

A Unified DERNet and YOLO11 Framework for Spinal Lesion Triage, Segmentation and Clinical Deployment

Anonymized Authors

Anonymized Affiliations
`email@anonymized.com`

Abstract. Automated analysis of spinal radiographs is critical to early diagnosis but is currently limited by severe class imbalance and the visual subtlety of lesions. This work introduces DERNet—an ensemble of DenseNet121, EfficientNetV2-S, and ResNet50—for highly sensitive classification, coupled with a YOLO11-L model for precise lesion detection. Our proposed method utilizes architectural diversity to overcome the sensitivity-specificity trade-off, along with a CSP-Darknet backbone to effectively capture the fine features of minute pathologies. The proposed framework is validated on the VinDr-SpineXR benchmark, yielding state-of-the-art AUROC of 91.03% and detection mAP@0.5 on 40.10%. In addition to the technical contributions, the authors address the translational challenge with the development of a functional web-based interface that can process various image formats, thereby facilitating the practical application of the proposed framework. All the code associated with this study is available at: GitHub

Keywords: Spinal Lesion · DERNet · Class Imbalance · Clinical Deployment.

1 Introduction

The spinal column is central to mechanical integrity yet vulnerable to pathologies ranging from degenerative diseases to structural anomalies [18,9]. While conventional radiography remains the screening gold standard, manual interpretation is confounded by anatomical superposition and inter-observer variability, necessitating high-precision automated assistance [12,7,1]. The absence of timely diagnosis frequently precipitates severe sequelae, including permanent neurological deficit [11,5]. To address this, we present a unified framework designed to resolve the critical trade-off between diagnostic sensitivity and localization specificity.

Our contribution is threefold. First, we establish Robust Triage via DERNet, a heterogeneous ensemble synergizing DenseNet121 [4], EfficientNetV2-S [17], and ResNet50 [3] to exploit diverse inductive biases, achieving a state-of-the-art AUROC of 91.03%. Second, we introduce High-Precision Localization using a

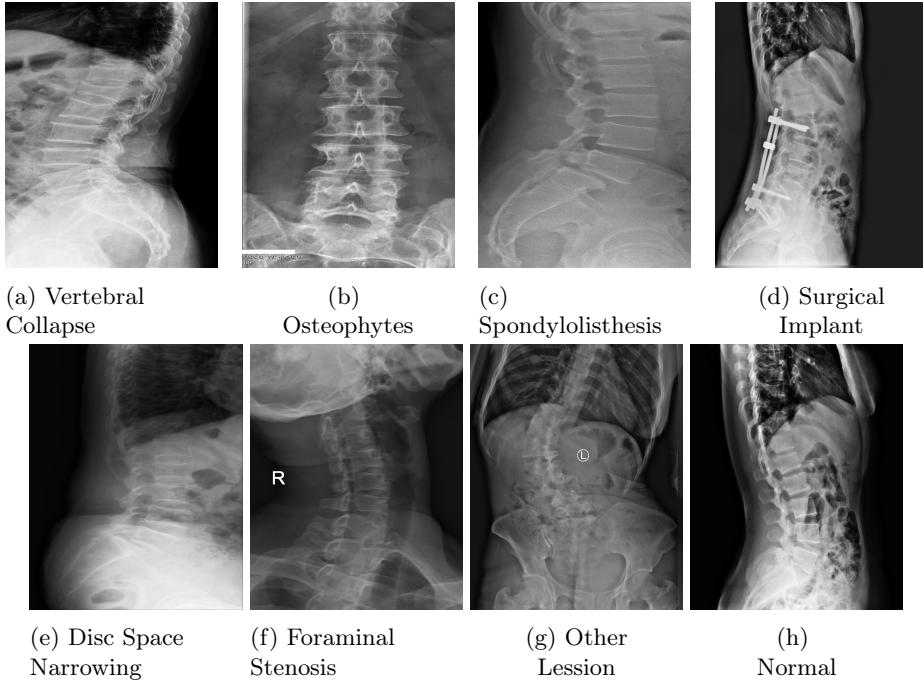


Fig. 1: Examples of detected spinal lesions in X-ray images. The figure highlights four distinct pathologies: (a) Vertebral Collapse, (b) Osteophytes, (c) Spondylolisthesis, (d) Surgical Implants, (e) Disc Space Narrowing, (f) Foraminal Stenosis, (g) Other Lesson and (h) Normal.

customized YOLO11-L detector optimized for class imbalance, securing a benchmark mAP@0.5 of 40.10% on subtle lesions like Foraminal Stenosis. Finally, we facilitate Explainable Clinical Deployment through an ONNX-accelerated web interface integrated with real-time Grad-CAM and LIME analysis, ensuring transparent and trustworthy clinical decision support.

