± 0,057), mittlerer Recall 0,627 (± 0,017) vs. 0,754 (± 0,051), mittlere Accuracy 0,660 (± 0,017) vs. 0,764 (± 0,052), mittlerer F1-score 0,615 (± 0,019) vs. 0,757 (± 0,053), mittlere PR AUC 0,682 (± 0,033) vs. 0,816 (± 0,062) und YOLOv8: mittlere Precision 0,724 (± 0,016) vs. 0,928 (± 0,018), mittlerer Recall 0,467 (± 0,022) vs. 0,858 (± 0,026), mittlere mAP50 0,553 (± 0,022)

vs. 0,932 (± 0,021), mittlere mAP50-95 0,253 (± 0,006) vs. 0,709 (± 0,024), mittlerer F1-score 0,568 (± 0,017) vs. 0,891 (± 0,022) und mittlerer PR AUC 0,735 (± 0,024) vs. 0,929 (± 0,029). Die Ergebnisse waren statistisch signifikant ($p < 0,05$).

Schlussfolgerung: Es wurde eine erheblich geringere diagnostische Leistung der Modelle EfficientNet und YOLOv8 samt Untervarianten bei der Klassifizierung und Erkennung von Frakturen im Testdatensatz mit schwierigen Fällen im Vergleich zum angepassten Datensatz nachgewiesen.

Abstract

Background: Pediatric fractures represent a significant health concern and are one of the most common reasons for emergency department visits. Outside the regular office hours of pediatric radiologists, trauma radiographs are initially reviewed by the Department of Pediatric and Adolescent Surgery. The following workday, these images are re-evaluated by physicians of the Division of Pediatric Radiology. Occasionally, discrepancies between the findings of the two departments arise. These discordant cases are documented on an internal list maintained by the Division of Pediatric Radiology.

Computer Vision models have demonstrated remarkable results in automated fracture detection. However, their performance may vary depending on the composition of the test set. We assume that the two AI-algorithms EfficientNet and YOLOv8 struggle significantly with increasing case complexity and demonstrate lower overall diagnostic performance in a test set comprising difficult cases in comparison to a matched set.

Material and Methods: Two retrospectively acquired datasets were crafted. One consisting of “difficult” radiographs, obtained from cases with initial discrepancies of the diagnostic findings (n=786) and one matched test set (n=786). Both test sets have been matched according to body region, age and fracture rate. The models were assessed in classification (EfficientNet) and object detection (YOLOv8), using common machine learning performance metrics: Precision, Recall, F1-score, Accuracy and PR AUC. Statistical significance was evaluated with the Wilcoxon Signed Rank Test.

Results: Diagnostic performance in the difficult test set vs. the matched test set was consistently lower across all subvariants of EfficientNet: mean Precision 0,683 (± 0,030) vs. 0,766 (± 0,057), mean Recall 0,627 (± 0,017) vs. 0,754 (± 0,051), mean Accuracy 0,660 (± 0,017) vs. 0,764 (± 0,052), mean F1-score 0,615 (± 0,019) vs. 0,757 (± 0,053), mean PR AUC 0,682 (± 0,033) vs. 0,816 (± 0,062) and YOLOv8: mean Precision 0,724 (± 0,016) vs. 0,928 (± 0,018), mean Recall 0,467 (± 0,022) vs. 0,858 (± 0,026), mean mAP50 0,553 (± 0,022) vs. 0,932 (± 0,021), mean mAP50-95 0,253 (± 0,006) vs. 0,709 (± 0,024), mean F1-score 0,568 (± 0,017) vs. 0,891 (±

0,022) and mean PR AUC 0,735 (± 0,024) vs. 0,929 (± 0,029), respectively. The results were statistically significant ($p < 0,05$).

Conclusion: We have demonstrated the overall lower diagnostic performance of the models EfficientNet and YOLOv8 across all subvariants on the tasks of fracture classification and fracture detection in the difficult set compared to the matched test set.

Disclosures

A scientific manuscript of the thesis topic entitled ***"Impact of Test Set Composition on AI Performance for Pediatric Radiograph Appendicular Skeleton Fracture Detection"*** has been published in the journal "Radiology" (1). Author list: Nikolaus Stranger, Mario Scherkl, Daniel Stütz, Michael Janisch, Georg Singer, Saskia Hankel, Holger Till, Maximilian Sackl, Franko Hrzic, Sereina Annik Herzog, and Sebastian Tschauner.

Table of Contents

Abbreviations

Artificial intelligence – AI

Machine Learning – ML

Computer Vision – CV

Deep learning – DL

Convolutional neural network – CNN

Path Aggregation Network – PANet

Mobile Inverted Bottleneck – MBConv

Ionizing Radiation – IR

Computed tomography – CT

As Low as Reasonably Achievable – ALARA

Magnetic Resonance Imaging – MRI

Elastic stable intramedullary nailing – ESIN

Area Under the Curve – AUC

Digital Imaging and Communications in Medicine – DICOM

Picture Archiving and Communication System – PACS

Portable Network Graphics – PNG

True Positive – TP

True Negative – TN

False Positive – FP

False Negative – FN

Precision-Recall Area Under the Curve – PR AUC

National Institute of Standards and Technology – NIST

Gradient-weighted Class Activation Mapping – Grad-CAM

Table of Figures

List of Tables

1 Introduction

1.1 Artificial Intelligence

Over the last years, artificial intelligence (AI) has gained widespread attention due to advancements in areas such as speech recognition (2) and natural language processing (3). These recent successes are largely attributable to the substantial increase in computational power and availability of data (4). The term ”Artificial Intelligence” itself is not well defined but can be broadly understood as the simulation of human intelligence.

Machine learning (ML) is a subfield within AI, which enables computers to learn from data from which they make predictions (5). However, these models were historically constrained in their capabilities, as they struggled with processing raw data as well as requiring complex engineering and extensive expertise for the development of feature extraction (6).

Computer Vision (CV) emerged as a specialized field of AI, that allows computers to learn from images using mainly ML techniques - in other words, enabling a machine to see. CV employs deep learning (DL) - methodologies, with convolutional neural networks (CNN) being the most prevalent type (7). DL, a sub-category of ML, is a form of representation learning where the computer acquires information automatically from raw data. It uses multilayered networks to perform predictions (6,7) and has drawn remarkable interest ever since the ImageNet competition in 2012, where Krizhevsky et al. proposed a DL-method based on CNNs, effectively reducing the prediction error rate to a yet unseen level (8).

Moreover, DL-models have demonstrated promising results in various medical applications, such as in the detection of diabetic retinopathy (9), the interpretation of mammographs (10), the assessment of lung cancer (11) and recognition the of skin cancer (12). CNNs are a particular type of DL-method, well-suited to handle image data, which renders them especially interesting in medical imaging.

These models are inspired by the visual cortex and designed to automatically identify discriminating aspects without the necessity of human intervention (13).

CV-models have shown successes in tasks like image classification, object detection and segmentation (6).

With respect to osseous injury and CV these terms would be:

- **Fracture classification**
- **Fracture detection**
- **Fracture segmentation**

Fracture classification refers to the task of predicting the presence or absence of a fracture (14). Fracture detection not only identifies a fracture, but also pinpoints its location with a bounding box (15), where fracture segmentation predicts the probability of a pixel belonging to the fracture (16).

1.2 The Training Process

Among the various approaches to training ML-models, supervised learning is the most common. In this method, the algorithm is provided with a set of labeled images that contain the ground truth, meaning the correct answer. In the case of fracture detection, the goal for the algorithm is to learn to predict whether a fracture is present or not (6). The model attempts to process distinguishing features from the data and makes its prediction based on them (4). The error between the ground truth and the prediction is measured and the algorithm subsequently adjusts its modifiable internal parameters (weights) to decrease it. In CNNs, this error is commonly quantified using the loss function, from which an optimization algorithm called gradient descent calculates the gradient and aims to reduce it. This operation is repeated, ideally until satisfactory performance is reached. In essence, the model tries to identify the most suitable setting for its weights with the aim of minimizing the discrepancy between the predicted outputs and the ground truth labels (17,18). Hyperparameters are important elements of an architecture but must be manually established prior to training. They include for example the learning rates or number of layers in the network, influencing the model's performance and often demanding high degrees of expertise for effective optimization (19).

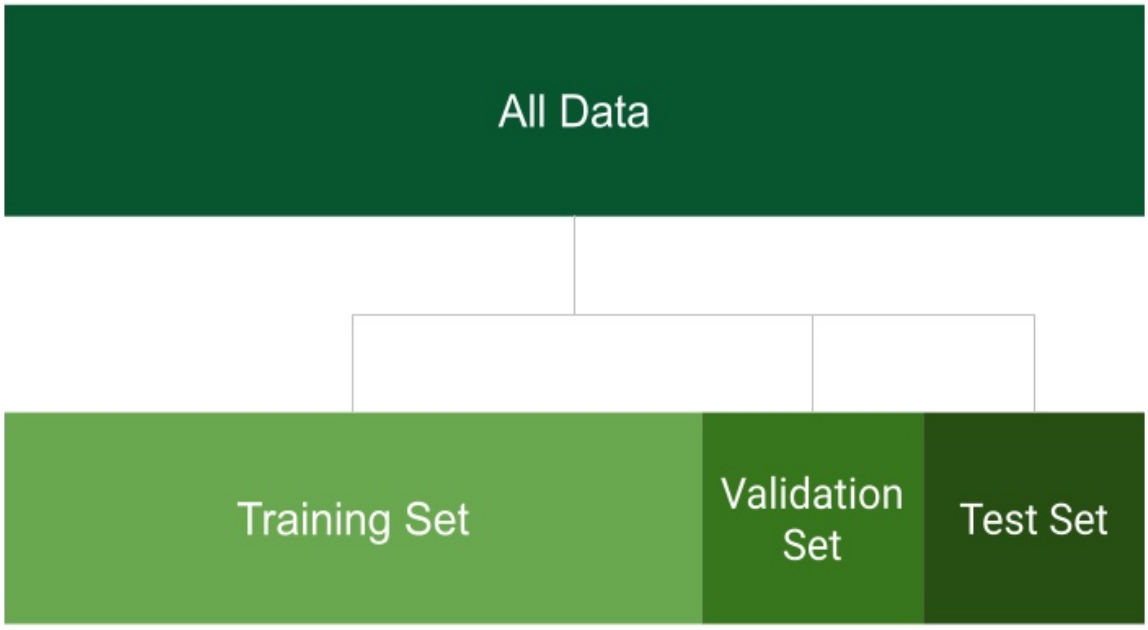

Figure 1: Representation of data separation. All relevant data is usually split up into three parts: the training, validation set and test set.
Image adapted from (20)

As illustrated in Figure 1, available data is conventionally split into three distinct data sets: the training set, the validation set and the test set. The training set makes up the largest portion of the data and is used to teach the algorithm to differentiate patterns. A separate validation set is employed to monitor performance and guide adjustments such as the fine tuning of hyperparameter during training. It is important to note that frequent validation can lead to information leakage to the model itself, thus causing it to overfit. For this reason, a test set comprising images that have never been previously seen is used following completion of the training process (17).

