

Deteksi Fraktur pada Citra Rontgen Wrist Pediatrik Berbasis YOLOv8: Evaluasi Performa dan Analisis Keterbatasan Generalisasi Model

Timothy Fajar Sinaga^{1*}, Hardy Gustino²

¹Kedokteran, Fakultas Kedokteran, Universitas Sumatera Utara.

²Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur.

Email: timothysinaga44@gmail.com ^{1*}, hardygustino885@gmail.com ²

*Penulis Korespondensi: timothysinaga44@gmail.com ¹

Abstract. Pediatric wrist fractures are common in emergency care and require specialized expertise for accurate radiographic interpretation. This study developed and evaluated a fracture-only detector for pediatric wrist radiographs using YOLOv8 and assessed its external performance. The public GRAZPEDWRI-DX dataset, comprising 20,327 images from 6,091 patients, was used, with 18,090 fracture bounding boxes retained and non-fracture images included as background. Images were normalized, enhanced using contrast limited adaptive histogram equalization (CLAHE), and resized to 640×640 pixels. Data were split by patient identity into training, validation, and test sets at a 70/20/10 ratio to prevent patient overlap. A pretrained YOLOv8s model was trained with early stopping. On the held-out test set of 1,918 images from 610 patients, the model achieved a precision of 0.9236, recall of 0.8607, mAP50 of 0.9275, and mAP50-95 of 0.5374, with no evidence of validation overfitting. However, zero-shot evaluation on the FracAtlas dataset showed very low performance, including an mAP50 of 0.0343 overall and 0.0336 on the hand/wrist subset. CLAHE matching and lower IoU thresholds produced minimal improvement. These findings indicate strong internal performance but poor external generalization. The study is limited by a single run, single architecture, single-institution dataset, and absence of comparison models or clinical validation.

Keywords: External Generalization; Fracture Detection; Pediatric Wrist; Radiograph Analysis; YOLOv8

Abstrak Fraktur pergelangan tangan pada anak merupakan salah satu kasus yang umum ditemukan dalam pelayanan gawat darurat dan memerlukan keahlian khusus untuk interpretasi radiografi yang akurat. Penelitian ini mengembangkan dan mengevaluasi detektor khusus fraktur pada radiografi pergelangan tangan anak menggunakan YOLOv8 serta menilai kinerjanya pada evaluasi eksternal. Dataset publik GRAZPEDWRI-DX yang terdiri atas 20.327 citra dari 6.091 pasien digunakan dengan mempertahankan 18.090 *bounding box* fraktur dan citra tanpa fraktur sebagai latar belakang. Citra dinormalisasi, ditingkatkan menggunakan *contrast limited adaptive histogram equalization* (CLAHE), dan diubah ukurannya menjadi 640×640 piksel. Data dibagi berdasarkan identitas pasien menjadi data pelatihan, validasi, dan pengujian dengan rasio 70/20/10 untuk mencegah tumpang tindih pasien. Model YOLOv8s dengan bobot pra-latih dilatih menggunakan *early stopping*. Pada data uji sebanyak 1.918 citra dari 610 pasien, model mencapai *precision* 0,9236, *recall* 0,8607, mAP50 0,9275, dan mAP50-95 0,5374 tanpa indikasi *overfitting* terhadap data validasi. Namun, evaluasi *zero-shot* pada FracAtlas menunjukkan kinerja sangat rendah, dengan mAP50 sebesar 0,0343 secara keseluruhan dan 0,0336 pada subset tangan/pergelangan tangan. Penyesuaian CLAHE dan pelonggaran ambang IoU hanya memberikan sedikit perbaikan. Hasil ini menunjukkan kinerja internal yang kuat, tetapi generalisasi eksternal yang buruk. Penelitian ini terbatas pada satu kali pengujian, satu arsitektur, satu dataset institusi, tanpa model pembandingan maupun validasi klinis.

Kata kunci: Analisis Radiografi; Deteksi Fraktur; Generalisasi Eksternal; Pergelangan Anak; YOLOv8

1. LATAR BELAKANG

Fraktur merupakan salah satu kasus yang sering diterima instalasi gawat darurat, dan pada anak-anak fraktur pergelangan tangan disebut menempati porsi terbesar di antara kasus tersebut (Ju & Cai, 2023). Radiografi digital adalah modalitas standar untuk mendiagnosis kasus ini, tetapi interpretasi citranya membutuhkan pelatihan khusus yang memakan waktu dan tidak selalu tersedia (Nagy et al., 2022). Ketergantungan pada keahlian tersebut menjelaskan

mengapa deteksi fraktur otomatis menjadi topik penelitian yang aktif dalam visi komputer medis.

Bukti yang ada menunjukkan bahwa kecerdasan buatan berpotensi menjadi alat bantu yang berguna. Kuo et al. (2022), dalam tinjauan sistematik dan meta-analisis terhadap 42 studi, melaporkan sensitivitas gabungan sebesar 92% dan 91% serta spesifisitas gabungan sebesar 91% dan 91% masing-masing pada validasi internal dan eksternal, yang sebanding dengan kinerja klinisi; namun hanya 13 studi di antaranya yang melakukan validasi eksternal. Jung et al. (2024) melaporkan sensitivitas dan spesifisitas gabungan tertinggi pada modalitas radiograf (94% dan 92%), sambil menekankan bahwa pemilihan pasien dan standar rujukan merupakan sumber bias yang utama serta bahwa pelaporan yang lebih transparan masih diperlukan sebelum penerapan klinis.

Di antara pendekatan deteksi objek, keluarga YOLO banyak digunakan karena mampu mendeteksi objek dalam satu tahap dengan kecepatan tinggi (Terven et al., 2023). Publikasi dataset GRAZPEDWRI-DX (Nagy et al., 2022), yang memuat 20.327 radiograf pergelangan tangan pediatrik dari 6.091 pasien, mendorong sejumlah penelitian lanjutan. Ju dan Cai (2023) melatih YOLOv8 pada dataset ini, Chien et al. (2024) menerapkan YOLOv9, Ahmed et al. (2024) membandingkan beberapa varian YOLO dengan Faster R-CNN, dan Scherkl et al. (2026) menelaah pengaruh penambahan citra pemeriksaan lanjutan terhadap kinerja model klasifikasi dan deteksi. Di Indonesia, Hidayat et al. (2026) melaporkan penggunaan YOLOv8 dengan prapemrosesan CLAHE untuk deteksi dan klasifikasi fraktur pada dataset radiograf yang berbeda.

Meskipun demikian, terdapat dua hal yang menurut penulis masih perlu diperhatikan. Pertama, dataset GRAZPEDWRI-DX tidak menyediakan pembagian data resmi, sementara satu pasien dapat memiliki lebih dari satu citra. Pembagian per citra berisiko menempatkan citra dari pasien yang sama pada data latih dan data uji. Pada citra OCT, pembagian yang tidak tepat terbukti menggelembungkan kinerja klasifikasi sebesar 0,07 hingga 0,43 pada *Matthews correlation coefficient* (Tampu et al., 2022); besarnya efek serupa pada radiograf pergelangan tangan tidak diukur dalam penelitian ini. Ju dan Cai (2023) membagi 20.327 citra secara acak menjadi 14.204 citra latih, 4.094 citra validasi, dan 2.029 citra uji, tanpa menyebut pembagian per pasien. Kedua, evaluasi terhadap dataset dari sumber lain masih jarang. Yu et al. (2022) hanya menemukan 83 studi *deep learning* radiologi yang melakukan validasi eksternal, dan sebagian besar algoritma yang dilaporkan (70 dari 86) mengalami penurunan kinerja pada data eksternal.

Berdasarkan hal tersebut, penelitian ini bertujuan membangun dan mengevaluasi model YOLOv8 untuk mendeteksi fraktur pada citra rontgen pergelangan tangan pediatrik dengan pembagian data per pasien, menganalisis kesalahan model pada data uji, serta menguji secara *zero-shot* (tanpa *fine-tuning*) kemampuan model pada dataset fraktur eksternal FracAtlas (Abdeen et al., 2023) dan melaporkan hasilnya apa adanya. Kebaruan penelitian ini terletak pada kombinasi pembagian per pasien, evaluasi pada data uji yang benar-benar terpisah, analisis kesalahan, dan pelaporan terbuka atas hasil evaluasi eksternal beserta uji kontrol praprosesnya. Penelitian ini tidak mengklaim kinerja yang lebih baik daripada studi sebelumnya karena pembagian data, jumlah kelas, dan metrik yang dipakai berbeda.

2. KAJIAN TEORITIS

Deteksi Fraktur Berbasis *Deep Learning* pada Radiograf

Deteksi fraktur otomatis dapat dirumuskan sebagai klasifikasi citra (ada atau tidaknya fraktur) atau sebagai deteksi objek (lokalisasi fraktur dengan *bounding box*). Scherkl et al. (2026) membandingkan keduanya pada GRAZPEDWRI-DX menggunakan EfficientNet untuk klasifikasi dan YOLOv8 untuk deteksi, dan menemukan bahwa penambahan citra pemeriksaan lanjutan tidak meningkatkan kinerja klasifikasi tetapi meningkatkan sejumlah metrik deteksi. Zech et al. (2023) dan Joshi et al. (2022) masing-masing meneliti deteksi objek dan lokalisasi fraktur pergelangan tangan pada radiograf. Sejumlah studi lain menilai bantuan algoritma bagi radiolog dan dokter gawat darurat pada fraktur dewasa maupun anak (Duron et al., 2021; Guermazi et al., 2022; Nguyen et al., 2022; Gasmi et al., 2023; Fu et al., 2024).