2 Related Work

Automated spinal analysis has undergone a paradigm shift toward Deep Learning [14], yet robust clinical adoption remains elusive. Although Nguyen et al. established the foundational *VinDr-SpineXR* benchmark [9], their baseline architectures exhibited sensitivity constraints that compromise the reliability required for primary clinical triage. Subsequent innovations, such as *HealNNet* [6], have attempted to mitigate this but often encounter diminishing returns or prohibitive computational overhead. We address these limitations by synergizing DenseNet, EfficientNet, and ResNet into a weighted ensemble that harmonizes feature reuse with parameter efficiency to maximize screening performance.

In lesion localization, efficacy is severely curtailed by extreme class imbalance (46.9:1) and minute pathology scales. Conventional two-stage detectors like Faster R-CNN [13] and Sparse R-CNN [15] frequently fail to resolve subtle feature maps, while edge-guided networks such as *EGCA-Net* [2] remain susceptible to complex background artifacts. Furthermore, recent YOLO adaptations [16] have been principally confined to fracture detection. We advance this domain by optimizing the YOLO11-L architecture specifically for the fine-grained, high-imbalance demands of multi-class spinal diagnostics.

3 Methodology

As illustrated in Fig. 2, we propose a unified hierarchical framework that synergizes a probabilistic DERNet ensemble for robust triage with a specialized YOLO11-L detector. This cascaded design effectively decouples high-sensitivity screening from the precise localization of fine-grained spinal pathologies.

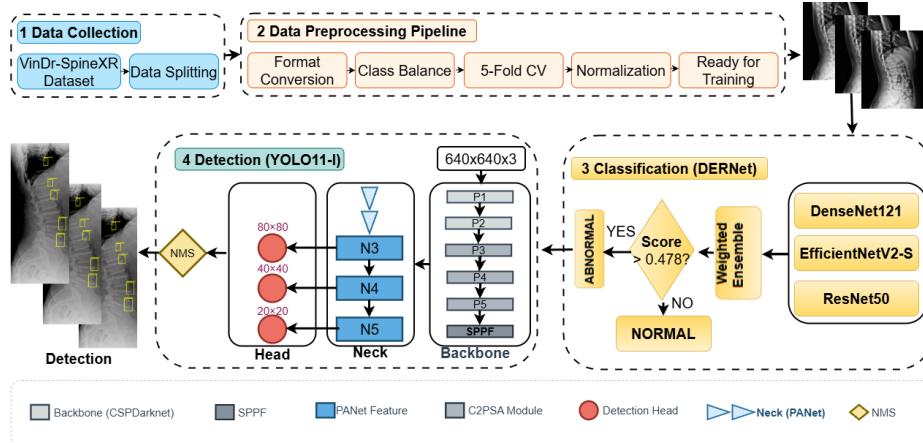


Fig. 2: Architectural Overview. The pipeline follows four stages: (1) Data Collection, (2) Adaptive Preprocessing Pipeline, (3) Probabilistic DERNet Classification for binary triage, and (4) YOLO11-L Detection optimized for fine-grained lesion localization.

3.1 Dataset and Adaptive Preprocessing

We leveraged the *VinDr-SpineXR* benchmark [10] ($N = 8,389$), addressing its severe 46.9:1 imbalance via a stratified, multi-stage preprocessing protocol. To mitigate radiographic exposure variance, Contrast Limited Adaptive Histogram

Equalization (CLAHE) computes enhanced intensity $I'(i, j)$ by mapping the local Cumulative Distribution Function (CDF):

$$I'(i, j) = \beta \left(\frac{CDF_{\Omega_k}(I(i, j)) - CDF_{min}}{|\Omega_k| - CDF_{min}} \right), \quad (1)$$

where $|\Omega_k|$ denotes local region cardinality and β the dynamic range factor. A clip limit $\tau = 2.0$ suppresses noise amplification in homogeneous bone structures prior to normalization and 5-fold cross-validation.