1.3 Convolutional Neural Networks

CNNs are composed of different building blocks called layers: the convolution layer, the pooling layer and the fully connected layer. The convolutional layer is the foundation of the CNN, basically facilitating all the computation. It consists of kernels (filters), which are essentially an array of adjustable weights usually in the size of 3x3. These glide across the image and identify relevant features, such as edges by performing convolutions. Convolutions are mathematical operations between two

matrices, one of which being the input and the other the kernel. The resulting outputs are activation maps, which are subsequently processed by the pooling layer. The pooling layer reduces the dimensionality of the data by performing down-sampling, which helps to minimize the effects of irregularities (noise) in the input information while also lowering computational demands, making networks more robust and efficient. Usually, the fully connected layer is placed at the end of a neural network architecture and performs the final prediction. The name derives from the fact, that this layer is connected to all the neurons of the previous layer (17,18).

1.4 Challenges of the Training Process

A frequent issue during training is the occurrence of overfitting. This phenomenon occurs when the model learns irrelevant details on the training set as opposed to general rules. When applied to new data, unsatisfactory results may be produced. Overfitting can be detected, by comparing the performance on the training and validation set. High performance in the training set and poor results in the validation set suggests an overfit model. Additionally, excessive use of the validation set can also cause the model to overfit to it. A common strategy to counter this is increasing the amount of training data, as DL-models generally scale well with larger datasets. However, collecting high-quality labeled data requires a considerable investment of time and financial resources, making this endeavor rather difficult in practice. For this reason, several countermeasures against this problem have been proposed: one method is data augmentation, where the input data is changed with random transformations such as rotation or skewing, without changing the integrity of the labels (4,17). Drop out is another technique that refers to the deactivation of random neurons to make the model less dependent on individual weights (21).
In contrast, a model can also be underfit if the training data is insufficient (7).
Similarly, overcoming the presence of selection bias, such as group differences regarding age or number of samples within the dataset is another challenge. Selection bias can negatively impact the performance of CV-models, making it critical to prevent this during training and even more when building the test set. Otherwise, inaccurate predictions could be the result (22).

1.5 YOLOv8 Architecture

YOLO is short for “You Only Look Once”, describing the capability of fulfilling predictions with only one pass through the network (23). It is a DL-based object detection model introduced by Joseph Redmon et al. (24), with YOLOv8 being the eighth iteration, published in 2023 by Ultralytics (25). YOLOv8 offers five different sizes: YOLOv8n (nano), YOLOv8s (small), YOLOv8m (medium), YOLOv8l (large), YOLOv8x (extra-large). Its architecture consists of three main components: a backbone, a neck and a head.

The backbone facilitates feature extraction, employing a modified version of CSPDarknet53, introducing the cross stage partial network method (26). It is based on the splitting of the feature map into two parts and merging them through a cross-stage hierarchy, reducing the computational costs, while enhancing accuracy (27). Additionally, a C2f module enables increased accuracy by combining high-level features with contextual information. The neck aggregates the feature maps utilizing the Path Aggregation Network (PANet) method allowing the recognition of details at multiple scales. Lastly, the head performs the final prediction and comprises numerous detection heads, yielding bounding boxes, class probabilities and objectness scores. The bounding box highlights the predicted area, class probability refers to the confidence of its prediction and the objectness score is the likelihood of the object being within the bounding box (26)

1.6 EfficientNet Architecture

EfficientNet is a type of CNN proposed by Tan and Le (28) and is based on the concept of compound scaling, which is essentially the combined balancing of network depth, width and resolution. Depth denotes the number of layers in a network; width refers to the amount of channels within a layer and resolution to the image size.

A major challenge in network scaling is that these parameters are dependent on each other and their optimal value varies under different resource constraints. For this reason, the typical approach so far has encompassed adjusting only one dimension, most commonly network depth. However, this method has turned out to

be quite inefficient. While adding more layers initially improves performance, beyond a certain threshold it has resulted in diminishing accuracy returns, merely increasing computational complexity without substantial accuracy improvement. Similar findings were observed when increasing width or resolution. EfficientNet addresses this issue by carefully scaling up all three values on a small baseline network (EfficientNet-B0), employing a fixed scaling coefficient.

The architecture is built upon the mobile inverted bottleneck (MBConv), which serves as the computational backbone of EfficientNet. With the application of the compound scaling strategy, larger variants have been created (B1-B7), decreasing the number of required parameters and necessary computational power while optimizing efficiency and test results (28).

2 Pediatric Fractures

Pediatric trauma represents a significant healthcare challenge. In the United States approximately 15% of emergency department visits in 2010 have been related to bone fractures (29). It is estimated that one in three children suffers from at least one fracture prior to reaching the age of 17 years. Typical incident peaks have been observed in boys at around 14 years and at 11 years in girls. The most prevalent location is the distal forearm in both genders, most frequently due to falling (30). Bone fractures can lead to immobilization, may require inpatient hospital treatment (31) and in rare cases can result in disability (32). Given the possible severe impact of bone injuries on children, maximizing efforts towards optimal diagnosis and treatment is vital.

2.1 The Pediatric Skeleton and Common Fractures

The pediatric skeleton features major differences in its constitution compared to the adult counterpart. It is thus important to keep these in mind to avoid drawing erroneous conclusions (33). Typical characteristics in children are the greater bone-forming potential and increased thickness of their periosteum, which play a critical role in providing more stability and in turn reducing displacement in the case of

fracture (34). Additionally, they demonstrate higher bone plasticity (35), superior bone healing and greater remodeling potential compared to adults (36).
One of the most notable distinctions is the presence of the growth plate. Also known as the physis or physeal plate, it is responsible for longitudinal bone growth and is located between the epiphysis and metaphysis. In larger bones, such as the humerus, they are featured on both ends while in shorter bones like the phalanges they are only found on one end (34).

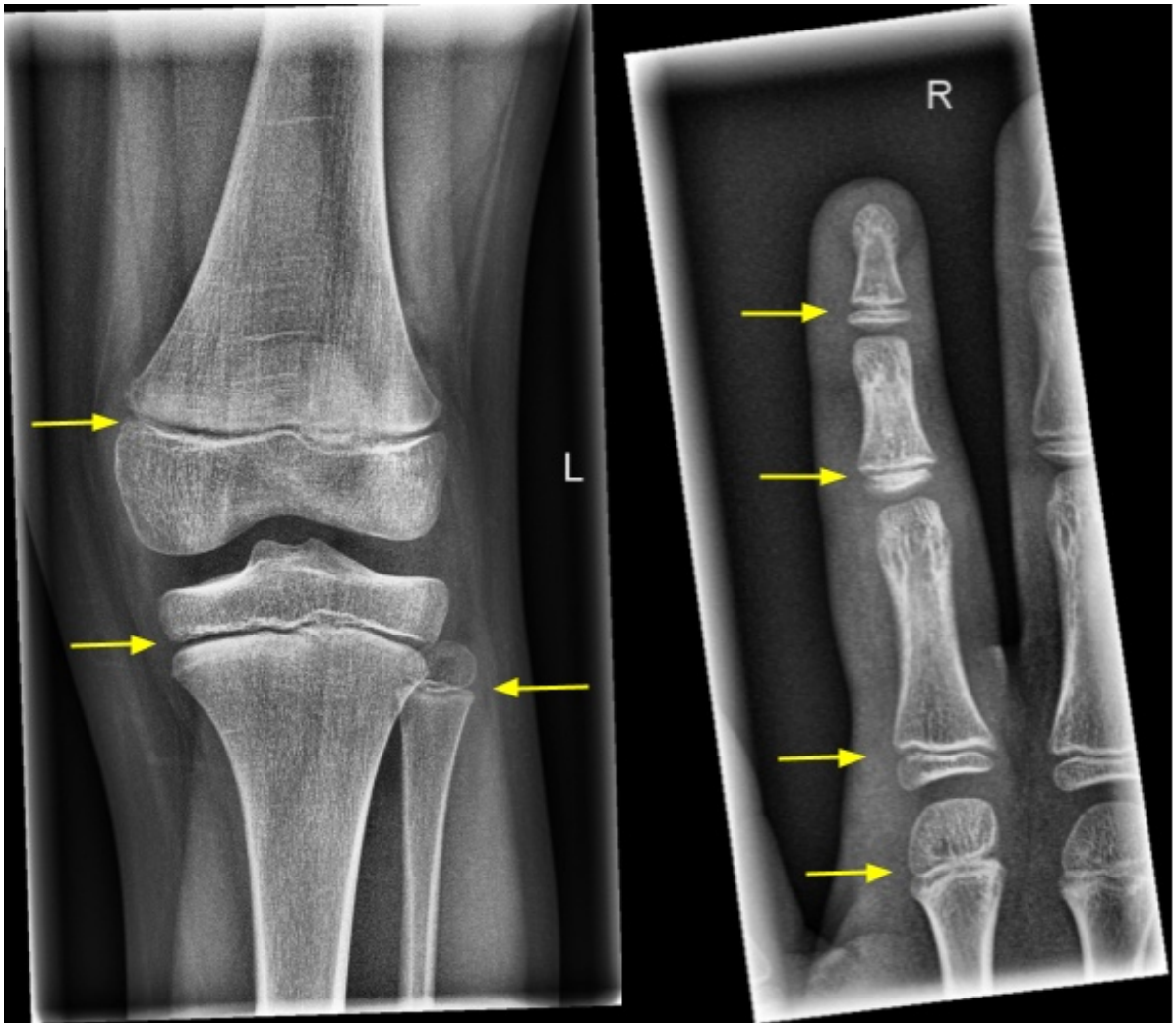

Figure 2: Sample x-rays of a knee (left) and a finger (right). The yellow arrows point to the location of the growth plates

Each growth plate contributes only a certain percentage to total growth. To illustrate this, the physis of the distal femur attributes around 70%, whereas the proximal growth plate of the same bone is only responsible for an estimated 30% of growth. The duration of physeal aperture has a direct influence on the healing process of

fractures, as growth plates with higher contribution remain open for a longer period of time and consequently have greater potential for displacement correction by means of remodeling (37). The greater remodeling potential largely decreases the need for exact anatomical reduction (34).

Described differences alter children's response to injuries, making them susceptible to fracture types that are exclusively found in their age group. Typical injuries, commonly observed in pediatric long bones include greenstick, buckle (torus), bowing (plastic) fractures and growth plate fractures (35). Resembling the breaking of a young tree branch, greenstick fractures refer to a type of incomplete fracture where the concave side of the bone remains intact while the convex side breaks. (38). A buckle fracture is characterized by a classic bulge and is commonly the result of compressive failure occurring through the metaphysis of the bone (39). Examples of the buckle and the greenstick fracture are depicted in Figure 3.

Bowing fractures are another type of incomplete fracture absent of visible cortex disruption. Instead, these fractures show a plastic deformation caused by extensive axial force (40), frequently resulting in misdiagnosis due to their subtleness (41).