YOLOv8 dan Metrik Evaluasi Deteksi Objek

YOLO (*You Only Look Once*) adalah keluarga detektor satu tahap yang memprediksi lokasi dan kelas objek secara langsung dari citra; perkembangan versinya dirangkum oleh Terven et al. (2023). YOLOv8 dikembangkan oleh Ultralytics (Jocher et al., 2023) dan tersusun atas *backbone*, *neck*, dan *head*. Model biasanya diinisialisasi dari bobot yang telah dilatih sebelumnya pada dataset besar (*transfer learning*).

Kinerja detektor dinilai dengan *intersection over union* (IoU), yaitu rasio luas irisan terhadap luas gabungan antara kotak prediksi dan kotak *ground truth*. Prediksi dianggap benar (*true positive*) bila IoU melampaui ambang tertentu; prediksi tanpa pasangan disebut *false positive*, dan objek *ground truth* yang tidak terdeteksi disebut *false negative*. *Precision* adalah proporsi prediksi yang benar terhadap seluruh prediksi, sedangkan *recall* adalah proporsi objek *ground truth* yang berhasil terdeteksi. *Average precision* (AP) adalah luas di bawah kurva *precision-recall*; mAP50 dihitung pada ambang IoU 0,5, sedangkan mAP50-95 merupakan

rata-rata pada ambang IoU 0,5 hingga 0,95 dengan kenaikan 0,05 (Terven et al., 2023). Karena penelitian ini hanya memodelkan satu kelas, mAP identik dengan AP kelas fraktur.

Peningkatan Kontras dengan CLAHE

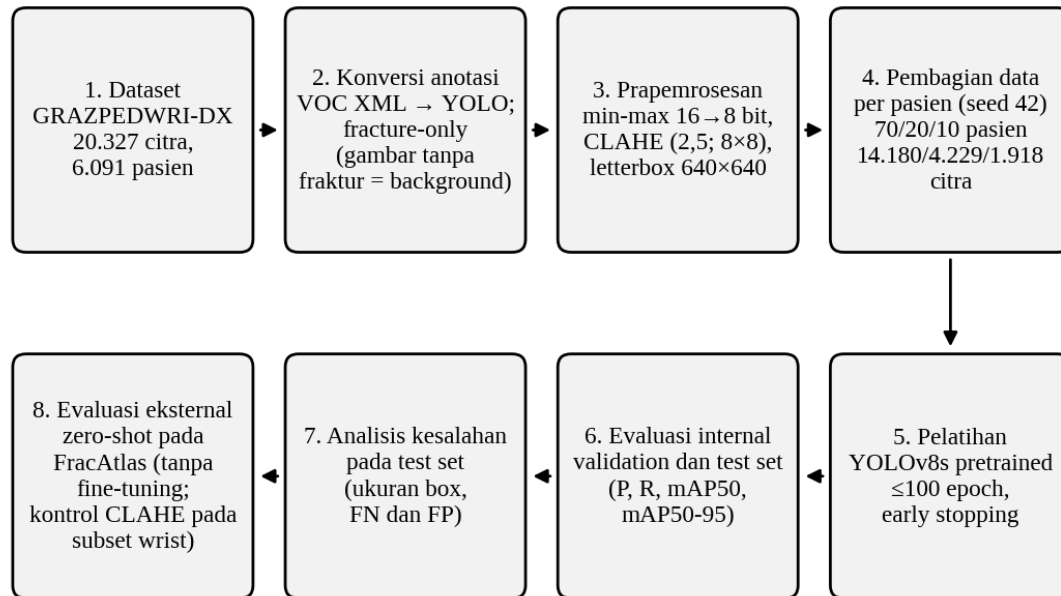
Contrast limited adaptive histogram equalization (CLAHE) meningkatkan kontras lokal dengan menyamakan histogram pada petak-petak kecil citra sambil membatasi amplifikasi derau melalui batas *clip* (Zuiderveld, 1994; Pizer et al., 1990). Teknik ini dipakai luas pada citra radiografi; misalnya Hidayat et al. (2026) menerapkannya bersama YOLOv8 pada deteksi fraktur. Pengaruh CLAHE terhadap hasil tidak diuji secara terpisah dalam penelitian ini.

Generalisasi dan Validasi Eksternal

Model yang berkinerja baik pada data pengembangan belum tentu berkinerja sama pada data dari sumber lain. Yu et al. (2022) menelaah 83 studi (86 algoritma) dan menemukan bahwa 81% melaporkan setidaknya sedikit penurunan kinerja pada data eksternal, 49% melaporkan penurunan sedikitnya moderat (0,05 pada skala unit), dan 24% melaporkan penurunan yang substansial (0,10 atau lebih); penyebab pastinya sebagian besar belum diketahui. Perbedaan populasi pasien, peralatan pencitraan, dan sebaran temuan pada data pengembangan merupakan sumber pergeseran data yang umum disebut (Yu et al., 2022). FracAtlas (Abedeen et al., 2023) adalah dataset radiograf muskuloskeletal dari tiga rumah sakit di Bangladesh dengan 4.083 citra, yang 717 citranya memuat 922 instans fraktur; berbeda dengan GRAZPEDWRI-DX, dataset ini mencakup berbagai lokasi tulang dan tidak dikhususkan untuk pasien pediatrik pergelangan tangan. Selain itu, evaluasi lintas dataset yang praprosesnya tidak setara dengan data pengembangan mengandung faktor pengganggu (*confound*), sehingga penurunan kinerja tidak dapat diatribusikan pada satu sebab saja.

3. METODE PENELITIAN

Penelitian ini merupakan studi eksperimental komputasional yang menggunakan data sekunder publik. Alur penelitian ditunjukkan pada Gambar 1 dan terdiri atas penyiapan data, pembagian data per pasien, pelatihan model, evaluasi internal, analisis kesalahan, dan evaluasi eksternal *zero-shot*.



Gambar 1. Alur penelitian.

Sumber: Hasil penelitian penulis (2026).

Dataset

Dataset utama adalah GRAZPEDWRI-DX (Nagy et al., 2022) yang tersedia publik melalui figshare, berisi 20.327 radiograf pergelangan tangan pediatrik dari 6.091 pasien yang dirawat di University Hospital Graz, Austria. Citra berupa PNG *grayscale* 16-bit. Anotasi asli berformat Pascal VOC XML dengan sembilan kelas (*boneanomaly*, *bonelesion*, *foreignbody*, *fracture*, *metal*, *periostealreaction*, *pronatorsign*, *softtissue*, dan *text*) dan dikonversi ke format YOLO (.txt). Hasil konversi berjumlah 47.444 *bounding box*, dengan 18.090 di antaranya berkelas *fracture* sesuai dengan angka rujukan dataset, dan 15 file label kosong. Berkas metadata (dataset.csv) memuat antara lain *patient_id*, usia, jenis kelamin, lateralitas, dan proyeksi, tetapi tidak memuat kolom pembagian resmi.

Pemeriksaan kualitas data menemukan 0 citra tanpa label, 0 label tanpa citra, dan 0 duplikat persis (dihitung dengan MD5 pada seluruh 20.327 citra). Validasi format label tidak menemukan masalah, tetapi dijalankan pada sampel dan bukan pada seluruh file.

Prapemrosesan Data

Penelitian ini memfokuskan diri pada deteksi fraktur, sehingga hanya kelas *fracture* yang dimodelkan (*class_id* 3 dipetakan menjadi 0) dan delapan kelas lainnya dibuang. Kelas *text* bukan temuan klinis, dan kelas lain sangat tidak seimbang (misalnya *foreignbody* hanya 8 *box*). Citra tanpa fraktur dipertahankan sebagai *background* (label kosong) untuk menekan *false positive*.

Citra 16-bit dikonversi menjadi 8-bit melalui normalisasi *min-max* yang dilanjutkan dengan CLAHE (*clipLimit* 2,5, petak 8×8); konversi langsung tanpa normalisasi menghasilkan citra rusak akibat *clipping* dan telah diperbaiki. Citra kemudian diubah ukurannya menjadi 640×640 piksel dengan *letterbox* (rasio aspek dipertahankan dan sisa area diisi hitam), dan koordinat *bounding box* diskalakan ulang sesuai transformasi tersebut.

Pembagian Data

Karena tidak ada pembagian resmi, data dibagi secara manual berdasarkan *patient_id* (bukan per citra) dengan *seed* 42 dan rasio 70/20/10 terhadap jumlah pasien, agar tidak terjadi kebocoran data antarsubset. Akibatnya, rasio jumlah citra tidak persis 70/20/10 (Tabel 1). Verifikasi menunjukkan 0 pasien yang muncul pada lebih dari satu subset. *Test set* memuat 703 citra *background*. Karena pembagian ini dibuat sendiri, angka yang dihasilkan tidak dapat dibandingkan langsung dengan penelitian lain yang memakai pembagian berbeda.

Tabel 1. Hasil pembagian data per pasien (*seed* 42).