3.2 Probabilistic DERNet Classification

To optimize normal study filtration, we construct DERNet, a heterogeneous ensemble (DenseNet121, EfficientNetV2-S, ResNet50) synergizing feature reuse, parameter efficiency, and residual flow. Given input \mathbf{x} , each backbone f_{θ_m} generates posterior $P(y = c|\mathbf{x}; \theta_m)$, which is fused via Weighted Soft-Voting (Eq. 2) to marginalize predictive uncertainty:

$$\hat{y} = {}_{c \in \{0,1\}} \left(\sum_{m=1}^{|\mathcal{E}|} \omega_m \cdot P(y = c|\mathbf{x}; \theta_m) \right), \quad (2)$$

subject to $\sum \omega_m = 1$. Empirical weights $\Omega = [0.42, 0.32, 0.26]$ prioritize DenseNet's feature propagation for subtle textures.

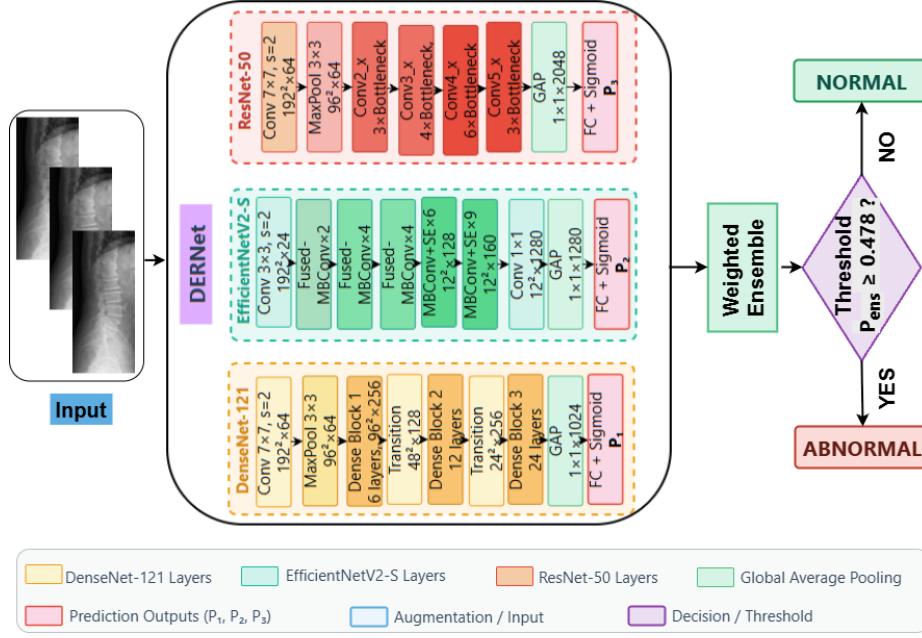


Fig. 3: **The DERNet Ensemble.** A weighted soft-voting framework synergizing DenseNet121, EfficientNetV2-S, and ResNet50 for high-sensitivity spinal triage.

3.3 Fine-Grained Localization (YOLO11-L)

We deploy YOLO11-L with a CSP-Darknet backbone and C2PSA attention to resolve minute lesions (< 1% FOV). The C2PSA block computes scaled dot-product attention (Eq. 3) to enhance discriminability, while PANet preserves high-frequency spatial features critical for stenosis detection.

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V. \quad (3)$$

Table 1: **End-to-End Pipeline Logic.** Hierarchical optimization for DERNet ensemble, mosaic-augmented detection, and clinical inference.

Algorithm 1: End-to-End Pipeline Logic

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, DERNet Ensemble \mathcal{E} , Detector \mathcal{Y}
Hyperparameters: $\eta = 1e^{-4}$ (LR), $\tau_{mosaic} = 25$ (Augmentation Cutoff)

Phase I: DERNet Optimization

```
[1]   for  $m \in \{1, \dots, 3\}$  do
[2]      $\theta_m \leftarrow \theta_{ImageNet}$  // Transfer Learning Initialization
[3]     while not converged do
[4]        $\mathcal{L}_{CE} \leftarrow -\sum_i [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$ 
[5]        $\theta_m \leftarrow \theta_m - \eta \nabla_{\theta_m} \mathcal{L}_{CE}$  // AdamW Update
[6]     end while
[7]   end for
```