Growth plate fractures are typically described with the Salter-Harris classification system and can lead to severe complications such as growth disturbances (42).

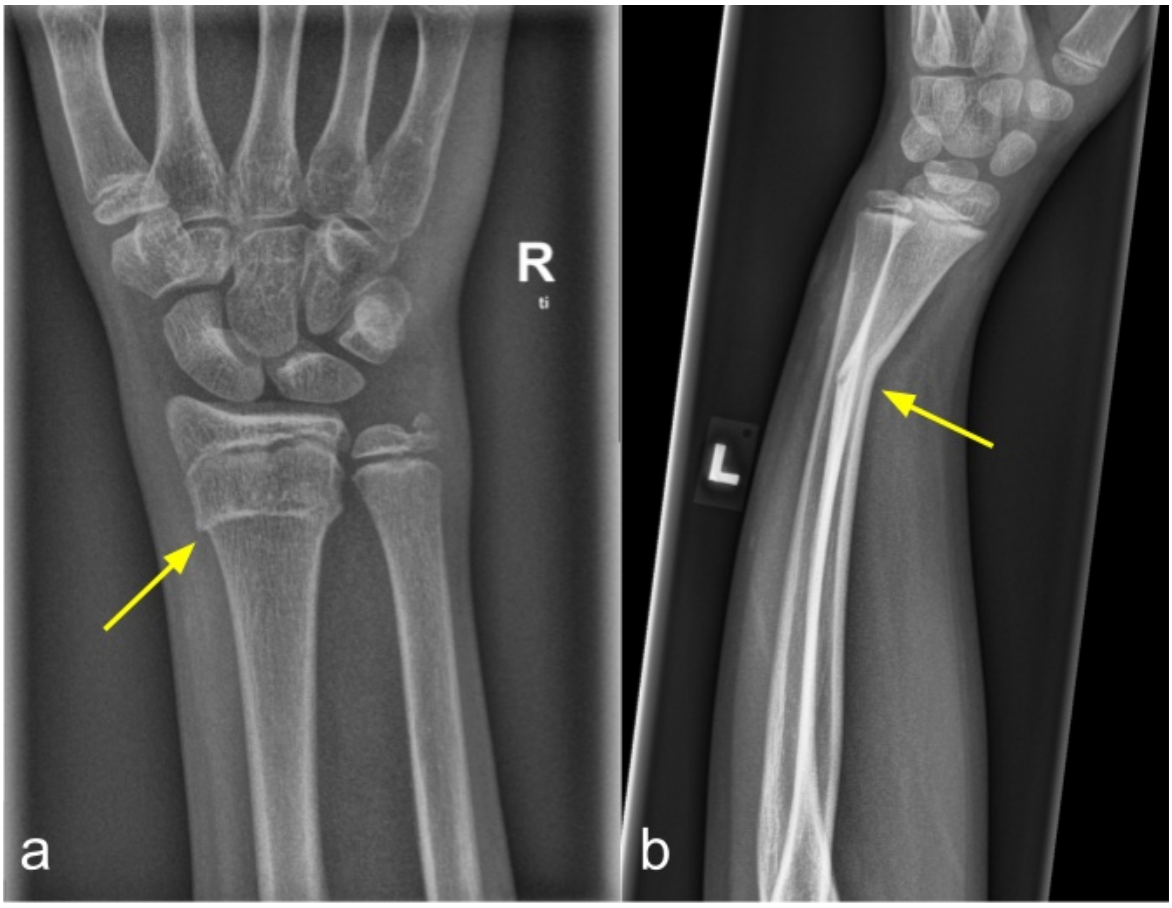

Figure 3: Examples of pediatric fractures: (a) buckle fracture and (b) greenstick fracture

Secondary ossification centers represent another distinctive feature of the developing skeleton, manifesting at specific locations and ages which vary by gender. The proper interpretation of pediatric radiographs demands a fundamental understanding of their expected appearance, thereby ensuring that physiologic development is not mistaken as a fracture (43).

Similarly, apophyseal avulsion fractures are injuries which are typically sports-related. They occur in adolescents as a consequence of sudden tensile force being exerted on the apophysis, causing a bone fragment to be torn away (44).

It is worth highlighting that not all fractures are necessarily the product of acute trauma. Underlying diseases can weaken the structural integrity of bones, which can result in pathologic fractures in response to inadequate force (45). Likewise, stress fractures are commonly observed among athletes due to increased remodeling activity reacting to repeated stress (46).

The management of pediatric trauma requires a solid grasp of the child's developmental process because, as lack of familiarity can easily lead to avoidable diagnostic pitfalls such as the abovementioned confusion of physiologic secondary ossification centers with avulsions. These unique anatomical features contribute to a fairly flat learning curve for less experienced physicians and can render the interpretation of pediatric trauma radiographs a challenging endeavor (47).

2.2 Pediatric Fracture Imaging

2.2.1 Important Considerations for Pediatric Fracture Imaging

An analysis of the effects of nuclear disasters has revealed the carcinogenic nature of Ionizing Radiation (IR). Exposure to IR in childhood has demonstrated that it heightens the risk of developing cancer. Despite the hazard, IR plays a crucial role in the medical setting. It has various useful applications, both diagnostically, through imaging modalities such as radiography or computed tomography (CT) and therapeutically in treatments like radiotherapy. However, concerns have been raised regarding the long-term outcomes of diagnostic and therapeutic exposure to IR (48). In the United States, as of the year 2006, an estimated 62 million CT scans have been conducted, with approximately 10% of these on children (49). In the light of potential health risks associated with IR usage in medical imaging, it is important to perform scans only when necessary, following the principle of as low as reasonably achievable (ALARA) (50).

2.2.2 Diagnostics

The common procedure of acute fracture assessment normally encompasses an effective anamnesis, thorough clinical examination and a plain radiograph in one or more projections (37). With this strategy, a diagnostic error rate of around 5% is reported (51). Still, misdiagnosis in the emergency department remains a prevalent concern. Errors are often caused by the subtleness of certain types of fractures, such as chip fractures in the phalanges. In addition, particular fractures manifest solely with indirect signs, i.e. the posterior fat pad sign (33). Consequences may

include wrong, unnecessary or delayed treatment (52), not to mention possible medicolegal concerns (53).

2.2.2.1 Ultrasound

Ultrasound has gained more attention as a diagnostic tool in the evaluation of pediatric fractures. It has demonstrated satisfactory performance in pediatric trauma, such as in the detection of clavicle and skull fractures (54,55) and may be a suitable, cost-effective and non-invasive alternative for the initial trauma x-ray. Nevertheless, its limitations include reduced efficacy near joints and difficulties with non-displaced epiphyseal fractures. Still, the diagnostic capability in the diaphyseal area of long bones has proven to be sufficient. (56,57). Given the number of negative findings of primary radiographic examinations, fears of unnecessary radiation are justified (58). In selected pediatric trauma cases, replacing the initial x-ray in favor for ultrasound could contribute to improved radiation protection (56,57). Additionally, ultrasound in the form of low-intensity pulsed ultrasound has proven its efficacy in assisting the healing of bone fractures and non-unions (59).

2.2.2.2 Computed Tomography

CT offers a high degree of diagnostic capability and quickly delivers detailed and consistent information about all organ systems in pediatric patients (60). CT exams are primarily utilized for the evaluation of injuries to the trunk but also in instances of complex tarsal or carpal fractures. CT scans can be a useful way to visualize fractures, determine the extent of dislocation and develop a treatment plan (37). While the advantages outweigh potential risks, a higher dose of IR is utilized compared to standard radiography (60). Given the increased vulnerability of children to the adverse effects of radiation, this modality should only employed with clear indications (61).

2.2.2.3 Magnetic Resonance Imaging

The use of magnetic resonance imaging (MRI) has demonstrated satisfactory results in the detection of fractures (62–64). It demonstrates excellent visualization of bone marrow, joint effusion, soft tissues such as tendons, periosteum, ligaments as well as edema and fracture lines (65). Nevertheless, its use in trauma diagnostics remains rather limited due to reduced availability and the frequently required sedation of younger children for successful imaging (37).

2.3 Treatment

Most pediatric fractures can be addressed non-operatively with cast immobilization. The main objectives are to provide stability, safeguard the fractured site, prevent further harm and most importantly, ensure alignment until healing is achieved. Treatment length differs according to the affected site and age of the patient. However, potential complications may occur from casting. I.e. pressure sores, resulting from an excessively tight cast or dermatitis from the presence of wetness beneath it. Furthermore, it is important to keep the immobilization period to a minimum, since longer duration may cause problems like joint stiffness, muscle atrophy or ligament weakening. Likewise, swelling after the application of a cast may lead to compartment syndrome. In such cases, interventions like splitting the cast and, if unsuccessful, complete cast removal should be considered (34).

While uncommon, certain types of fractures necessitate surgical attention. The decision to operate, whether through open or closed reduction, is dependent on the type of injury. In cases where the affected bone shows complete dislocation of either the dia- or metaphysis and is unable to be realigned through conservative means, closed reduction is indicated. Methods to facilitate this include the use of percutaneous K-Wire, external fixators or elastic stable intramedullary nailing (ESIN). In regard to open reductions, the main indications include dislocated articular fractures, dislocations accompanied by osseous or ligamentous injuries, open fractures, fractures involving nerve and vessel damage as well as other types of fractures which cannot be addressed by closed reduction (37).

2.4 Complications

Most fractures in children heal without long-term consequences largely due to their remarkable bone healing and remodeling potential (36), however, some fracture types can have severe consequences for the injured child.

The physis, as a structural weak point in the pediatric bone is particularly prone to fractures. Injury to the open growth plate has the potential to induce growth disturbances such as limb length discrepancies and angular deformities, even possible premature growth arrest (66). Moreover, open fractures are especially susceptible to infections, possibly leading to the grave consequence of osteomyelitis (67). Osseous non-unions, while rare, occur if fracture fragments fail to reconnect in the expected time of the healing process, specifically in areas with limited vascular supply such as the scaphoid (68). Malunions are the result of fracture healing with deformity and special attention must be given to rotational deformities, as their remodeling potential is unsatisfactory and may require surgical fixation (37)

3 Hypothesis

The detection of fractures in children poses a unique set of challenges, largely attributable to anatomical differences and the subtle nature of certain fractures. Specialized pediatric radiologists may not always be immediately available, such as during night shifts or in rural areas, and errors in diagnosis can have severe consequences for example inadequate or delayed treatment.

AI-based CV-models have demonstrated promising results in automated fracture detection. However, their performance may vary depending on the composition of the underlying test set. We assume that the AI-algorithms struggle with classification and object detection of more complex cases and demonstrate lower overall diagnostic performance when using test sets comprising difficult cases in comparison to a matched set.