Subset	Pasien	Citra	Box fraktur
<i>Train</i>	4.263	14.180	12.640
<i>Validation</i>	1.218	4.229	3.835
<i>Test</i>	610	1.918	1.615
Total	6.091	20.327	18.090

Model dan Konfigurasi Pelatihan

Model yang digunakan adalah YOLOv8s (Jocher et al., 2023) dengan bobot awal *pretrained* (yolov8s.pt), sehingga pelatihan bersifat *transfer learning*. Model hasil *fuse* memiliki 72 lapisan, 11.125.971 parameter, dan 28,4 GFLOPs. Konfigurasi pelatihan dan lingkungan komputasi dirangkum pada Tabel 2. Pelatihan dijalankan di Google Colab dengan GPU Tesla T4 dan sempat terputus pada *epoch* 57 karena batas sesi, lalu dilanjutkan dari *checkpoint* terakhir (last.pt) pada sesi berikutnya (*epoch* 58 sampai 78). *Early stopping* aktif pada *epoch* 78, dan bobot terbaik (best.pt) berasal dari *epoch* 58. Versi perangkat lunak pada *epoch* 1 sampai 57 tidak tercatat, sehingga tidak dapat dipastikan sama dengan sesi lanjutan.

Tabel 2. Konfigurasi Pelatihan dan Lingkungan Komputasi.

Parameter	Nilai
Arsitektur dan bobot awal	YOLOv8s, yolov8s.pt (<i>pretrained</i>)
Epoch maksimum / <i>patience</i>	100 / 20 (<i>early stopping</i>)
Ukuran citra / <i>batch</i> / <i>seed</i>	640 / 16 / 42
Optimizer / <i>learning rate</i>	auto (dipilih otomatis oleh Ultralytics) / lr0 = 0,01, lrf = 0,01; <i>weight decay</i> 0,0005; <i>warmup</i> 3 epoch
Augmentasi	Bawaan Ultralytics: <i>mosaic</i> 1,0; <i>flip horizontal</i> 0,5; <i>translate</i> 0,1; <i>scale</i> 0,5; HSV (h/s/v) 0,015/0,7/0,4; <i>random erasing</i> 0,4; RandAugment; <i>mosaic</i> dimatikan pada 10 epoch terakhir
Pelatihan: perangkat keras	Google Colab, GPU Tesla T4
Pelatihan <i>epoch</i> 58-78: perangkat lunak	Ultralytics 8.4.162, PyTorch 2.11.0 (CUDA 12.8), Python 3.13.15
Pelatihan <i>epoch</i> 1-57: perangkat lunak	Tidak tercatat
Evaluasi <i>test set</i> internal dan FracAtlas tanpa CLAHE	CPU Intel Xeon 2,20 GHz; Ultralytics 8.4.163
Evaluasi ulang FracAtlas (<i>hand/wrist</i>) dengan CLAHE	CPU Intel Xeon 2,20 GHz; Ultralytics 8.4.118

Evaluasi Model dan Analisis Kesalahan

Kinerja dinilai dengan *precision*, *recall*, mAP50, dan mAP50-95 pada *validation set* dan *test set*; *test set* tidak digunakan dalam pelatihan. Angka resmi diambil dari prosedur evaluasi Ultralytics. Sebagai gambaran titik operasi, matriks konfusi dan jumlah *true positive*, *false positive*, serta *false negative* dihitung pada ambang *confidence* 0,25 dan IoU 0,45 (bawaan Ultralytics), sehingga angkanya berbeda dari metrik resmi yang dihitung pada ambang berbeda. Analisis tingkat citra dilakukan dengan skrip sendiri (IoU 0,5; *confidence* 0,25). Kecepatan inferensi hanya diperkirakan dari log: sekitar 4,6 ms per citra pada validasi di GPU T4 dan sekitar 740 ms per citra pada CPU; angka ini bukan *benchmark* formal. Evaluasi pada *test set* internal dan pada FracAtlas tanpa CLAHE dijalankan di CPU (Intel Xeon 2,20 GHz) dengan Ultralytics 8.4.163, sedangkan pelatihan dan validasi dijalankan di GPU Tesla T4; evaluasi ulang FracAtlas dengan CLAHE memakai Ultralytics 8.4.118.

Analisis kesalahan pada *test set* mencakup *recall* per kuartil ukuran *box* (IoU 0,5; *confidence* 0,25) dan inspeksi kualitatif terhadap enam sampel acak *false negative* dan enam sampel acak *false positive*. Inspeksi kualitatif ini tidak dikuantifikasi.

Evaluasi Eksternal pada FracAtlas

Model diuji tanpa *fine-tuning* (*zero-shot*) pada FracAtlas (Abdeen et al., 2023) menggunakan bobot terbaik yang sama. Digunakan seluruh citra folder *Fractured* yang memiliki label YOLO valid dan citra *Non_fractured* sebagai *background*. Beberapa file JPG korup (*truncated*) dilewati [PERLU DIVERIFIKASI: jumlah dan asal file yang dilewati; sementara diturunkan dari log sebanyak 61, yaitu 2 dari *Fractured* dan 59 dari *Non_fractured*, belum dihitung langsung]. Pada evaluasi utama, CLAHE tidak diterapkan pada FracAtlas, berbeda dari data latih: citra JPG hanya dibuka, dikonversi ke RGB, dan diberi *letterbox*

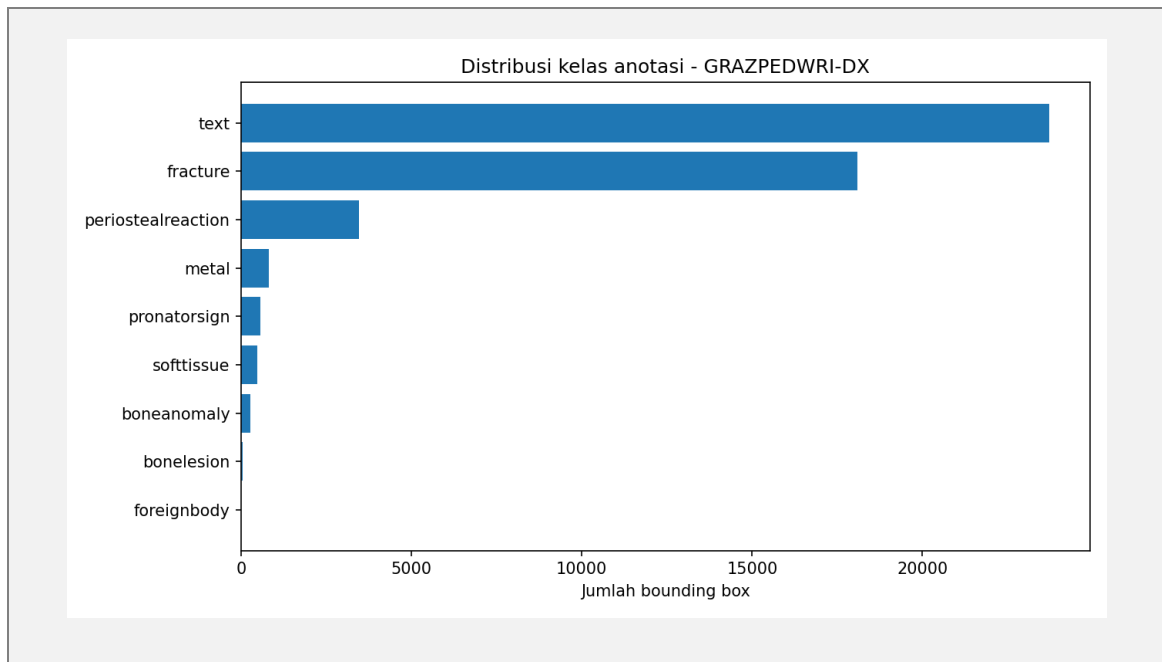
640×640. Perbedaan ini bukan persoalan kedalaman bit, melainkan distribusi kontras yang berbeda dari data latih. Sebagai kontrol parsial, subset *hand/wrist* dievaluasi ulang dengan prosedur yang disamakan dengan data latih (normalisasi *min-max*, CLAHE dengan *clipLimit* 2,5 dan petak 8×8, lalu *letterbox* 640×640). Kontrol ini bersifat parsial karena sumber citra tetap berbeda, yaitu PNG 16-bit pada data latih dan JPG 8-bit terkompresi pada FracAtlas. Evaluasi dilakukan pada seluruh FracAtlas (tanpa CLAHE) dan pada subset *hand/wrist* (tanpa dan dengan CLAHE), yakni citra dengan kolom *hand* bernilai 1; penanda ini mencakup tangan, jari, dan telapak, sehingga tidak identik dengan pergelangan tangan pada GRAZPEDWRI-DX. Sebagai uji tambahan, ambang IoU dilonggarkan pada citra fraktur subset *wrist* (549 *box ground truth*), dihitung secara manual pada *confidence* 0,25; pada ambang 0,5; 0,3; 0,2; dan 0,1 tanpa CLAHE serta pada ambang 0,5 dan 0,1 dengan CLAHE, untuk menilai seberapa banyak kegagalan berkaitan dengan perbedaan ukuran atau gaya kotak anotasi.

4. HASIL DAN PEMBAHASAN

Data penelitian berasal dari GRAZPEDWRI-DX yang dikumpulkan di University Hospital Graz, Austria, pada rentang 2008-2018 (Nagy et al., 2022), dan dari FracAtlas yang dikumpulkan dari tiga rumah sakit di Bangladesh (Abedein et al., 2023). Pelatihan dijalankan di Google Colab dengan GPU Tesla T4, sedangkan evaluasi *test set* internal dan FracAtlas tanpa CLAHE dijalankan di CPU dengan rentang waktu 3 hari. Subbagian berikut memaparkan hasil pelatihan, kinerja internal, analisis kesalahan, dan evaluasi eksternal, masing-masing disertai pembahasan; perbandingan dengan penelitian lain dan keterbatasan dibahas pada subbagian terakhir.