Phase II: Detection with Mosaic Decay

```
[8]   for epoch  $t = 1 \rightarrow T_{max}$  do
[9]     if  $t < \tau_{mosaic}$  then
[10]        $x' \leftarrow MosaicAug(x) \sim U(0.3W, 0.7W)$  // Synth. context
[11]     else
[12]        $x' \leftarrow x$  // Disable Mosaic for distribution alignment
[13]     end if
[14]     Update  $\theta_Y$  minimizing  $\mathcal{L}_{total}$  (Eq. 4)
[15]   end for
```

Phase III: Clinical Inference Logic

```
[16]   Score  $\mathcal{S} \leftarrow \sum_{m=1}^3 \omega_m P_m(x)$ 
[17]   if  $\mathcal{S} > \tau_{optimal}$  then
[18]      $\mathcal{B} \leftarrow \text{NMS}(\mathcal{Y}(x), \text{IoU}_{thresh} = 0.5)$  // Filter Overlaps
[19]     return  $\mathcal{B}$ 
[20]   else
[21]     return  $\emptyset$  (Normal Study)
[22]   end if
```

Loss Optimization. The network minimizes a composite objective \mathcal{L}_{total} (Eq. 4), balancing geometric alignment ($\lambda_{box} = 7.5$), classification ($\lambda_{cls} = 0.5$), and dis-

tribution precision ($\lambda_{dfl} = 1.5$):

$$\mathcal{L}_{total} = \lambda_{box}\mathcal{L}_{CIoU} + \lambda_{cls}\mathcal{L}_{Focal} + \lambda_{dfl}\mathcal{L}_{DFL}. \quad (4)$$

Precise localization utilizes Complete-IoU (Eq. 5) to penalize geometric variance, where ρ is center distance and v is aspect ratio consistency:

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha v. \quad (5)$$

Simultaneously, Focal Loss (Eq. 6) addresses class imbalance ($\gamma = 2.0$), complemented by Distribution Focal Loss (\mathcal{L}_{DFL}) for ambiguous boundaries:

$$\mathcal{L}_{Focal} = -\alpha_t(1 - p_t)^\gamma \log(p_t). \quad (6)$$

Clinical Deployment. To operationalize the logic defined in Algorithm 1, we engineered a -optimized inference engine using ONNX Runtime. This system provides modality-agnostic support (DICOM, PNG) and ensures sub-second diagnostic throughput on standard clinical workstations, effectively bridging the gap between research benchmarks and point-of-care utility.

4 Results & Discussions

In this section, we present a comprehensive evaluation of the proposed framework on the VinDr-SpineXR dataset. We first detail the experimental protocols and evaluation metrics, followed by a quantitative analysis of the DERNet classification performance and YOLO11-L localization precision. Finally, we provide qualitative visualizations to demonstrate the model’s clinical applicability.

4.1 Experimental Setup

We applied VinDr-SpineXR protocols [9] ($N = 2,078$) with 5-Fold Stratified Cross-Validation. Implementation: PyTorch 2.0.1, NVIDIA RTX 3050 (8GB). Reproducibility: fixed seeds ($seed = 42$) and deterministic algorithms. Training: AdamW optimization. Classification: 60 epochs (Cosine Annealing); YOLO11-L: 55 epochs (Batch 12, Mosaic Augmentation). Time: ≈ 45 h. Validation: Bootstrap resampling ($B = 1000$, 95% CI) and paired t-tests ($p < 0.05$).

4.2 Experimental results

Classification Performance. Table 2 benchmarks our framework against established baselines [9,6]. While individual backbones exhibited distinct biases EfficientNetV2-S favoring specificity (91.12%) versus ResNet50’s recall priority—our Weighted Soft-Voting strategy successfully harmonized these representations. Consequently, DERNet achieved a state-of-the-art AUROC of 91.03%, significantly outperforming the VinDr-SpineXR baseline (88.61%) while maintaining a robust clinical equilibrium (Sensitivity: 84.91%, Specificity: 81.68%).

Table 2: Classification performance comparison. Best results are in **bold**.

Method	AUROC (%)	F1 (%)	Sens. (%)	Spec. (%)
EfficientNetV2-S	89.44	79.34	70.80	91.12
ResNet50	88.88	80.15	82.72	78.13
DenseNet121	86.93	79.55	80.39	79.32
VinDr Ensemble [9]	88.61	81.06	83.07	79.32
HealNNet [6]	88.84	81.20	-	-
DERNet	91.03	83.09	84.91	81.68

Detection Performance. Table 3 benchmarks the proposed YOLO11-L framework against incumbent models using mean Average Precision (mAP) at an IoU threshold of 0.5. Our architecture establishes a new localization standard, achieving a mAP@0.5 of 40.10% and demonstrating robust efficacy across heterogeneous spinal lesions. Notably, the model excels in fine-grained categories, attaining an AP of 41.40% for Foraminal Stenosis (LT4). These results validate the capacity of the optimized CSP-Darknet backbone to preserve high-frequency anatomical features essential for small-object localization.