4 Aims and Scope of the Study

The main objective of this study is to evaluate the diagnostic performance of two CV-models, namely YOLOv8 and EfficientNet. Each model, along with its subvariants YOLOv8-n YOLOv8-s, YOLOv8-m, YOLOv8-l, YOLOv8-x and EfficientNet B0-B7 undergo training and validation. Testing is conducted on two separate test sets: one consisting of difficult cases and another matched set designed to evaluate fracture classification and detection performance in pediatric trauma cases of the appendicular skeleton. Using standard machine learning metrics such as Precision, Recall, Accuracy and F1-score, we conduct a comparative analysis of the two test sets to provide an understanding of the response of the AI-models to varying levels of case complexity. In addition, we utilize Precision-Recall-Curves with their respective area under the curve (PR AUC), to graphically represent the results.

5 Materials and Methods

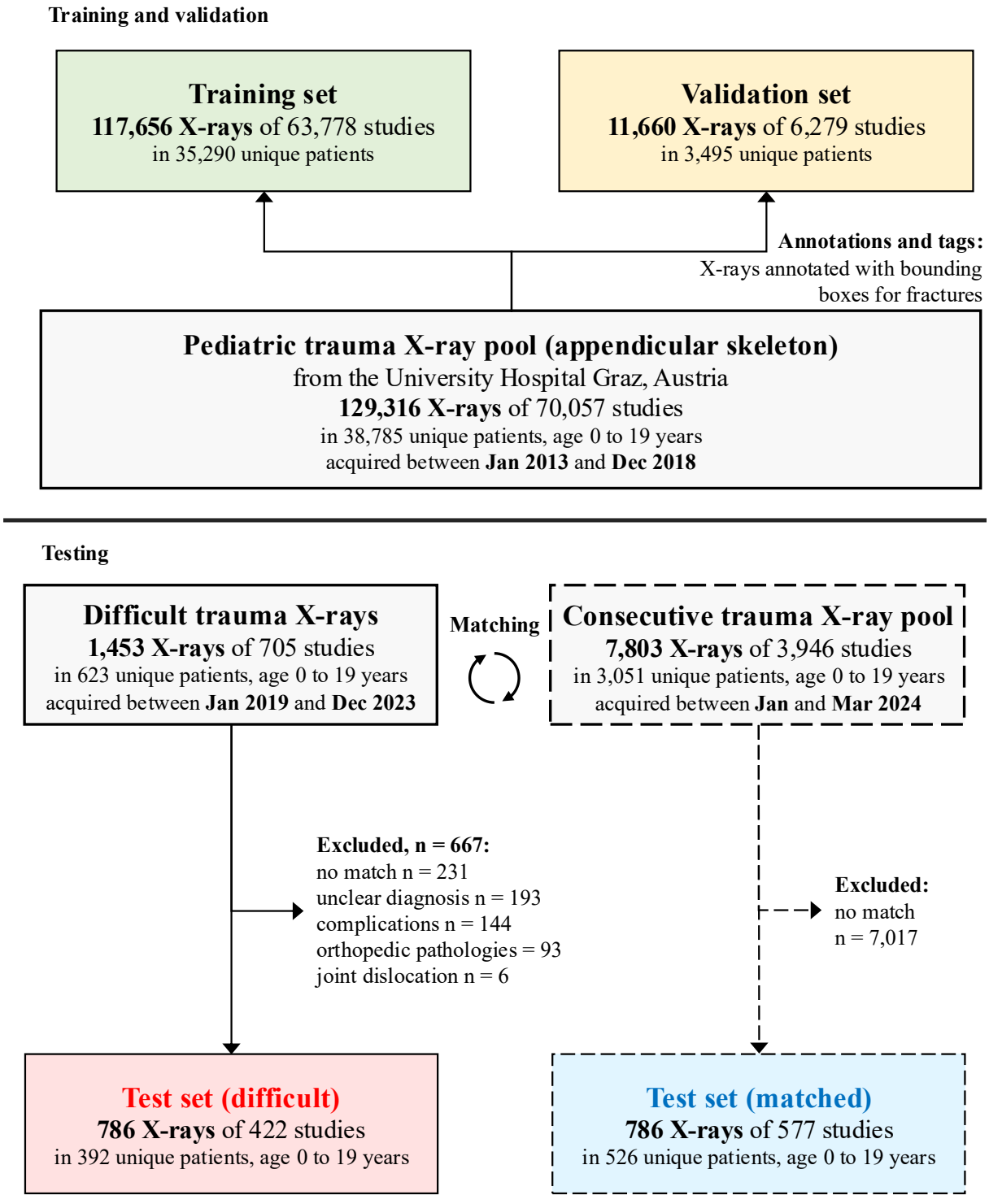

Figure 4: Flowchart depicting the allocation of each radiograph to the training, validation and both test sets

5.1 Training and Validation Set

This study employed a total of 129,316 radiographs from the pediatric trauma x-ray pool of the appendicular skeleton from the University Hospital Graz, Austria. The images were obtained between January 2013 and December 2018 and include data from 38,785 unique patients. Of these, 117,656 x-rays (~91%) of 35,290 patients were allocated to the training set. 11,660 radiographs (~9%) of 3,495 patients were separated for the validation set. These images were annotated with bounding boxes for fractures as seen in Figure 5. The selection process of the different sets is depicted in Figure 4.

5.2 Test Sets

Outside regular working hours of pediatric radiologists, initial evaluation of x-rays is performed by the Department of Pediatric and Adolescent Surgery. On the following workday, these images are reassessed by physicians from the Division of Pediatric Radiology. Occasionally, discrepancies arise between the findings of the two departments. Such discordant cases are kept on an internal list and are classified as “difficult”.
In the period from 2019 to 2023, a total number of 1,453 trauma radiographs of 623 patients with conflicting findings were identified and included.
These studies were reviewed again and with the aid of radiological reports and clinical information, divided into different groups:

- “missed”
- “overdiagnosed”
- “complications”
- “joint dislocation”
- “orthopedic pathologies”
- “unclear“

In instances where a radiograph was marked as “missed”, the fracture was overseen in the initial trauma report. The tag “overdiagnosed” was designated to images in which a fracture was primarily diagnosed but not identified in the subsequent review. The label “complication” was assigned to images showing trauma-related complications, such as secondary displacement. “joint dislocation” was used due to present dislocations of joints. The tag “orthopedic pathologies” was allocated if there was no fracture present, but pathologies like bone lesions occurred. Finally, cases that could not be reliably reconstructed were labeled as “unclear”. Only the cases tagged “overdiagnosed” and “missed” were included in the composition of the difficult test set. X-rays marked with other tags as well as cases without matching radiographs were excluded, totaling 667 images. Consequently, 786 x-rays of 392 unique patients were chosen for the difficult test set.

Out of 7,803 radiographs, obtained from the consecutive trauma x-ray pool of the appendicular skeleton in the period between January and March 2024, 786 x-rays of 526 patients were included for the matching process. Thus, 7,017 were not suitable for matching and therefore excluded.

Figure 5: Sample images of annotated fractures

5.3 Ethics approval

A vote of the ethics committee of the Medical University of Graz was requested. It was not necessary to obtain individual informed consent from the patients and their legal guardians. The reference numbers of the votes were “31-108 ex 18/19” and “1052/2025”. Furthermore, this study was conducted in accordance with the Declaration of Helsinki.

5.4 Image acquisition

The radiographic images utilized in this retrospective study were obtained from the Department of Pediatric Radiology at the Medical University of Graz, Austria. Those images were stored locally as Digital Imaging and Communications in Medicine (DICOM) in the Picture Archiving and Communication System (PACS).

5.5 Image processing and Neural Network Training

To prepare the DICOM studies for the training process, the data was arranged as follows: the radiographs used to conduct this study were stored within the local PACS. Upon retrieval, they were converted to portable network graphics (PNG) format. The DICOMs exhibited grayscale values of either 12-bit or 16-bit and varying dimensions. Next, the data was standardized to 16-bit PNGs with the assistance of the “pydicom” and “cv2” packages in Python, while maintaining the initial pixel dimensions. Through the “exposure” module of the Python “scikit-image” – package, the radiographs were subjected to post-processing. This involved transforming the 16-bit inputs to float64. In this process, the greyscales were divided by 65,535. Subsequently, the upper and lower 0,05% percentiles of the image histograms were cropped, resulting in a rescaling of the intensity. Moreover, the “exposure adapthist” was employed to enhance contrasts with the standard settings. The acquired float64 images were converted to 8-bit outputs and saved. (69)

For the neural network training, a Linux workstation with an Intel Core i7 13700KF processor and 64GB of random access memory, two Nvidia Geforce RTX 4090

graphics cards, both with a video memory capacity of 24 GB were used. All submodels of YOLOv8 and EfficientNet were trained with 50 epochs utilizing pre-trained weights. In this study PyTorch 2.4.0 with Nvidia CUDA 12.1 was employed and batch sizes were adjusted to fit the training data within the GPU memory.

5.6 Test parameters

For the comparison of the different models, standard machine learning test-metrics were employed:

$$Accuracy = \frac{TP*TN}{TP+FP+FN+TN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1\text{-}score = 2\ \frac{(Precision*Recall)}{(Precision+Recall)}$$

A true positive (TP) is defined as the correct prediction of a positive outcome, while a true negative (TN) is described as the correct prediction of a negative result. A false positive (FP) occurs when a positive result is incorrectly predicted, whereas a false negative (FN) is characterized as the wrong prediction of a negative outcome. Accuracy indicates the number of correct predictions of all samples. Precision expresses the number of true positives of all positively classified predictions. Recall measures the ability to correctly recognize the actual positives. The F1-score is the harmonic mean of Precision and Recall (70).

5.7 Statistics

Statistical analysis was conducted using IBM SPSS Statistics (version 28) (IBM Corp., Armonk, N.Y., USA)

5.8 Data anonymization

For the training of the computer algorithms, the x-ray images are exported and subsequently anonymized. The radiographs themselves do not display any identifiable patient information. These are stored in the metadata of the DICOM files. The new file name is formed from the following metadata values:

- SHA-3(256) hash National Institute of Standards and Technology (NIST), 2015 of the patient ID
- Unix-timestamp of the time of examination – (minus) 9 integer hash of the patient ID (to anonymize the time of the examination without changing the chronological order of the examinations
- Serial number
- Region code
- Gender
- Age (rounded to 1 decimal)

For example, the file name may be expressed as follows:

“0a168d681a40aac6f04c6aeb07703cc084e323f8813f525cb1f0f20aa80ecc0b_060 6582086_01-01_TOE-N0_F014-4.png“

The following values are stored in an encrypted database on a password-protected computer at the Division of Pediatric Radiology:

- The file name as previously described
- Age in years
- Sex

- Body region
- SHA – 3 (256) hash of the patient number (patient ID)
- SHA – 3 (512) hash of the accession number
- SHA – 3 (512) hash of the SOP Class UID of the image
- Serial number
- Acquisition number
- Image dimensions, pixel sizes and gray value range
- Radiological findings
- Referral diagnosis
- Findings text
- Question
- Traumatological findings:
    - X-ray findings
    - Diagnostic text
- Diagnostic text

Collected values:

- Fractures, pathologies, marked structures
- Image classifiers

The aforementioned measures render de-anonymization of personal information a highly impractical endeavor. It would require a disproportionate amount of time, resources and personnel.