Karakteristik Data dan Hasil Pembagian

Pembagian per pasien menghasilkan 14.180 citra latih (4.263 pasien), 4.229 citra validasi (1.218 pasien), dan 1.918 citra uji (610 pasien) seperti terlihat pada Tabel 1. Distribusi kelas anotasi pada dataset asli ditunjukkan pada Gambar 2; kelas fraktur mendominasi dengan 18.090 *box* dari 47.444, sedangkan kelas lain seperti *foreignbody* sangat sedikit (8 *box*). Kondisi ini mendasari keputusan memodelkan fraktur saja. Ketiadaan pasien pada lebih dari satu subset memastikan bahwa kinerja pada *validation set* dan *test set* diukur pada pasien yang belum pernah dilihat model.

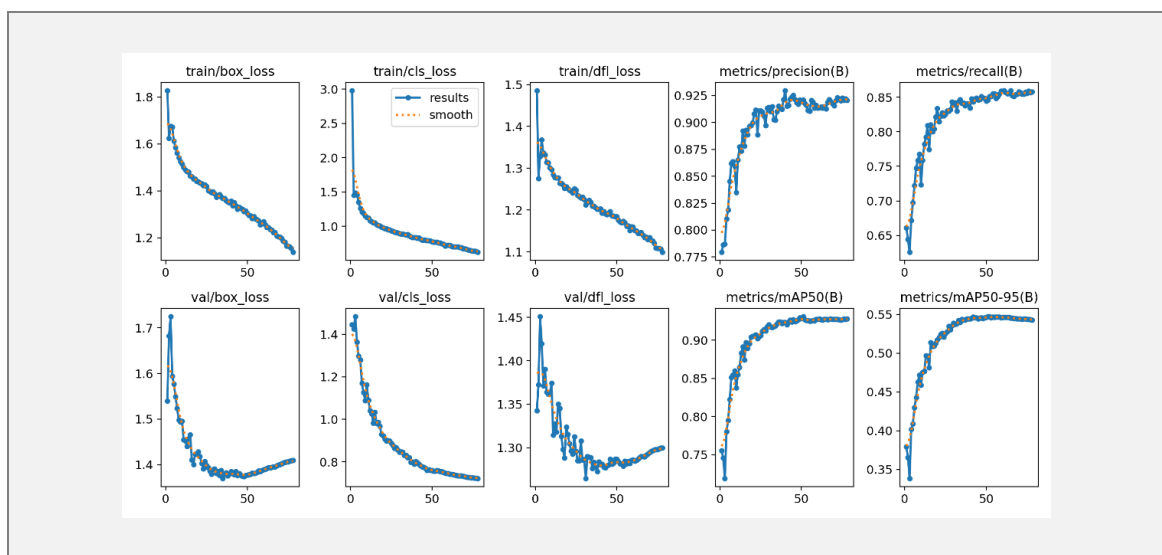


Gambar 2. Distribusi kelas anotasi pada GRAZPEDWRI-DX.

Sumber: Hasil penelitian penulis (2026), data dari Nagy et al. (2022).

Proses Pelatihan

Kurva pelatihan (Gambar 3) menunjukkan *loss* pelatihan dan validasi menurun secara stabil. *Loss* validasi naik tipis setelah *epoch* sekitar 50, yang lazim terjadi dan menjadi alasan *early stopping* aktif pada *epoch* 78 dengan *epoch* terbaik 58. Tidak tampak tanda *overfitting* berat pada kurva ini. Pelatihan yang dilanjutkan dari *checkpoint* setelah terputus pada *epoch* 57 perlu dicatat sebagai bagian dari prosedur.



Gambar 3. Kurva pelatihan YOLOv8s (*loss* dan metrik per *epoch*).

Sumber: Hasil penelitian penulis (2026).

Kinerja pada *Validation Set* dan *Test Set*

Tabel 3 menyajikan kinerja model terbaik. Pada *test set* yang terdiri atas 1.918 citra dan sepenuhnya belum pernah dilihat model, diperoleh *precision* 0,9236, *recall* 0,8607, mAP50 0,9275, dan mAP50-95 0,5374. Nilai ini sangat dekat dengan hasil *validation set* (mAP50 0,927), sehingga tidak ada indikasi *overfitting* terhadap *validation set*. Perlu dicatat bahwa hasil berasal dari satu kali percobaan dengan satu *seed* dan tanpa interval kepercayaan, sehingga selisih kecil antarsubset tidak dapat ditafsirkan lebih jauh.

Tabel 3. Kinerja YOLOv8s pada *validation set* dan *test set* (best.pt, epoch 58).

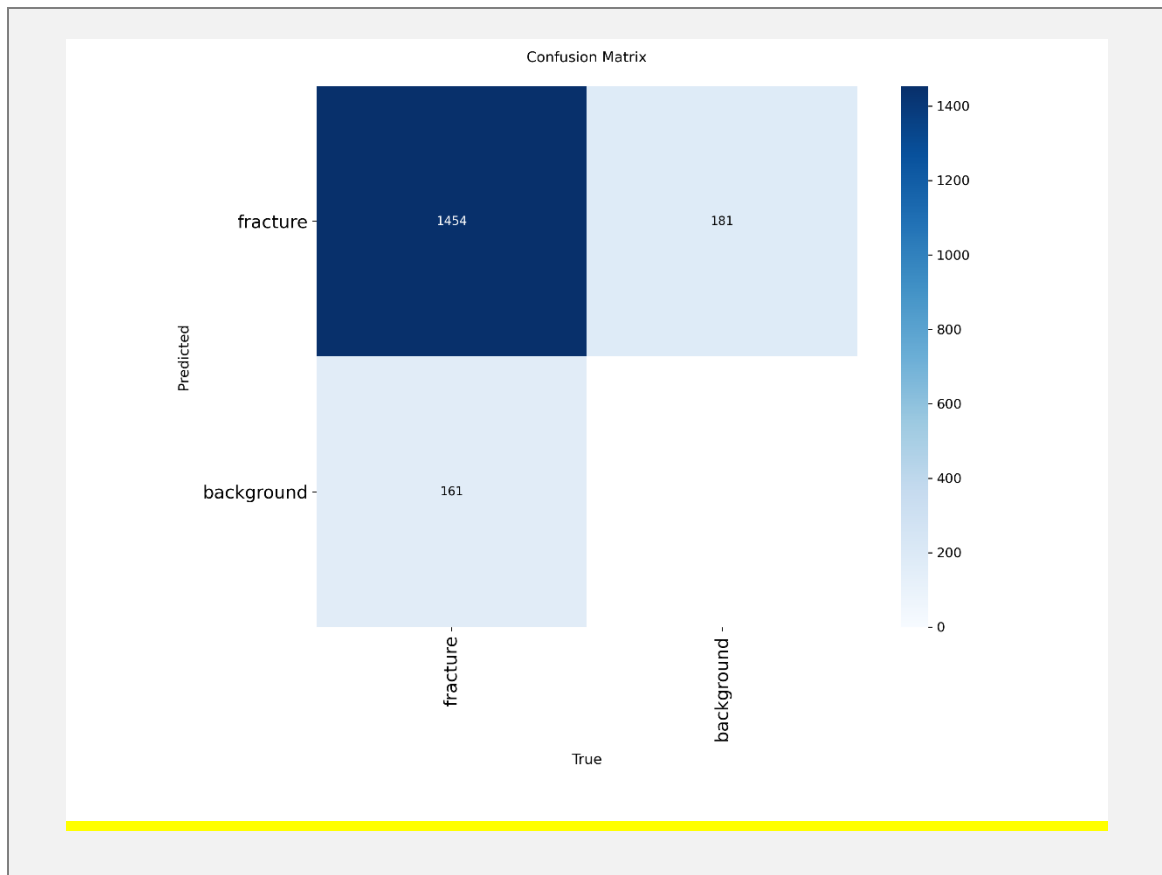
Subset	Precision	Recall	mAP50	mAP50-95
Validation	0,917	0,852	0,927	0,547
Test (angka utama)	0,9236	0,8607	0,9275	0,5374

Nilai mAP50-95 (0,5374) jauh lebih rendah daripada mAP50 (0,9275) karena mAP50-95 merata-ratakan ambang IoU yang lebih ketat hingga 0,95. Interpretasi ini menunjukkan bahwa keberadaan fraktur lebih baik terdeteksi daripada ketepatan tepi kotaknya, dan sejalan dengan observasi kualitatif pada subbagian analisis kesalahan.

Pada titik operasi *confidence* 0,25 dan IoU 0,45, *test set* menghasilkan 1.454 *true positive*, 181 *false positive*, dan 161 *false negative* (Tabel 4; Gambar 4), atau *recall* sekitar 90% dan *precision* sekitar 89%. Angka ini berbeda dari metrik resmi pada Tabel 3 karena ambang yang berbeda. Kurva F1 pada *test set* mencapai maksimum sekitar 0,89 pada *confidence* sekitar 0,377, dan kurva *precision-recall* ditunjukkan pada Gambar 5. Pada tingkat citra, 167 dari 1.918 citra uji memuat sedikitnya satu fraktur yang terlewat dan 180 dari 1.918 citra memuat sedikitnya satu salah deteksi (IoU 0,5; *confidence* 0,25). Kecepatan inferensi diperkirakan dari log sekitar 4,6 ms per citra pada validasi di GPU T4 dan sekitar 740 ms per citra pada CPU (Intel Xeon 2,20 GHz); keduanya perkiraan, bukan *benchmark* formal. Gambar 6 memperlihatkan contoh perbandingan prediksi dan *ground truth*.