Table 3: Comparison results of different methods. The detection performance for each type of spinal lesion is evaluated by AP (%) at 0.5 IoU threshold. Best results are in **bold**.

Method	LT2 ^(*)	LT4	LT6	LT8	LT10	LT11	LT13	mAP@0.5
Dino [19]	16.58	22.87	28.53	32.71	59.78	41.28	3.24	29.28
RetinaNet [8]	14.53	25.35	41.67	32.14	65.49	51.85	5.30	28.09
Faster R-CNN [13]	22.66	35.99	49.24	31.68	65.22	51.68	2.16	31.83
Sparse R-CNN [15]	20.09	32.67	48.16	45.32	72.20	49.30	5.41	33.15
VinDr-SpineXR [9]	21.43	27.36	34.78	41.29	62.53	43.39	4.16	33.56
EGCA-Net [2]	22.36	29.75	36.73	44.69	66.58	50.41	2.09	36.09
Ours (YOLO11-L)	26.70	41.40	40.60	54.80	74.10	51.20	2.99	40.10

^(*) LT2, LT4, LT6, LT8, LT10, LT11, LT13 denotes for disc space narrowing, foraminal stenosis, osteophytes, spondylolisthesis, surgical implant, vertebral collapse and other lesions, respectively, following the same indexing in Table 2.

Qualitative Analysis. Fig. 4 validates the clinical relevance of DERNet. LIME (a) confirms the model’s focus on *meaningful structures* rather than background artifacts. Grad-CAM (b) ensures *transparent validation*, with heatmaps aligning precisely to pathological lesions. Finally, Segmentation (c) demonstrates superior *anatomical fidelity*, preserving fine details often missed by baseline models.

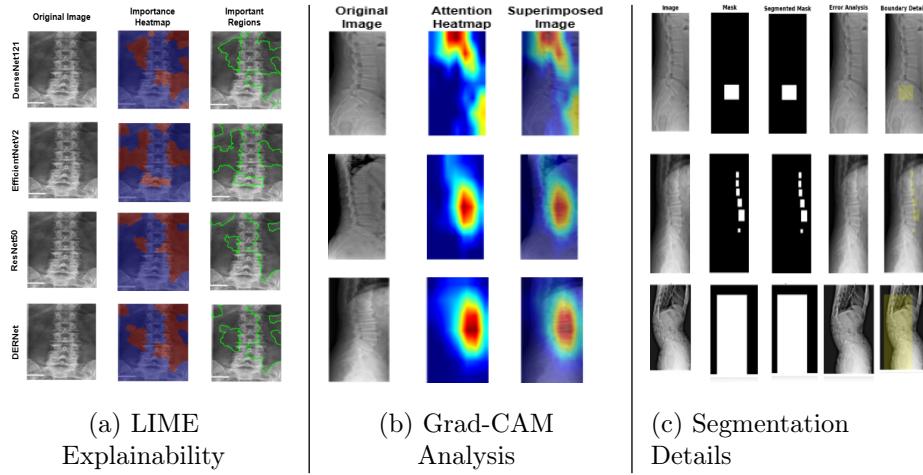


Fig. 4: Performance Evaluation. (a) **Explainability:** LIME confirms focus on meaningful structures. (b) **Transparency:** Grad-CAM heatmaps validate lesion localization. (c) **Fidelity:** Segmentation masks show high overlap with ground truth.

4.3 Discussion

Our results highlight the efficacy of integrating diverse inductive biases within a hierarchical decision framework. The Weighted DERNet effectively resolves the sensitivity-specificity trade-off by fusing EfficientNetV2-S’s high specificity with ResNet50’s robust recall, establishing a clinical equilibrium superior to single-stream baselines. Regarding localization, the CSP-Darknet backbone demonstrates superior feature preservation compared to explicit edge priors (e.g., EGCA-Net), evidenced by a +11.65% AP gain in detecting subtle Foraminal Stenosis. Crucially, the pipeline functions as a conditional hard-attention gate; by leveraging high-specificity screening (81.68%) to preemptively filter normal studies, we significantly mitigate detector false positives, driving the system-wide mAP@0.5 to 40.10%.

