A Cascaded DERNet and YOLO11 Framework for Spinal Lesion Triage and Localization

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Automated analysis of spinal radiographs is critical for early diagnostic triage but remains severely constrained by the visual subtlety of minute lesions. The purpose of this study is to improve screening sensitivity and localization specificity by engineering a unified, cascaded artificial intelligence framework. For this methodology, we decouple the diagnostic workflow into two hierarchical stages: First, we use DERNet, which is a combination of EfficientNetV2-S, DenseNet121, and ResNet50 - a highly sensitive binary triage and secondly, routing abnormal cases to a customized YOLO11-L detector to detect precisely. According to the experimental results on the VinDr-SpineXR benchmark, our approach is more effective, achieving AUROC of 91.03% for the image-level classification task and a mAP@0.5 of 40.10% for the lesion-level localization task. This research is necessary and beneficial to the medical imaging research community, as it is through LIME, Grad-CAM and Qualitative Visualization that the framework ensures diagnostic transparency, bridging the critical gap between high-performance deep learning and trustworthy clinical deployment. All code associated with this study is available at: GitHub.

Keywords: Spine X-rays, Radiographs, DERNet, Explainability, YOLO11

1 Introduction

Spinal radiographs serve as the primary screening modality for evaluating a wide spectrum of spinal pathologies [1,2]. However, the manual diagnostic workflow is intrinsically challenging; radiologists must identify visually subtle, minute lesions as illustrated in Fig. 1 that are frequently obscured by the complex anatomical superposition of dense, three-dimensional bone structures onto a two-dimensional plane [3,4]. This visual ambiguity not only induces high inter-observer variability but also elevates the risk of delayed diagnoses, which can precipitate irreversible neurological sequelae [5,6]. Consequently, there is an urgent clinical mandate to develop robust, high-precision automated diagnostic frameworks to standardize radiographic triage and mitigate human perceptual errors.

Deep Learning fundamentally transforms medical image analysis by extracting high-dimensional, non-linear features from complex radiographs [7,8]. In

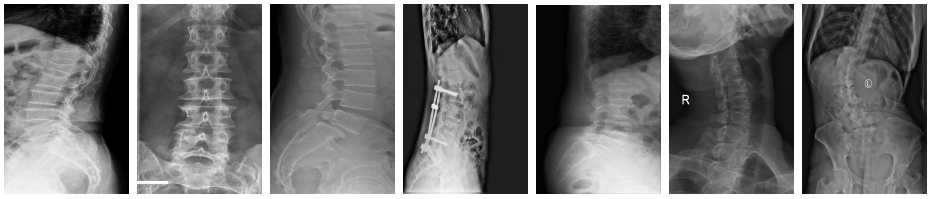


Fig. 1: Representative examples of spinal radiographs from the VinDr-SpineXR benchmark. The figure illustrates the heterogeneity and visual subtlety of seven lesion categories: (a) Vertebral Collapse, (b) Osteophytes, (c) Spondylolisthesis, (d) Surgical Implants, (e) Disc Space Narrowing, (f) Foraminal Stenosis, and (g) Other Lesions. Many lesions occupy less than 1% of the field-of-view and are often confounded by overlapping anatomical structures. This variability and small object scale exacerbate class imbalance (46.9:1 ratio) and motivate the need for a cascaded triage–localization framework.

spinal lesions, image-based AI provides objective, high-throughput triage that significantly reduces radiologist fatigue. More importantly, advanced convolutional architectures can isolate early-stage, visually ambiguous structural anomalies that frequently evade human visual perception.

Despite these advancements, clinical AI adoption is severely hindered by long-tailed pathological distributions and the persistent sensitivity-specificity trade-off [2]. Traditional monolithic detectors [9,10] fail to account for this statistical skew, leading to excessive false positives or missed diagnoses. To resolve this, we propose a unified, cascaded AI framework that structurally decouples the diagnostic workflow into two hierarchical phases: first, DERNet performs highly sensitive binary triage; second, abnormal radiographs are dynamically routed to a customized YOLO11-L detector to localize fine-grained spinal pathologies precisely. This approach effectively neutralizes data skewness without compromising computational efficiency.