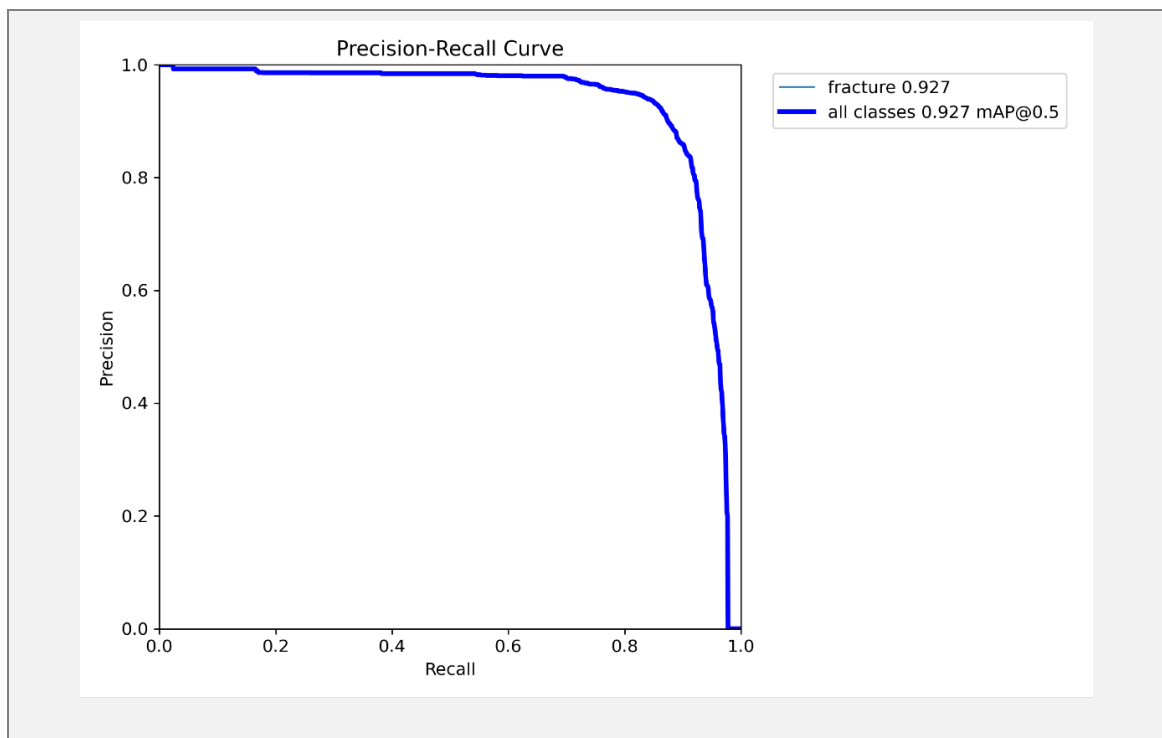
Tabel 4. Jumlah *true positive* (TP), *false positive* (FP), dan *false negative* (FN) pada *confidence* 0,25 dan IoU 0,45.

Subset	TP	FP	FN
Validation	3.391	427	444
Test	1.454	181	161



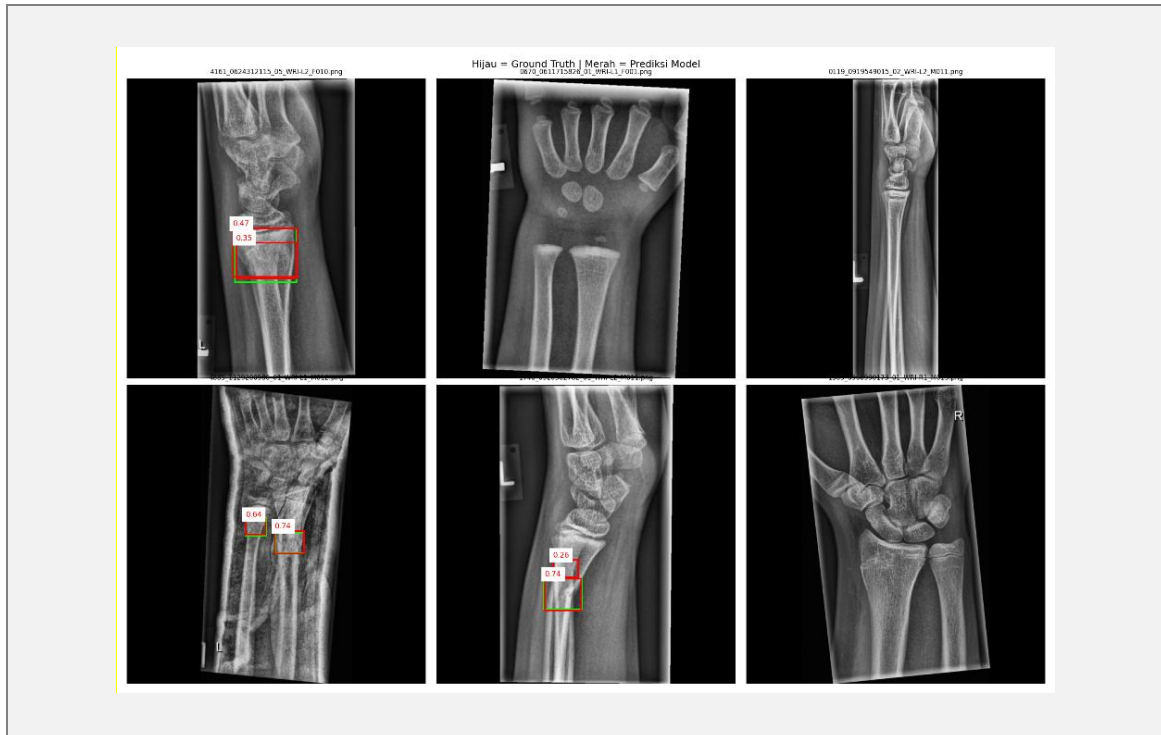
Gambar 4. Matriks Konfusi pada *Test Set* (*confidence* 0,25; *IoU* 0,45).

Sumber: Hasil penelitian penulis (2026).



Gambar 5. Kurva *Precision-Recall* pada *Test Set*.

Sumber: Hasil penelitian penulis (2026).



Gambar 6. Contoh prediksi model dibandingkan dengan *ground truth* pada *test set*.

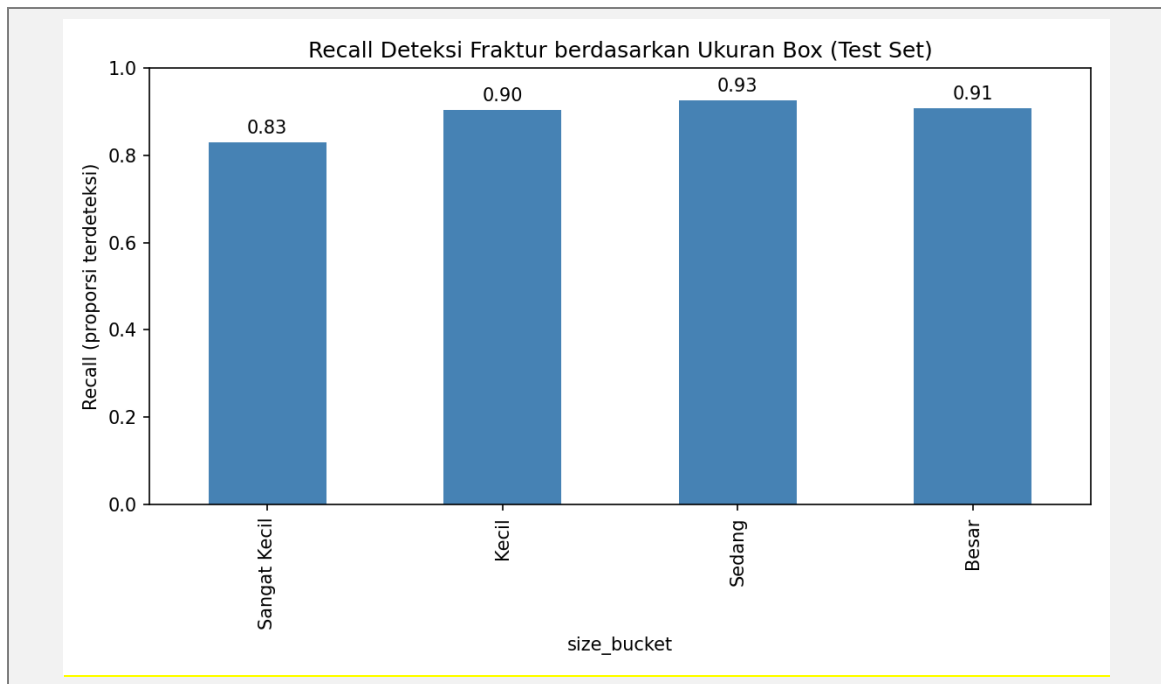
Sumber: Hasil penelitian penulis (2026).