6 Results

Body region	Number of cases		Mean age		Fracture rate	
	difficult	*matched*	*difficult*	*matched*	*difficult*	*matched*
Ankle	103	103	11.48	11.54	16.5%	16.5%
Clavicle	8	8	7.30	7.06	62.5%	62.5%
Elbow	64	64	7.85	8.18	60.9%	60.9%
Finger	169	169	10.67	10.63	60.4%	60.4%
Foot	61	61	10.11	10.08	41.0%	41.0%
Forearm	34	34	7.12	7.23	64.7%	64.7%
Hand	31	31	12.78	12.92	41.9%	41.9%
Knee	65	65	12.90	13.02	12.3%	12.3%
Lower leg	13	13	7.98	7.98	38.5%	38.5%
Pelvis	6	6	11.85	12.15	16.7%	16.7%
Shoulder	7	7	14.97	14.11	14.3%	14.3%
Thigh	3	3	5.17	5.90	0.0%	0.0%
Toe	82	82	11.38	11.22	51.2%	51.2%
Upper arm	6	6	12.73	11.82	33.3%	33.3%
Wrist	134	134	11.91	11.92	46.3%	46.3%
Total	**786**	**786**	**10.87**	**10.88**	**43.8%**	**43.8%**

Table 1: Representation of the number of body regions including number of cases, mean age and fracture rate. Every region has been matched adequately.

6.1 Results EfficientNet

Model	Precision		Recall		Accuracy		F1-score		PR AUC	
	diff.	*mat.*	*diff.*	*mat.*	*diff.*	*mat.*	*diff.*	*mat.*	*diff.*	*mat.*
B0	0,643	0,676	0,609	0,674	0,639	0,682	0,600	0,675	0,631	0,728
B1	0,653	0,714	0,613	0,713	0,644	0,719	0,602	0,714	0,653	0,747
B2	0,674	0,714	0,608	0,708	0,645	0,718	0,587	0,709	0,673	0,761
B3	0,665	0,752	0,618	0,750	0,650	0,756	0,606	0,751	0,665	0,809
B4	0,689	0,813	0,644	0,798	0,673	0,809	0,638	0,802	0,690	0,860
B5	0,695	0,830	0,640	0,815	0,672	0,826	0,631	0,820	0,705	0,899
B6	0,744	0,837	0,655	0,826	0,691	0,835	0,643	0,830	0,745	0,896
B7	0,703	0,793	0,631	0,750	0,667	0,771	0,616	0,754	0,697	0,829
Mean	**0,683**	**0,766**	**0,627**	**0,754**	**0,660**	**0,764**	**0,615**	**0,757**	**0,682**	**0,816**
SD	**± 0,030**	**± 0,057**	**± 0,017**	**± 0,051**	**± 0,017**	**± 0,052**	**± 0,019**	**± 0,053**	**± 0,033**	**± 0,062**
	P= 0,012		**P= 0,012**		**P= 0,012**		**P= 0,012**		**P= 0,012**	

Table 2: Results of each EfficientNet-variant. The results were statistically significant ($p < 0,05$) with the Wilcoxon test. Note: diff. = difficult, mat. = matched.

When comparing diagnostic performance on the difficult test set versus the matched test set, the mean Precision was 0,683 (± 0,030) vs. 0,766 (± 0,057, average Recall 0,627 (± 0,017) vs. 0,754 (± 0,051), mean Accuracy 0,660 (± 0,017) and 0,764 (± 0,052), respectively. The average F1-score was 0,615 (± 0,019) in the challenging set vs. 0,757 (± 0,053) in the matched test set.

Analysis of the PR AUC values further highlight this performance gap with a mean PR AUC 0,682 (± 0,033) vs. 0,816 (± 0,062). Across models, PR AUC values generally improved with increasing model size on the difficult test set, except for the drop in EfficientNet-B7 (PR AUC 0,697). Each EfficientNet-model has consistently yielded lower results in the classification task on the difficult set in comparison to the matched test set. Figure 6 shows the corresponding PR AUC curves. The results of the comparison were statistically significant ($p < 0,05$).

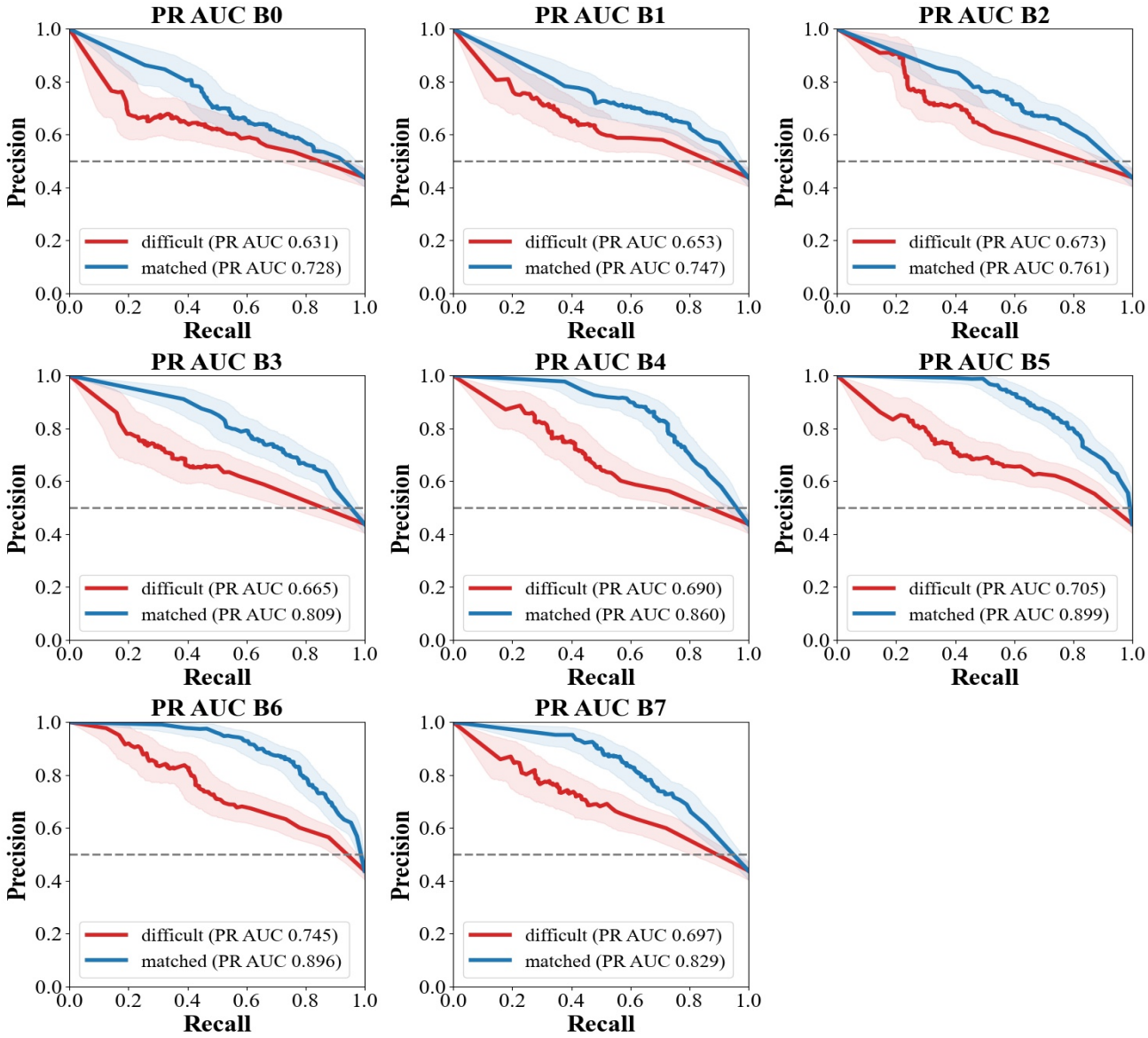

Figure 6: Representation of each EfficientNet model's PR AUC curve with the 95% confidence intervals. The difficult test set in red, and the matched test set in blue. The confidence intervals are marked as the highlighted area. The results were statistically significant with p = 0,012.

Body region	Samples	Precision		Recall		Accuracy		F1-Score	
		Testset		*Testset*		*Testset*		*Testset*	
	n	*difficult*	*matched*	*difficult*	*matched*	*difficult*	*matched*	*difficult*	*matched*
Ankle	103	0.639	0.870	0.502	0.719	0.824	0.891	0.468	0.761
Clavicle	8	0.546	0.714	0.504	0.642	0.547	0.688	0.497	0.644
Elbow	64	0.639	0.872	0.633	0.881	0.596	0.873	0.595	0.870
Finger	169	0.614	0.631	0.574	0.631	0.507	0.606	0.485	0.605
Foot	61	0.559	0.705	0.544	0.703	0.586	0.697	0.532	0.693
Forearm	34	0.565	0.685	0.568	0.684	0.533	0.695	0.529	0.667
Hand	31	0.753	0.715	0.686	0.715	0.722	0.710	0.685	0.705
Knee	65	0.815	0.810	0.601	0.693	0.888	0.898	0.624	0.726
Lower leg	13	0.628	0.646	0.564	0.644	0.625	0.625	0.542	0.618
Pelvis	6	0.573	0.647	0.488	0.413	0.729	0.688	0.461	0.400
Shoulder	7	0.410	0.815	0.385	0.813	0.661	0.857	0.397	0.716
Thigh	3	0.813	0.625	0.938	0.708	0.875	0.417	0.775	0.375
Toe	82	0.556	0.667	0.537	0.663	0.530	0.663	0.499	0.661
Upper arm	6	0.783	0.683	0.719	0.625	0.667	0.583	0.627	0.550
Wrist	134	0.807	0.910	0.767	0.903	0.779	0.906	0.768	0.904
Total	**786**	**0.647**	**0.733**	**0.601**	**0.696**	**0.671**	**0.720**	**0.566**	**0.660**
		±0.117	**±0.093**	**±0.130**	**±0.112**	**±0.123**	**±0.136**	**±0.108**	**±0.139**
Significance (Wilcoxon test)		*p = 0.036**		*p = 0.031**		*p = 0.074*		*p = 0.027**	

*Table 3: Representation of the results of the different test metrics of each region of EfficientNet: The numbers highlighted with * denote the significance calculated with the Wilcoxon test.*

6.2 Results YOLOv8

Across all models, the performance metrics were consistently higher on the matched test set compared to the difficult set. Table 4 summarizes these results of the YOLOv8 variants on the object detection task. On average, Precision was 0,724 (± 0,016) on the difficult set vs. 0,928 (± 0,018) on the matched set. Mean Recall was 0,467 (± 0,022) vs 0,858 (± 0,026), mean mAP50 0,553 (± 0,022) vs. 0,932 (± 0,021), mean mAP50-95 0,253 (± 0,006) vs. 0,709 (± 0,024) and mean F1-score 0,568 (± 0,017) vs. 0,891 (± 0,022) on the difficult compared to the matched set, respectively. Also, the PR AUC curves as seen in Figure 7 show the inferior performance across all YOLOv8-variants when tested on complex cases with a mean PR AUC 0,735 (± 0,024) vs. 0,929 (± 0,029). All results were statistically significant ($p < 0,05$).