5 Conclusion

This study addressed the critical bottleneck in automated spinal analysis: the trade-off between high-sensitivity screening and the precise localization of subtle pathologies. By synergizing a probabilistic DERNet ensemble with a specialized YOLO11-L detector, we successfully established a unified framework that overcomes severe class imbalance without compromising diagnostic precision. Empirical validation on the VinDr-SpineXR benchmark confirms that exploiting architectural heterogeneity significantly enhances triage performance (AUROC

91.03%), while the optimized CSP-Darknet backbone proves essential for resolving minute lesions like Foraminal Stenosis (mAP@0.5 40.10%). Beyond algorithmic advancements, the successful deployment of our ONNX-integrated web interface bridges the translational gap, offering a scalable, low-latency solution to augment radiologist decision-making in real-world clinical workflows.

References

1. Cheng, L.W., Chou, H.H., Cai, Y.X., et al.: Automated detection of vertebral fractures from x-ray images: A novel machine learning model and survey of the field. *Neurocomputing* **566**, 126946 (2024)
2. Guo, L., Peng, B., Lei, J., Zhang, X., Zhao, J., Huang, Q.: Spinal lesion detection in X-Ray images via edge guidance and context aggregation. *IEEE Transactions on Instrumentation and Measurement* **74**, 1–11 (2025). <https://doi.org/10.1109/TIM.2025.3527484>
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
5. Huang, J., Zhu, X., Chen, Z., Lin, G., Huang, M., Feng, Q.: Pathological priors inspired network for vertebral osteophytes recognition. *IEEE Transactions on Medical Imaging* **43**(7), 2522–2536 (2024)
6. Khan, A., Gupta, S., Arsalan, M.: Healnnet-lesions: A deep learning framework for spinal lesion detection demonstrating x-ray images as a screening alternative. *Preprints* (2025). <https://doi.org/10.20944/preprints202510.0094.v1>
7. Krupinski, E.A., Berbaum, K.S., Caldwell, R., Schartz, K., Kim, J.: Long radiology workdays reduce detection and accommodation accuracy. *Journal of the American College of Radiology* **7**(9), 698–704 (2010)
8. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
9. Nguyen, H.T., Pham, H.H., Nguyen, N.T., Nguyen, H.Q., Huynh, T.Q., Dao, M., Vu, V.: Vindr-spinexr: A deep learning framework for spinal lesions detection and classification from radiographs. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 291–301. Springer (2021)
10. Pham, H.H., Nguyen Trung, H., Nguyen, H.Q.: VinDr-SpineXR: A large annotated medical image dataset for spinal lesions detection and classification from radiographs. *PhysioNet* (Aug 2021). <https://doi.org/10.13026/q45h-5h59>, <https://doi.org/10.13026/q45h-5h59>, version 1.0.0
11. Pinto, A., Beritto, D., Russo, A., Ricciello, F., Caruso, M., Belfiore, M.P., Papapietro, V.R., Carotti, M., Pinto, F., Giovagnoni, A., et al.: Traumatic fractures in adults: missed diagnosis on plain radiographs in the emergency department. *Acta Bio Medica: Atenei Parmensis* **89**(Suppl 1), 111 (2018)
12. Priolo, F., Cerase, A.: The current role of radiography in the assessment of skeletal tumors and tumor-like lesions. *European Journal of Radiology* **27**, S77–S85 (1998)
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **28** (2015)

14. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. *Annual review of biomedical engineering* **19**, 221–248 (2017)
15. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14454–14463 (2021)
16. Sutradhar, D., et al.: Cervical spine fracture detection utilizing yolov8 and deep attention-based vertebrae classification ensuring xai. *Biomedical Signal Processing and Control* **101**, 107228 (2025)
17. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: *International conference on machine learning*. pp. 10096–10106. PMLR (2021)
18. Zhang, D., Chen, B., Li, S.: Sequential conditional reinforcement learning for simultaneous vertebral body detection and segmentation with modeling the spine anatomy. *Medical Image Analysis* **67**, 101861 (2021). <https://doi.org/10.1016/j.media.2020.101861>
19. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: *International Conference on Learning Representations (ICLR)*. pp. 1–12 (2023)