Analisis Kesalahan

Recall per kuartil ukuran *box* pada *test set* ditunjukkan pada Tabel 5 dan Gambar 7. *Box* berukuran sangat kecil memang paling sering terlewat (*recall* 0,829), tetapi selisihnya dengan kuartil lain moderat dan tidak linear: *recall* kuartil kecil (0,903), sedang (0,926), dan besar (0,908) tidak naik secara monoton terhadap ukuran. Dengan demikian, ukuran *box* bukan faktor dominan tunggal penyebab kegagalan deteksi.

Tabel 5. *Recall* per kuartil ukuran *box* pada *test set* (IoU 0,5; *confidence* 0,25).

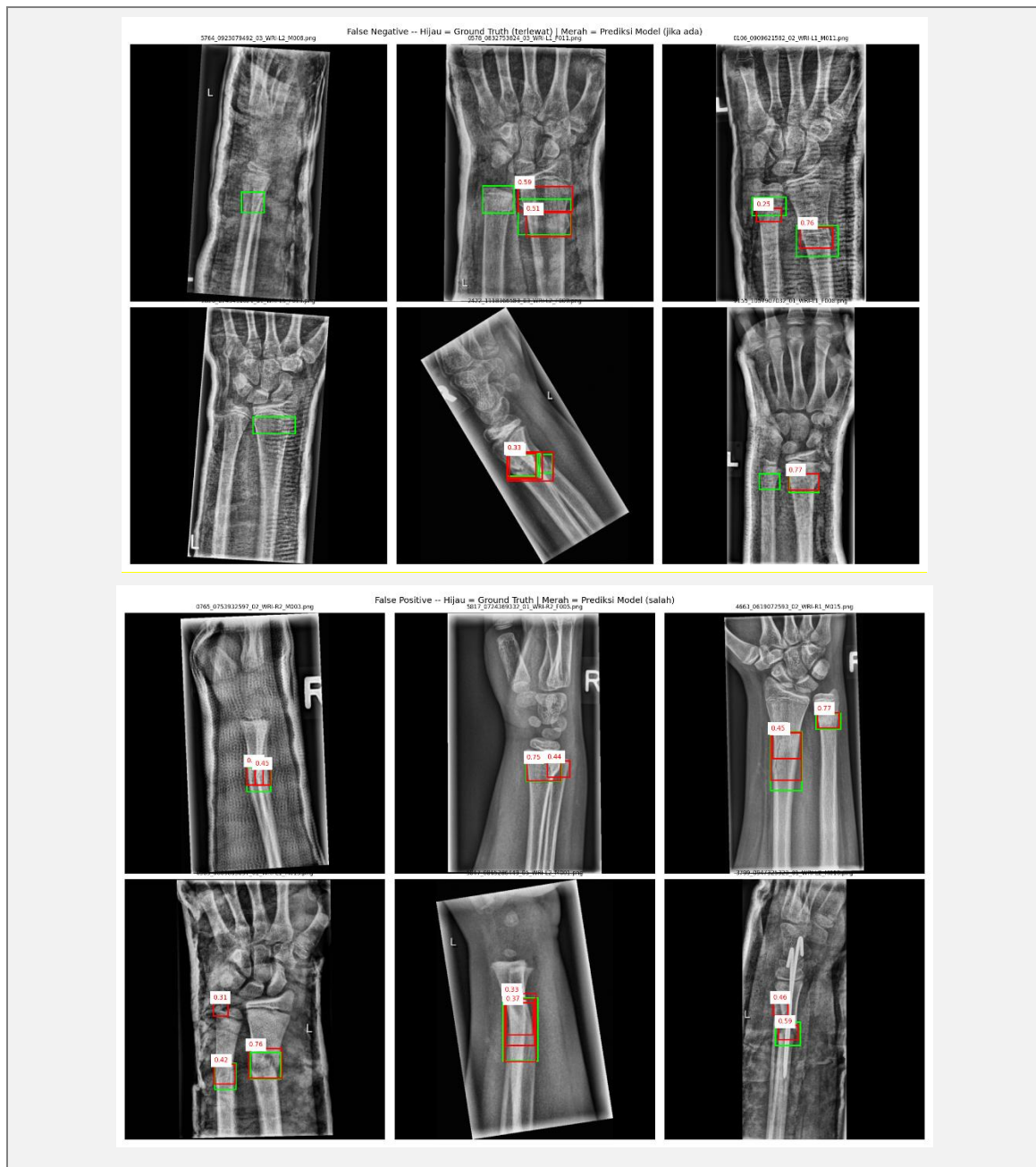
Kuartil ukuran <i>box</i>	Jumlah <i>box</i>	<i>Recall</i>
Sangat kecil	404	0,829
Kecil	404	0,903
Sedang	403	0,926
Besar	404	0,908



Gambar 7. Recall per kuartil ukuran box pada test set.

Sumber: Hasil penelitian penulis (2026).

Inspeksi kualitatif terhadap enam sampel acak *false negative* dan enam sampel acak *false positive* (Gambar 8) memperlihatkan bahwa banyak kasus yang tercatat sebagai kegagalan ternyata memiliki kotak prediksi yang berdekatan atau tumpang tindih dengan *ground truth*, tetapi dengan IoU di bawah 0,5. Pola ini mengarah pada masalah presisi lokalisasi, bukan kegagalan deteksi total. Observasi ini bersifat kualitatif, berbasis sampel yang sangat kecil, dan tidak dikuantifikasi, sehingga tidak dapat dinyatakan sebagai temuan kuantitatif.



Gambar 8. Contoh *false negative* dan *false positive* pada *test set* (sampel acak).

Sumber: Hasil penelitian penulis (2026).

Evaluasi Eksternal *Zero-Shot* pada FracAtlas

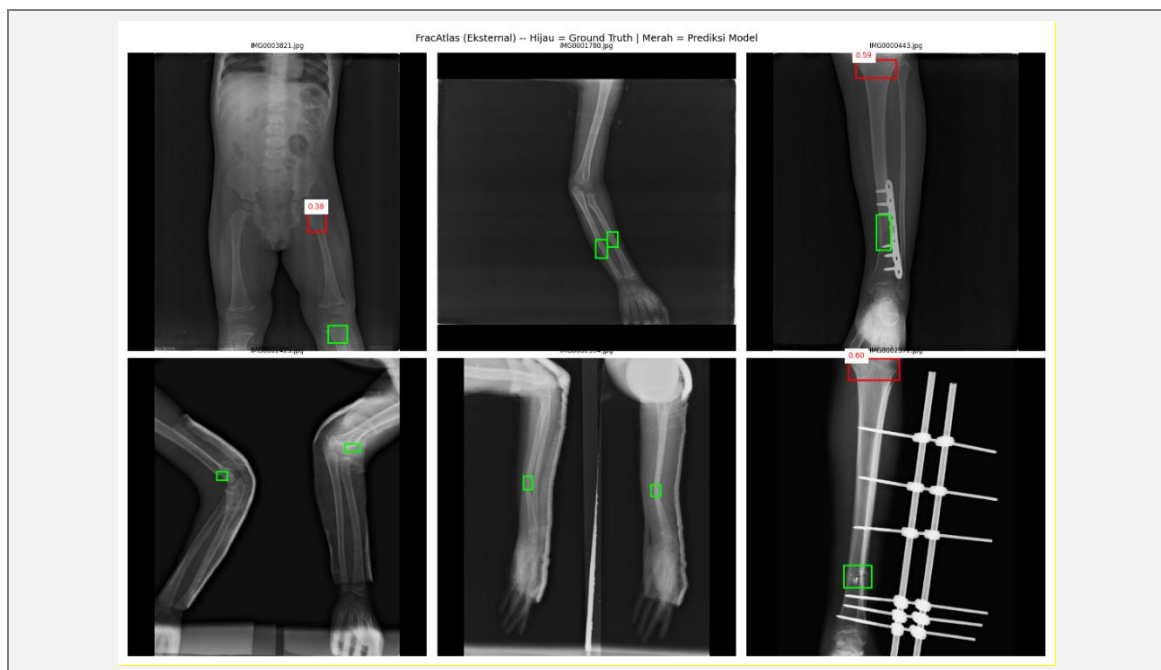
Hasil pada seluruh FracAtlas dan subset *hand/wrist*

Log Ultralytics mencatat 4.024 citra FracAtlas yang dievaluasi, terdiri atas 3.307 citra *background* dan 717 citra fraktur dengan 922 instans *ground truth*. Subset *hand/wrist* terdiri atas 1.510 citra (438 citra fraktur dan 1.072 citra *background*) dengan 549 instans. Tanpa CLAHE, performa *zero-shot* model pada kedua cakupan tersebut sangat rendah (Tabel 6; Gambar 9): mAP50 hanya 0,0343 pada seluruh citra dan 0,0336 pada subset *hand/wrist*, dibandingkan 0,9275 pada *test set* internal.

Untuk memeriksa apakah ketidakcocokan CLAHE menjadi penyebab, subset *hand/wrist* dievaluasi ulang dengan CLAHE yang disamakan dengan data latih (Tabel 6). Hasilnya tidak berubah secara berarti: mAP50 0,0384 (sebelumnya 0,0336), mAP50-95 0,0133 (0,0118), dan *recall* 0,0710 (0,0674) sedikit naik, sedangkan *precision* turun dari 0,1733 menjadi 0,1318. Arah perubahannya tidak konsisten antarmetrik dan seluruh nilainya tetap sangat rendah. Karena tidak ada uji statistik, perubahan ini hanya dideskripsikan. Ketidakcocokan CLAHE dengan demikian bukan penyebab utama, tetapi kontrol ini parsial karena sumber citra (PNG 16-bit dibandingkan JPG 8-bit terkompresi) tetap berbeda. Versi Ultralytics yang dipakai juga berbeda antarevaluasi (Tabel 6).