YOLO variant	Test set	Precision	Recall	mAP50	mAP50-95	F1-score	PR AUC
YOLOv8-n	Difficult	0,715	0,430	0,514	0,242	0,537	0,689
YOLOv8-s	Difficult	0,726	0,457	0,565	0,256	0,561	0,759
YOLOv8-m	Difficult	0,755	0,476	0,566	0,253	0,584	0,741
YOLOv8-l	Difficult	0,713	0,481	0,566	0,254	0,574	0,750
YOLOv8-x	Difficult	0,711	0,492	0,579	0,261	0,582	0,736
Mean SD		**0,724 ± 0,016**	**0,467 ± 0,022**	**0,553 ± 0,022**	**0,253 ± 0,006**	**0,568 ± 0,017**	**0,735 ± 0,024**
YOLOv8-n	Matched	0,900	0,812	0,893	0,666	0,854	0,873
YOLOv8-s	Matched	0,916	0,853	0,927	0,699	0,883	0,926
YOLOv8-m	Matched	0,949	0,884	0,941	0,727	0,915	0,949
YOLOv8-l	Matched	0,934	0,856	0,942	0,720	0,893	0,947
YOLOv8-x	Matched	0,942	0,884	0,955	0,732	0,912	0,950
Mean SD		**0,928 ± 0,018**	**0,858 ± 0,026**	**0,932 ± 0,021**	**0,709 ± 0,024**	**0,891 ± 0,022**	**0,929 ± 0,029**
Significance (Wilcoxon test)		**P= 0,043**	**P= 0,043**	**P= 0,042**	**P= 0,043**	**P= 0,043**	**P= 0,043**

Table 4: Representation of the different results across each YOLOv8 variant on the difficult and the matched test set

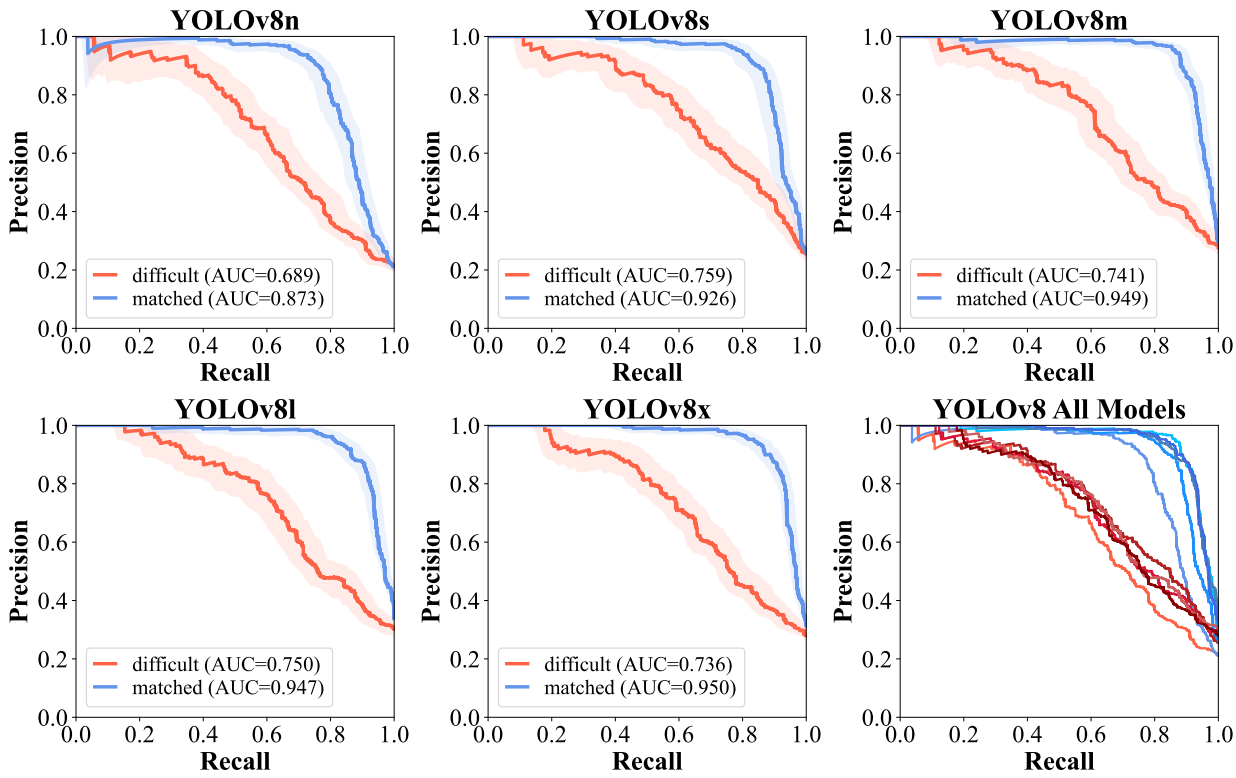

Figure 7: PR curves of all YOLOv8 variants.

7 Discussion

7.1 Summary of results

This study evaluated the diagnostic performance of two AI-models: EfficientNet and YOLOv8 with their subvariants using two distinct test sets: one set consisting of challenging cases from pediatric trauma routine and another one composed of cases that have been matched for mean age, body region and fracture rate. Both models have consistently yielded inferior diagnostic performance in the difficult test set when compared to the matched set. As demonstrated in Table 2 and 4, mean F1-scores for EfficientNet were 0,615 (± 0,019) in the difficult set vs. 0,757 (± 0,053) in the matched test set and for YOLOv8 mean F1-scores were 0,568 (± 0,017) vs. 0,891 (± 0,022), respectively. These results highlight the critical impact the underlying test set composition has on the actual performance of DL-models.

With this study, we underscore a critical issue in the field of computer vision: the overall robustness of AI-algorithms. Challenging cases occur less commonly in the clinical routine and may therefore be underrepresented in the training data, potentially influencing the performance of AI-models. Nonetheless, due to the large size of the data set in this study and the carefully constructed test sets, this effect is largely mitigated. Still, our findings suggest that AI-models struggle with increasing difficulty and subtleness of cases, exactly where AI-assistance would be most beneficial. It appears that standalone evaluation of pediatric fractures through AI-algorithms is still too risky and necessitates further investigation. Next to the evaluation of the influence of case difficulty on standardized performance metrics, a key novelty of this work was the inclusion of false positive “overdiagnosed” cases, creating a test environment that closely mirrors real-world clinical conditions.
As Table 1 depicts, the test sets have been adequately matched. Nonetheless, some body regions are numerically more represented than others, such as the wrist (n=134) compared to the thigh (n=3). This may lead to skewed results when investigating region specific performance (i.e. F1-score of the thigh of 0,775 vs. 0,375 in the difficult vs. the matched test set). This underlines the data acquisition problem within ML research, as gathering samples of such relatively rare fractures would take an unfeasible amount of time.

Recently, the number of findings of exceptional diagnostic accuracy of AI-models have surged in fracture detection (71). Previous works also reported high overall performance of AI-algorithms in fracture detection in pediatric patients (72–74) Others have investigated AI-aided fracture detection like Lindsey et al. who demonstrated increased diagnostic accuracy and reduction of misdiagnosis of emergency physicians without subspecialized orthopedic expertise (75). Nguyen et al. reported a mean increase of 10% in sensitivity of human experts in the detection of appendicular pediatric skeletal fractures (76). DL-models also have shown on par diagnostic performance to that of human experts (77). However, there is critique that the underlying data sets are not selected entirely without bias, possibly skewing test results to a more favorable light. Often, challenging cases are not given enough consideration and are subsequently underrepresented, producing non-realistic results. Raisuddin et al. highlighted this in their study, leveraging wrist fractures that

required further CT examination as ground truth for their difficult test set, similarly concluding significantly lower performance on the challenging test set in comparison to a general independent test set (78). However, using CT as ground truth for the difficult test set comes with the problem that the number of samples is relatively small (n=105 cases), thus possibly leading to bias.

7.2 Limitations

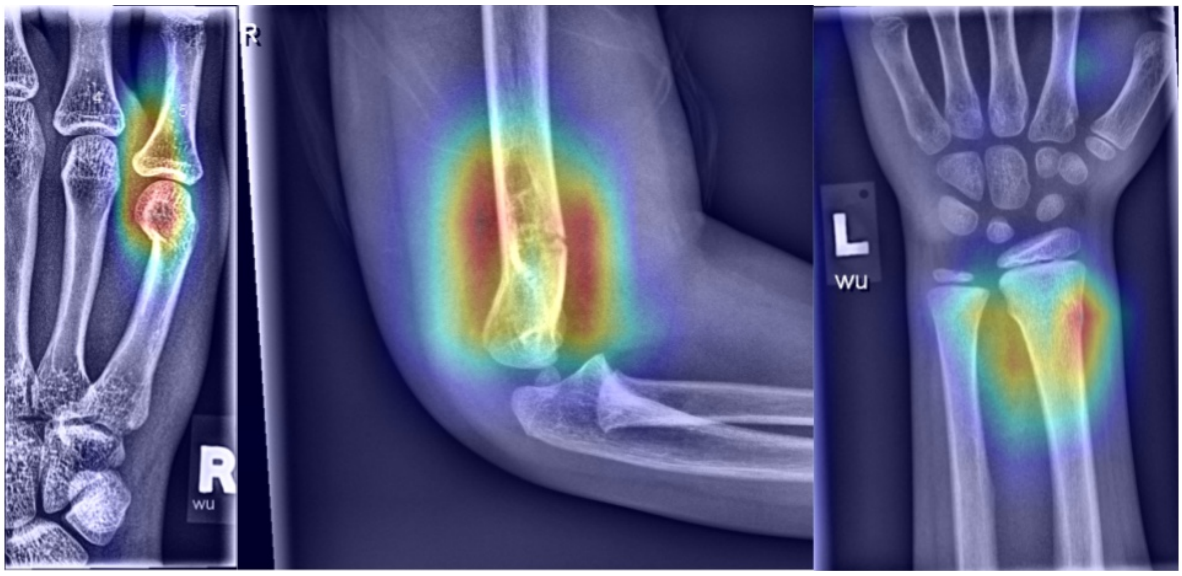

Figure 8: Grad-CAM examples highlighting the region leading to the prediction

This study has several limitations. Firstly, the lack of explainability in the realm of AI remains a major topic. DL-models are often critiqued for being opaque in their decision making due to their high complexity, for which they are commonly referred to as a "blackbox". Simpler models such as decision support trees would be more transparent but yield lower predictive results. For AI-solutions to be implemented into medical practice, the question of explainability must be addressed to ensure trust and understandable decision-making (79). One widely used technique for enhancing explainability of CNN-based models is Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM offers a visual explanation by highlighting the area that has led to the model's prediction through heatmaps (80). An example can be seen in Figure 8.

Another limitation is the presence of irregularities in the training data. The annotation process relied on expert knowledge and radiological reports, which are subject to

human error. As a possible solution to yield highly qualitative labels, multiple independent experts could be leveraged to annotate the data. However, this would be difficult to facilitate due to the substantial investment of time and financial resources needed for this endeavor. Likewise, more advanced imaging studies like CT could be used to serve as ground truth. Although, this would drastically decrease the available amount of data for model development, since not every fracture is assessed with further imaging and potentially lead to bias.

Additionally, all data has been derived from a single pediatric hospital, which in turn diminishes the generalizability due to regional differences of the population sustaining fractures. Hence, future work should focus on acquiring data from multiple centers to make algorithms more robust against selection bias regarding factors such as gender, sex and age. Moreover, it is uncertain if these models are applicable to other age groups, since training was conducted only on the pediatric population.

7.3 Relevance for Clinical Practice

With the increasing demand in medical imaging (81) and yet an ongoing shortage of radiologists (82), advances in AI-research in radiology is a subject of high interest. Implementation into clinical practice becomes an increasingly prevalent topic as the use of ML-models has major benefits. Machines neither suffer from fatigue nor sickness and are not constrained by lack of concentration and produce consistent results. The implementation of AI-solutions into clinical practice can aid the scarcity of specialized pediatric radiologists particularly in rural regions without round the clock coverage (83) or help physicians lacking subspecialized knowledge and experience. A well rounded and robust CV-model could support clinicians in the detection of subtle fractures, reducing the number of errors occurring in everyday clinical practice (33).

False negative cases, where a fracture is present but undetected, are a considerable risk to patient safety, as these could result in severe complications such as growth disturbances (66) and sometimes even disability (32). False positives “overdiagnosed” are cases in which a fracture is initially diagnosed but later ruled out upon further review. Usually, the consequences are less severe but

still can cause unnecessary interventions, higher costs and discomfort to the young patient (51).
Furthermore, implementation of AI-solutions into clinical practice could severely contribute to IR protection by reducing the need for lateral projections in radiographs in minor injuries (69).
Other applications of DL-models could be in the role of a trainer, supporting the education of junior doctors.

7.4 Future work

Undoubtedly, AI is expected to have a considerable impact on medicine with the field of radiology being at the forefront. Radiologists may be freed up from repetitive and draining tasks and can shift their attention to more complex cases. AI could take a supporting role and augment the accuracy of physicians, where patients may profit from diagnostics and treatment tailored to their specific needs.
This thesis has demonstrated the influence of case complexity on DL-algorithms measured on common performance metrics, emphasizing the necessity for further research in this area prior to safe integration into clinical practice. Future training data sets should deliberately implement a higher number of complex cases such as subtle injuries including common “overdiagnosed” fracture sites for better generalization. The results of this study suggest that AI is not the replacement of radiologists but rather serves as a tool to augment specialists and minimize possible misdiagnosis.
Furthermore, there is a necessity for a clear regulatory framework to tackle ethical, legal and data privacy issues associated with AI in radiology. Building trust in AI-solutions requires both clinicians and patients to understand how these systems operate. In this regard, the blackbox-nature of DL-models poses a significant challenge as it limits the capability to understand, explain and correct errors. Ensuring transparency is therefore critical to prevent erroneous results in the future. In addition, questions revolving around liability in cases of misdiagnosis remains unsolved and have yet to be addressed.
Since this study was based on plain radiographs for the detection of pediatric fractures, future work could focus on the combining of clinical context, further

imaging and combining x-ray views to make a more accurate and robust prediction. Moreover, further exploration of explainability tools could be beneficial to enhance trust in AI-solutions. Additionally, there is a need for external validation of AI-algorithms to safeguard trust and demonstrate the ability to generalize on different patient demographics, bringing implementation into clinical practice one step closer to reality.

7.5 Conclusion

In this work, we showed inferior predictive performance of two AI-models in a difficult test set in comparison to a matched set in fracture classification and detection. The results underscore the need for further investigation regarding integrating difficult cases in model development to ensure robustness prior to safe clinical implementation.

8 References

1. Stranger N, Scherkl M, Stütz D, Janisch M, Singer G, Hankel S, et al. Impact of Test Set Composition on AI Performance for Pediatric Radiograph Appendicular Skeleton Fracture Detection. Radiology. 2026 Feb 1;318(2):e250540. doi:10.1148/radiol.250540

2. Graves A, Mohamed A rahman, Hinton G. Speech Recognition with Deep Recurrent Neural Networks [Internet]. arXiv; 2013 [cited 2025 May 7]. Available from: http://arxiv.org/abs/1303.5778 doi:10.48550/arXiv.1303.5778

3. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv; 2018 [cited 2025 May 7]. Available from: https://arxiv.org/abs/1810.04805 doi:10.48550/ARXIV.1810.04805

4. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep Learning: A Primer for Radiologists. RadioGraphics. 2017 Nov;37(7):2113–31. doi:10.1148/rg.2017170077

5. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol. 2019 Mar 4;28(2):73–81. doi:10.1080/13645706.2019.1575882

6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28;521(7553):436–44. doi:10.1038/nature14539

7. Saba L, Biswas M, Kuppili V, Cuadrado Godia E, Suri HS, Edla DR, et al. The present and future of deep learning in radiology. Eur J Radiol. 2019 May;114:14–24. doi:10.1016/j.ejrad.2019.02.038

8. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2012. Available from: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

9. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016 Dec 13;316(22):2402–10. doi:10.1001/jama.2016.17216

10. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans Med Imaging. 2020 Apr;39(4):1184–94. doi:10.1109/TMI.2019.2945514

11. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019 Jun;25(6):954–61. doi:10.1038/s41591-019-0447-x PubMed PMID: 31110349.

12. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb 2;542(7639):115–8. doi:10.1038/nature21056 PubMed PMID: 28117445; PubMed Central PMCID: PMC8382232.

13. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data. 2021 Mar 31;8(1):53. doi:10.1186/s40537-021-00444-8

14. Rayan JC, Reddy N, Kan JH, Zhang W, Annapragada A. Binomial Classification of Pediatric Elbow Fractures Using a Deep Learning Multiview Approach Emulating Radiologist Decision Making. Radiol Artif Intell. 2019 Jan;1(1):e180015. doi:10.1148/ryai.2019180015

15. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs. Radiol Artif Intell. 2019 Jan;1(1):e180001. doi:10.1148/ryai.2019180001

16. Jin L, Yang J, Kuang K, Ni B, Gao Y, Sun Y, et al. Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet. eBioMedicine. 2020 Dec 1;62. doi:10.1016/j.ebiom.2020.103106 PubMed PMID: 33186809.

17. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018 Aug;9(4):611–29. doi:10.1007/s13244-018-0639-9

18. O'Shea K, Nash R. An Introduction to Convolutional Neural Networks [Internet]. arXiv; 2015 [cited 2024 Jun 22]. Available from: https://arxiv.org/abs/1511.08458 doi:10.48550/ARXIV.1511.08458

19. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing. 2020 Nov;415:295–316. doi:10.1016/j.neucom.2020.07.061

20. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018 Aug;9(4):611–29. doi:10.1007/s13244-018-0639-9

21. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting.

22. Burlina P, Joshi N, Paul W, Pacheco KD, Bressler NM. Addressing Artificial Intelligence Bias in Retinal Diagnostics. Transl Vis Sci Technol. 2021 Feb 5;10(2):13. doi:10.1167/tvst.10.2.13 PubMed PMID: 34003898; PubMed Central PMCID: PMC7884292.

23. Terven J, Córdova-Esparza DM, Romero-González JA. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. Mach Learn Knowl Extr. 2023 Nov 20;5(4):1680–716. doi:10.3390/make5040083

24. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection [Internet]. arXiv; 2016 [cited 2025 Mar 25]. Available from: http://arxiv.org/abs/1506.02640 doi:10.48550/arXiv.1506.02640

25. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8 [Internet]. 2023. Available from: https://github.com/ultralytics/ultralytics

26. YOLOv8 Architecture; Deep Dive into its Architecture -Yolov8 [Internet]. 2024 Jan 15 [cited 2025 May 5]. Available from: https://yolov8.org/yolov8-architecture/

27. Wang CY, Liao HYM, Yeh IH, Wu YH, Chen PY, Hsieh JW. CSPNet: A New Backbone that can Enhance Learning Capability of CNN [Internet]. arXiv; 2019 [cited 2025 May 7]. Available from: http://arxiv.org/abs/1911.11929 doi:10.48550/arXiv.1911.11929

28. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks [Internet]. arXiv; 2020 [cited 2025 May 2]. Available from: http://arxiv.org/abs/1905.11946 doi:10.48550/arXiv.1905.11946

29. Naranje SM, Erali RA, Warner WC, Sawyer JR, Kelly DM. Epidemiology of Pediatric Fractures Presenting to Emergency Departments in the United States. J Pediatr Orthop. 2016 Jun;36(4):e45–8. doi:10.1097/BPO.0000000000000595

30. Cooper C, Dennison EM, Leufkens HG, Bishop N, Van Staa TP. Epidemiology of Childhood Fractures in Britain: A Study Using the General Practice Research Database. J Bone Miner Res. 2004 Dec;19(12):1976–81. doi:10.1359/jbmr.040902

31. Kopjar B, Wickizer TM. Fractures among children: incidence and impact on daily activities. Inj Prev. 1998 Sep;4(3):194–7. doi:10.1136/ip.4.3.194

32. Barker M, Power C, Roberts I. Injuries and the risk of disability in teenagers and young adults. Arch Dis Child. 1996 Aug;75(2):156–8. doi:10.1136/adc.75.2.156 PubMed PMID: 8869200; PubMed Central PMCID: PMC1511629.

33. Mounts J, Clingenpeel J, McGuire E, Byers E, Kireeva Y. Most Frequently Missed Fractures in the Emergency Department. Clin Pediatr (Phila). 2011 Mar;50(3):183–6. doi:10.1177/0009922810384725

34. Waters PM, Skaggs DL, Flynn JM, editors. Rockwood and Wilkins' fractures in children. Tenth edition. Philadelphia, PA: Wolters Kluwer; 2025.

35. Randsborg PH, Sivertsen EA. Distal radius fractures in children: substantial difference in stability between buckle and greenstick fractures. Acta Orthop. 2009 Oct;80(5):585–9. doi:10.3109/17453670903316850

36. Wilkins KE. Principles of fracture remodeling in children. Injury. 2005 Feb;36(1):S3–11. doi:10.1016/j.injury.2004.12.007

37. Laer L von, Schneidmüller D, Hell AK. Frakturen und Luxationen im Wachstumsalter. 7., vollständig überarbeitete Auflage. Stuttgart New York: Georg Thieme Verlag; 2020. 485 p. doi:10.1055/b-006-160378

38. Noonan KJ, Price CT. Forearm and Distal Radius Fractures in Children. JAAOS - J Am Acad Orthop Surg. 1998 Jun;6(3):146.