Tabel 6. Hasil evaluasi *zero-shot* pada FracAtlas (tanpa *fine-tuning*), tanpa dan dengan CLAHE.

Cakupan dan praproses	<i>Precision</i>	<i>Recall</i>	mAP50	mAP50-95	Ultralytics
Seluruh FracAtlas (4.024 citra), tanpa CLAHE	0,1119	0,0824	0,0343	0,0138	8.4.163
<i>Hand/wrist</i> (1.510 citra), tanpa CLAHE	0,1733	0,0674	0,0336	0,0118	8.4.163
<i>Hand/wrist</i> (1.510 citra), dengan CLAHE	0,1318	0,0710	0,0384	0,0133	8.4.118



Gambar 9. Contoh prediksi model dibandingkan dengan *ground truth* pada FracAtlas (tanpa CLAHE).

Sumber: Hasil penelitian penulis (2026), data dari Abedeen et al. (2023).

Uji pelonggaran IoU

Untuk menilai seberapa banyak kegagalan berkaitan dengan perbedaan ukuran atau gaya kotak anotasi, ambang IoU dilonggarkan pada citra fraktur subset *wrist* (Tabel 7). Tanpa CLAHE, *precision* naik moderat dari 0,343 (IoU 0,5) menjadi 0,519 (IoU 0,1), sedangkan *recall* hanya naik dari 0,067 menjadi 0,102; *true positive* naik dari 37 menjadi 56 dan *false negative* turun dari 512 menjadi 493 dari 549 *box ground truth*. Dengan CLAHE polanya serupa: *precision* naik dari 0,289 menjadi 0,485, *recall* dari 0,051 menjadi 0,086, dan *true positive* dari 28 menjadi 47. Ini merupakan observasi, bukan penjelasan sebab: pada kedua pengaturan sebagian besar *ground truth* tidak terdeteksi sama sekali, dan ketidakcocokan ukuran atau gaya kotak hanya menjelaskan sebagian kecil kegagalan. Pada uji manual ini (*confidence* 0,25, hanya citra fraktur), *recall* dengan CLAHE justru lebih rendah daripada tanpa CLAHE pada kedua ambang, berlawanan dengan arah *recall* pada metrik resmi (Tabel 6), yang sejalan dengan tidak konsistennya pengaruh CLAHE.

Tabel 7. Uji pelonggaran IoU pada citra fraktur subset *wrist* FracAtlas (549 *box ground truth*), tanpa dan dengan CLAHE (P: *precision*; R: *recall*; –: tidak tersedia).

IoU	P tanpa CLAHE	R tanpa CLAHE	TP tanpa CLAHE	P dengan CLAHE	R dengan CLAHE	TP dengan CLAHE
0,5	0,343	0,067	37	0,289	0,051	28
0,3	0,444	0,087	48	–	–	–
0,2	0,491	0,097	53	–	–	–
0,1	0,519	0,102	56	0,485	0,086	47

Penurunan tetap terjadi setelah CLAHE disamakan, sehingga ketidakcocokan CLAHE bukan penyebab utama rendahnya performa ini. Penyebab sebenarnya belum dapat ditentukan; dugaan yang konsisten dengan data mencakup (i) karakteristik citra, termasuk kompresi JPG 8-bit pada FracAtlas dibandingkan PNG 16-bit pada data latih, (ii) perbedaan populasi dan cakupan anatomi, karena FracAtlas mencakup berbagai lokasi tulang dan penanda *hand* tidak identik dengan pergelangan tangan pada GRAZPEDWRI-DX, serta (iii) perbedaan skala dan gaya anotasi, karena kotak *ground truth* FracAtlas sangat kecil. Kontrol CLAHE bersifat parsial: evaluasi ulangnya hanya pada subset *hand/wrist*, sumber citra tetap berbeda, dan tidak ada eksperimen terisolasi yang memisahkan ketiga faktor tersebut, sehingga kontribusi masing-masing tidak dapat dipisahkan. *Fine-tuning* pada FracAtlas tidak dilakukan, sehingga penelitian ini tidak membuat klaim tentang hasil adaptasi model.

Konteks terhadap Penelitian Sebelumnya

Hasil penelitian ini hanya diletakkan sebagai konteks terhadap studi lain pada GRAZPEDWRI-DX, tanpa perbandingan peringkat. Ju dan Cai (2023) melaporkan mAP50 sebesar 0,638 pada *validation set*, dirata-ratakan atas sembilan kelas, untuk model terbaik mereka pada ukuran citra 1024; tabel per kelas untuk YOLOv8s pada ukuran citra yang sama memuat mAP50 0,947 dan mAP50-95 0,569 untuk kelas fraktur saja. Angka kelas fraktur tersebut secara nominal lebih tinggi daripada mAP50 penelitian ini (0,927 pada *validation set* dan 0,9275 pada *test set*), tetapi keduanya tidak dapat dibandingkan langsung: pembagian data mereka acak atas citra sedangkan penelitian ini per pasien, ukuran citra masukan berbeda (1024 dibandingkan 640), dan angka mereka diukur pada *validation set*. Ahmed et al. (2024) melaporkan mAP50 kelas fraktur sebesar 0,95 untuk YOLOv8m, YOLOv8l, dan YOLOv8x (sensitivitas fraktur tertinggi 0,92 pada YOLOv8m) serta mAP50 seluruh kelas tertinggi sebesar 0,77 pada YOLOv8x. Angka kelas fraktur ini juga secara nominal lebih tinggi daripada mAP50 penelitian ini, tetapi tidak dapat dibandingkan langsung: data mereka dibagi acak atas citra (15.245 latih, 4.066 validasi, dan 1.016 uji), model dilatih dari awal (*from scratch*) selama 100 *epoch*, dan makalah tersebut tidak menyebutkan secara eksplisit apakah angka pada tabel hasilnya berasal dari *validation set* atau *test set*. Karena itu, penelitian ini tidak mengklaim kinerja yang lebih baik atau tergolong *state-of-the-art*.

Rendahnya performa pada data eksternal searah dengan temuan umum bahwa sebagian besar algoritma *deep learning* radiologi berkinerja lebih rendah pada data eksternal (Yu et al., 2022). Pada penelitian ini, penurunan tetap terjadi setelah CLAHE disamakan pada subset *hand/wrist*; penyebabnya diduga terkait karakteristik citra dan kompresi, cakupan anatomi, serta skala dan gaya anotasi, tetapi kontribusi masing-masing tidak dapat dipisahkan. Besarnya penurunan juga tidak setara dengan besaran yang dilaporkan tinjauan tersebut, sebab penelitian ini menilai deteksi dengan mAP pada anatomi dan gaya anotasi yang berbeda. Scherkl et al. (2026) juga mencatat bahwa data satu institusi dan ketiadaan validasi eksternal membatasi generalisasi temuan, dan keterbatasan yang sama berlaku di sini.

Keterbatasan dan Implikasi

Penelitian ini memiliki beberapa keterbatasan. Pertama, seluruh pelatihan, validasi, dan pengujian internal memakai satu dataset publik dari satu institusi dan satu populasi, bukan data klinis primer. Kedua, performa *zero-shot* pada FracAtlas sangat rendah (mAP50 0,0343) dan tetap rendah setelah CLAHE disamakan pada subset *hand/wrist* (mAP50 0,0384), tetapi penyebabnya tidak terisolasi. Kontrol praproses ini hanya parsial: sumber citra (PNG 16-bit dibandingkan JPG 8-bit terkompresi) tetap berbeda, dan evaluasi ulang hanya dilakukan pada

subset *hand/wrist*. Penelitian ini tidak dapat memisahkan kontribusi karakteristik citra, cakupan anatomi, serta skala dan gaya anotasi, sehingga kemampuan model pada data dari sumber lain tidak dapat dinilai lebih jauh dari hasil yang dilaporkan. Ketiga, hanya satu arsitektur (YOLOv8s) yang diuji, tanpa *ablation* dan model pembandingan, dengan satu *seed* dan tanpa interval kepercayaan; dengan demikian kontribusi CLAHE maupun kontribusi pembagian per pasien terhadap kinerja tidak dapat dipisahkan. Keempat, hanya kelas fraktur yang dimodelkan; delapan kelas anotasi lain tidak dipelajari. Kelima, belum ada validasi klinis atau *review* oleh radiolog, sehingga hasil ini bukan klaim kegunaan klinis. Keenam, analisis kesalahan kualitatif berbasis sampel kecil. Ketujuh, pembagian dibuat sendiri per pasien dan bukan pembagian resmi dataset, sehingga angka tidak dapat dibandingkan langsung dengan makalah lain. Kedelapan, versi perangkat lunak pada *epoch* 1 sampai 57 tidak tercatat, evaluasi *test set* serta FracAtlas tanpa CLAHE dijalankan pada perangkat (CPU) yang berbeda dari pelatihan (GPU), dan versi Ultralytics berbeda antarevaluasi (8.4.163 tanpa CLAHE dan 8.4.118 dengan CLAHE).