39. Davidson JS, Brown DJ, Barnes SN, Bruce CE. Simple treatment for torus fractures of the distal radius. J Bone Joint Surg Br. 2001 Nov;83-B(8):1173–5. doi:10.1302/0301-620X.83B8.0831173

40. Crowe J, Swischuk L. Acute bowing fractures of the forearm in children: a frequently missed injury. Am J Roentgenol. 1977 Jun;128(6):981–4. doi:10.2214/ajr.128.6.981

41. Jadhav SP, Swischuk LE. Commonly missed subtle skeletal injuries in children: a pictorial review. Emerg Radiol. 2008 Nov 1;15(6):391–8. doi:10.1007/s10140-008-0733-2

42. SALTER RB, HARRIS WR. Injuries Involving the Epiphyseal Plate. JBJS [Internet]. 1963;45(3). Available from: https://journals.lww.com/jbjsjournal/fulltext/1963/45030/injuries_involving_the_epiphyseal_plate.19.aspx

43. DeFroda SF, Hansen H, Gil JA, Hawari AH, Cruz AI. Radiographic Evaluation of Common Pediatric Elbow Injuries. Orthop Rev. 2017 Feb 20;9(1):7030. doi:10.4081/or.2017.7030 PubMed PMID: 28286625; PubMed Central PMCID: PMC5337779.

44. McKinney BI, Nelson C, Carrion W. Apophyseal Avulsion Fractures of the Hip and Pelvis. Orthopedics. 2009 Jan;32(1):42–8. doi:10.3928/01477447-20090101-12

45. Ortiz EJ, Isler MH, Navia JE, Canosa R. Pathologic Fractures in Children. Clin Orthop Relat Res. 2005 Mar;432:116. doi:10.1097/01.blo.0000155375.88317.6c

46. Fredericson M, Jennings F, Beaulieu C, Matheson GO. Stress Fractures in Athletes. Top Magn Reson Imaging. 2006 Oct;17(5):309–25. doi:10.1097/RMR.0b013e3180421c8c

47. Li W, Stimec J, Camp M, Pusic M, Herman J, Boutis K. Pediatric Musculoskeletal Radiographs: Anatomy and Fractures Prone to Diagnostic Error Among Emergency Physicians. J Emerg Med. 2022 Apr 1;62(4):524–33. doi:10.1016/j.jemermed.2021.12.021 PubMed PMID: 35282940.

48. Kutanzi K, Lumen A, Koturbash I, Miousse I. Pediatric Exposures to Ionizing Radiation: Carcinogenic Considerations. Int J Environ Res Public Health. 2016 Oct 28;13(11):1057. doi:10.3390/ijerph13111057

49. Brenner DJ, Hall EJ. Computed Tomography — An Increasing Source of Radiation Exposure. N Engl J Med. 2007 Nov 29;357(22):2277–84. doi:10.1056/NEJMra072149

50. Pearce MS, Salotti JA, Little MP, McHugh K, Lee C, Kim KP, et al. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. The Lancet. 2012 Aug;380(9840):499–505. doi:10.1016/S0140-6736(12)60815-0

51. Kargl S, Pumberger W, Luczynski S, Moritz T. Assessment of interpretation of paediatric skeletal radiographs in the emergency room. Clin Radiol. 2019 Feb 1;74(2):150–3. doi:10.1016/j.crad.2018.06.024

52. Raine JE. An analysis of successful litigation claims in children in England. Arch Dis Child. 2011 Sep;96(9):838–40. doi:10.1136/adc.2011.212555 PubMed PMID: 21685505.

53. Selbst SM, Friedman MJ, Singh SB. Epidemiology and etiology of malpractice lawsuits involving children in US emergency departments and urgent care centers. Pediatr Emerg Care. 2005 Mar;21(3):165–9. PubMed PMID: 15744194.

54. Chien M, Bulloch B, Garcia-Filion P, Youssfi M, Shrader MW, Segal LS. Bedside Ultrasound in the Diagnosis of Pediatric Clavicle Fractures: Pediatr Emerg Care. 2011 Nov;27(11):1038–41. doi:10.1097/PEC.0b013e318235e965

55. Choi JY, Lim YS, Jang JH, Park WB, Hyun SY, Cho JS. Accuracy of Bedside Ultrasound for the Diagnosis of Skull Fractures in Children Aged 0 to 4 Years. Pediatr Emerg Care. 2020 May;36(5):e268–73. doi:10.1097/PEC.0000000000001485

56. Barata I, Spencer R, Suppiah A, Raio C, Ward MF, Sama A. Emergency Ultrasound in the Detection of Pediatric Long-Bone Fractures: Pediatr Emerg Care. 2012 Nov;28(11):1154–7. doi:10.1097/PEC.0b013e3182716fb7

57. Hübner U, Schlicht W, Outzen S, Barthel M, Halsband H. Ultrasound in the diagnosis of fractures in children. J Bone Joint Surg Br. 2000 Nov;82-B(8):1170–3. doi:10.1302/0301-620X.82B8.0821170

58. Alzen G, Duque-Reina D, Urhahn R, Solbach G. Röntgenuntersuchung bei Traumen im Kindesalter: Klinische und juristische Überlegung bei der Indikationsstellung. DMW - Dtsch Med Wochenschr. 2008 Mar 25;117(10):363–7. doi:10.1055/s-2008-1062320

59. Gebauer D, Mayr E, Orthner E, Ryaby JP. Low-intensity pulsed ultrasound: Effects on nonunions. Ultrasound Med Biol. 2005 Oct;31(10):1391–402. doi:10.1016/j.ultrasmedbio.2005.06.002

60. Brody AS, Frush DP, Huda W, Brent RL, and the Section on Radiology. Radiation Risk to Children From Computed Tomography. Pediatrics. 2007 Sep 1;120(3):677–82. doi:10.1542/peds.2007-1910

61. Mathews JD, Forsythe AV, Brady Z, Butler MW, Goergen SK, Byrnes GB, et al. Cancer risk in 680 000 people exposed to computed tomography scans in childhood or adolescence: data linkage study of 11 million Australians. BMJ. 2013 May 21;346(may21 1):f2360–f2360. doi:10.1136/bmj.f2360

62. Gufler H, Schulze CG, Wagner S, Baumbach L. MRI for occult physeal fracture detection in children and adolescents. Acta Radiol. 2013 May;54(4):467–72. doi:10.1177/0284185113475606

63. Spence LD, Savenor A, Nwachuku I, Tilsley J, Eustace S. MRI of fractures of the distal radius: comparison with conventional radiographs. Skeletal Radiol. 1998 May 18;27(5):244–9. doi:10.1007/s002560050375

64. Pudas T, Hurme ,T., Mattila ,K., and Svedström E. Magnetic Resonance Imaging in Pediatric Elbow Fractures. Acta Radiol. 2005 Jan 1;46(6):636–44. doi:10.1080/02841850510021643 PubMed PMID: 16334848.

65. Pai DR, Strouse PJ. MRI of the Pediatric Knee. Am J Roentgenol. 2011 May;196(5):1019–27. doi:10.2214/AJR.10.6117

66. Basener CJ, Mehlman CT, DiPasquale TG. Growth Disturbance After Distal Femoral Growth Plate Fractures in Children: A Meta-Analysis. J Orthop Trauma. 2009 Oct;23(9):663. doi:10.1097/BOT.0b013e3181a4f25b

67. Kortram K, Bezstarosti H, Metsemakers WJ, Raschke MJ, Van Lieshout EMM, Verhofstad MHJ. Risk factors for infectious complications after open fractures; a systematic review and meta-analysis. Int Orthop. 2017 Oct 1;41(10):1965–82. doi:10.1007/s00264-017-3556-5

68. Nandra R, Grover L, Porter K. Fracture non-union epidemiology and treatment. Trauma. 2016 Jan 1;18(1):3–11. doi:10.1177/1460408615591625

69. Janisch M, Apfaltrer G, Hržić F, Castellani C, Mittl B, Singer G, et al. Pediatric radius torus fractures in x-rays—how computer vision could render lateral projections obsolete. Front Pediatr. 2022 Dec 14;10:1005099. doi:10.3389/fped.2022.1005099

70. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep. 2024 Mar 13;14(1):6086. doi:10.1038/s41598-024-56706-x

71. Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. Radiology. 2022 Jul;304(1):50–62. doi:10.1148/radiol.211785

72. Hayashi D, Kompel AJ, Ventre J, Ducarouge A, Nguyen T, Regnard NE, et al. Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning. Skeletal Radiol. 2022 Nov 1;51(11):2129–39. doi:10.1007/s00256-022-04070-0

73. Ju RY, Cai W. Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm. Sci Rep. 2023 Nov 16;13(1):20077. doi:10.1038/s41598-023-47460-7

74. Zech JR, Carotenuto G, Igbinoba Z, Tran CV, Insley E, Baccarella A, et al. Detecting pediatric wrist fractures using deep-learning-based object detection. Pediatr Radiol. 2023 May 1;53(6):1125–34. doi:10.1007/s00247-023-05588-8

75. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 2018 Nov 6;115(45):11591–6. doi:10.1073/pnas.1806905115 PubMed PMID: 30348771; PubMed Central PMCID: PMC6233134.

76. Nguyen T, Maarek R, Hermann AL, Kammoun A, Marchi A, Khelifi-Touhami MR, et al. Assessment of an artificial intelligence aid for the detection of appendicular skeletal fractures in children and young adults by senior and junior radiologists. Pediatr Radiol. 2022 Oct 1;52(11):2215–26. doi:10.1007/s00247-022-05496-3

77. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs: Deep learning algorithms—are they on par with humans for diagnosing fractures? Acta Orthop. 2017 Nov 2;88(6):581–6. doi:10.1080/17453674.2017.1344459

78. Raisuddin AM, Vaattovaara E, Nevalainen M, Nikki M, Järvenpää E, Makkonen K, et al. Critical evaluation of deep neural networks for wrist fracture detection. Sci Rep. 2021 Mar 16;11(1):6006. doi:10.1038/s41598-021-85570-2

79. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Min Knowl Discov. 2019 Jul;9(4):e1312. doi:10.1002/widm.1312

80. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization [Internet]. 2016. doi:10.48550/ARXIV.1610.02391

81. Smith-Bindman R, Kwan ML, Marlow EC, Theis MK, Bolch W, Cheng SY, et al. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. JAMA. 2019 Sep 3;322(9):843–56. doi:10.1001/jama.2019.11456

82. Bhargavan M, Sunshine JH, Schepps B. Too Few Radiologists? Am J Roentgenol. 2002 May;178(5):1075–82. doi:10.2214/ajr.178.5.1781075

83. Farmakis SG, Chertoff JD, Barth RA. Pediatric Radiologist Workforce Shortage: Action Steps to Resolve. J Am Coll Radiol. 2021 Dec 1;18(12):1675–7. doi:10.1016/j.jacr.2021.07.026 PubMed PMID: 34547272.

9 Appendix

The following tool was used to optimize the language of the text:

DeepL Write, DeepL SE
14.12.2024 – 28.09.2025
URL: https://www.deepl.com/de/write