Secara terapan, hasil ini dapat dipakai sebagai landasan awal bagi pengembangan lanjutan pendeteksi fraktur pergelangan tangan pediatrik pada data dari satu domain, dengan catatan bahwa penerapan pada data dari sumber lain memerlukan validasi tambahan dan kemungkinan adaptasi. Secara teoretis, pelaporan hasil *zero-shot* yang terbuka, beserta kontrol CLAHE-nya, menunjukkan bahwa menyamakan CLAHE tidak cukup untuk memulihkan performa pada FracAtlas; kinerja tinggi pada satu dataset tidak dengan sendirinya menjamin kinerja pada dataset lain, sejalan dengan literatur validasi eksternal.

5. KESIMPULAN DAN SARAN

Penelitian ini membangun model YOLOv8s untuk mendeteksi fraktur pada citra rontgen pergelangan tangan pediatrik dengan pembagian data per pasien. Pada *test set* yang belum pernah dilihat model, diperoleh *precision* 0,9236, *recall* 0,8607, mAP50 0,9275, dan mAP50-95 0,5374, yang sangat dekat dengan hasil *validation set* dan tidak menunjukkan indikasi *overfitting* terhadapnya. Analisis kesalahan menunjukkan bahwa *box* berukuran sangat kecil paling sering terlewat (*recall* 0,829), tetapi ukuran bukan faktor dominan tunggal, dan observasi kualitatif mengarah pada masalah presisi lokalisasi pada sebagian kasus. Performa *zero-shot* pada FracAtlas sangat rendah (mAP50 0,0343 pada seluruh citra dan 0,0336 pada subset *hand/wrist*) dan tetap rendah setelah CLAHE disamakan pada subset *hand/wrist* (mAP50 0,0384). Penyebabnya diduga terkait karakteristik citra dan kompresi, cakupan anatomi, serta skala dan gaya anotasi, tetapi kontribusi masing-masing tidak dapat dipisahkan

karena tidak ada eksperimen terisolasi. Hasil ini berasal dari satu kali percobaan pada satu dataset dan tidak merupakan klaim kegunaan klinis maupun klaim generalisasi.

Untuk penelitian selanjutnya disarankan untuk mengulang pelatihan dengan beberapa *seed* dan melaporkan interval kepercayaan, menambahkan model pembanding dan *ablation* (termasuk pengaruh CLAHE), mengkuantifikasi kesalahan lokalisasi secara sistematis, melakukan eksperimen terisolasi untuk memisahkan pengaruh karakteristik citra (misalnya kompresi JPG), cakupan anatomi, serta skala dan gaya anotasi pada evaluasi eksternal, menguji *fine-tuning* atau pelatihan multi-pusat untuk menilai dan memperbaiki generalisasi, serta melakukan *review* oleh radiolog dan validasi klinis sebelum penerapan. Penelitian berikutnya juga dapat memodelkan kelas anotasi lain yang tidak dicakup di sini dan melaporkan hasil pada pembagian yang dapat dibandingkan dengan penelitian lain.

UCAPAN TERIMA KASIH

Penulis berterima kasih kepada tim Medical University of Graz yang menyediakan dataset GRAZPEDWRI-DX dan kepada penyusun dataset FracAtlas atas ketersediaan data secara publik.

DAFTAR REFERENSI

- Abdedeen, I., Rahman, M. A., Prottyasha, F. Z., Ahmed, T., Chowdhury, T. M., & Shatabda, S. (2023). FracAtlas: A dataset for fracture classification, localization and segmentation of musculoskeletal radiographs. *Scientific Data*, 10(1), 521. <https://doi.org/10.1038/s41597-023-02432-4>
- Ahmed, A., Imran, A. S., Manaf, A., Kastrati, Z., & Daudpota, S. M. (2024). Enhancing wrist abnormality detection with YOLO: Analysis of state-of-the-art single-stage detection models. *Biomedical Signal Processing and Control*, 93, 106144. <https://doi.org/10.1016/j.bspc.2024.106144>
- Chien, C.-T., Ju, R.-Y., Chou, K.-Y., & Chiang, J.-S. (2024). YOLOv9 for fracture detection in pediatric wrist trauma X-ray images. *Electronics Letters*, 60(11), e13248. <https://doi.org/10.1049/ell2.13248>
- Duron, L., Ducarouge, A., Gillibert, A., Lainé, J., Allouche, C., Cherel, N., ... Feydy, A. (2021). Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: A multicenter cross-sectional diagnostic study. *Radiology*, 300(1), 120–129. <https://doi.org/10.1148/radiol.2021203886>
- Fu, T., Viswanathan, V., Attia, A., Zerbib-Attal, E., Kosaraju, V., Barger, R., ... Faraji, N. (2024). Assessing the potential of a deep learning tool to improve fracture detection by radiologists and emergency physicians on extremity radiographs. *Academic Radiology*, 31(5), 1989–1999. <https://doi.org/10.1016/j.acra.2023.10.042>

- Gasmi, I., Calinghen, A., Parienti, J.-J., Belloy, F., Fohlen, A., & Pelage, J.-P. (2023). Comparison of diagnostic performance of a deep learning algorithm, emergency physicians, junior radiologists and senior radiologists in the detection of appendicular fractures in children. *Pediatric Radiology*, 53(8), 1675–1684. <https://doi.org/10.1007/s00247-023-05621-w>
- Guerhazi, A., Tannoury, C., Kompel, A. J., Murakami, A. M., Ducarouge, A., Gillibert, A., ... Hayashi, D. (2022). Improving radiographic fracture recognition performance and efficiency using artificial intelligence. *Radiology*, 302(3), 627–636. <https://doi.org/10.1148/radiol.210937>
- Hidayat, J. J., Anshor, A. H., & Anwar, M. S. (2026). Pemodelan deteksi dan klasifikasi fraktur tulang pada radiografi X-ray menggunakan YOLOv8 dan preprocessing CLAHE. *Jurnal FASILKOM*, 16(1). <https://doi.org/10.37859/jf.v16i1.11241>
- Jocher, G., Chaurasia, A., & Qiu, J. (2023). *Ultralytics YOLO* [Computer software]. Retrieved from <https://github.com/ultralytics/ultralytics>
- Joshi, D., Singh, T. P., & Joshi, A. K. (2022). Deep learning-based localization and segmentation of wrist fractures on X-ray radiographs. *Neural Computing and Applications*, 34(21), 19061–19077. <https://doi.org/10.1007/s00521-022-07510-z>
- Ju, R.-Y., & Cai, W. (2023). Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm. *Scientific Reports*, 13(1), 20077. <https://doi.org/10.1038/s41598-023-47460-7>
- Jung, J., Dai, J., Liu, B., & Wu, Q. (2024). Artificial intelligence in fracture detection with different image modalities and data types: A systematic review and meta-analysis. *PLOS Digital Health*, 3(1), e0000438. <https://doi.org/10.1371/journal.pdig.0000438>
- Kuo, R. Y. L., Harrison, C., Curran, T.-A., Jones, B., Freethy, A., Cussons, D., ... Furniss, D. (2022). Artificial intelligence in fracture detection: A systematic review and meta-analysis. *Radiology*, 304(1), 50–62. <https://doi.org/10.1148/radiol.211785>
- Nagy, E., Janisch, M., Hrzić, F., Sorantin, E., & Tschauner, S. (2022). A pediatric wrist trauma X-ray dataset (GRAZPEDWRI-DX) for machine learning. *Scientific Data*, 9(1), 222. <https://doi.org/10.1038/s41597-022-01328-z>
- Nguyen, T., Maarek, R., Hermann, A.-L., Kammoun, A., Marchi, A., Khelifi-Touhami, M. R., ... Le Pointe, H. D. (2022). Assessment of an artificial intelligence aid for the detection of appendicular skeletal fractures in children and young adults by senior and junior radiologists. *Pediatric Radiology*, 52(11), 2215–2226. <https://doi.org/10.1007/s00247-022-05496-3>
- Pizer, S. M., Johnston, R. E., Ericksen, J. P., Yankaskas, B. C., & Muller, K. E. (1990). Contrast-limited adaptive histogram equalization: Speed and effectiveness. In *Proceedings of the First Conference on Visualization in Biomedical Computing* (pp. 337–345). IEEE. <https://doi.org/10.1109/VBC.1990.109340>
- Scherkl, M., Stranger, N., Ciornei-Hoffman, A., Singer, G., Till, T., Till, H., ... Tschauner, S. (2026). Automated AI fracture detection in initial presentation pediatric wrist X-rays: Effects and benefits of adding follow-up examinations. *La Radiologia Medica*, 131(3), 458–469. <https://doi.org/10.1007/s11547-025-02153-1>
- Tampu, I. E., Eklund, A., & Haj-Hosseini, N. (2022). Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Scientific Data*, 9(1), 580. <https://doi.org/10.1038/s41597-022-01618-6>

- Terven, J., Córdova-Esparza, D.-M., & Romero-González, J.-A. (2023). A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, 5(4), 1680–1716. <https://doi.org/10.3390/make5040083>
- Yu, A. C., Mohajer, B., & Eng, J. (2022). External validation of deep learning algorithms for radiologic diagnosis: A systematic review. *Radiology: Artificial Intelligence*, 4(3), e210064. <https://doi.org/10.1148/ryai.210064>
- Zech, J. R., Carotenuto, G., Igbinoba, Z., Tran, C. V., Insley, E., Baccarella, A., & Wong, T. T. (2023). Detecting pediatric wrist fractures using deep-learning-based object detection. *Pediatric Radiology*, 53(6), 1125–1134. <https://doi.org/10.1007/s00247-023-05588-8>
- Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. In *Graphics Gems IV* (pp. 474–485). Academic Press.