Our primary contributions to automated spinal radiograph analysis are as follows:

(a) Our proposed DERNet, which is a combination of DenseNet121 [11], EfficientNetV2-S [12], and ResNet50 [13] to achieving a screening AUROC of 91.03%.
(b) We localize the spinal cord lesion level using a customized YOLO11-L detector (CSP-Darknet with C2PSA) director. Through which we achieve mAP@0.5 of 40.10%.
(c) We deploy a latency-optimized, ONNX-accelerated inference engine integrating LIME, Grad-CAM and Qualitative Visualization mapping for transparent, clinically verifiable decision support.

2 Related Work

Deep learning [7] drives automated spinal analysis, yet robust clinical adoption remains limited by architectural constraints. Highlighting foundational contri-

butions, Nguyen et al. [2] and Pham et al. [14] established the *VinDr-SpineXR* benchmark and proposed baseline classifiers. However, their models exhibited severe sensitivity constraints on minority classes, compromising clinical triage. To address this, Upadhyay [15], Turk et al. [16], and Sutradhar et al. [17] proposed specialized deep learning frameworks. Despite improving feature extraction, these works present critical research gaps: Upadhyay's *HealNNet* [15] incurs prohibitive computational overhead in high-throughput settings, while Turk et al. and Sutradhar et al. are strictly confined to fracture detection, lacking the multi-class capabilities required for comprehensive triage. We resolve this by synergizing DenseNet, EfficientNet, and ResNet into our DERNet ensemble to maximize screening performance with parameter efficiency.

For lesion localization, extreme class imbalance (e.g., 46.9:1) and minute pathology scales severely curtail diagnostic efficacy. Addressing this, Ren et al. [9], Sun et al. [10], and Guo et al. [18] proposed diverse object detectors, ranging from conventional two-stage models to the edge-guided *EGCA-Net*[18]. A major limitation of these architectures is their failure to resolve subtle fine-grained features; they remain highly susceptible to background artifacts, frequently misidentifying overlapping normal anatomy as pathological lesions. We advance this domain by proposing an optimized YOLO11-L equipped with a CSP-Darknet backbone and C2PSA attention, explicitly engineered to suppress artifacts, neutralize severe class imbalance, and precisely localize minute pathologies.

3 Methodology

As illustrated in Fig. 2, our cascaded framework structurally decouples the diagnostic workflow to resolve the sensitivity-specificity trade-off. Specifically, a probabilistic DERNet ensemble first performs high-sensitivity binary triage to filter normal studies, dynamically routing only the abnormal candidates to a customized YOLO11-L detector for precise lesion localization. This architectural isolation explicitly neutralizes false-positive amplification and mitigates severe pathological class imbalance.

3.1 Dataset and Adaptive Preprocessing

We leveraged the *VinDr-SpineXR*[14] benchmark, utilizing 8,389 images for training and 2,077 for testing, 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 the enhanced intensity $I'(i,j)$ by mapping the local Cumulative Distribution Function (CDF), as formulated in Eq. 1:

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

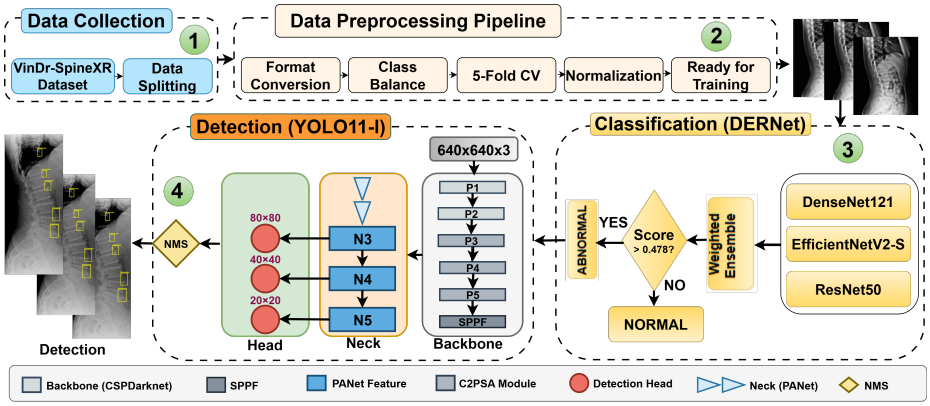


Fig. 2: Overview of the cascaded framework for spinal lesion triage and localization.

where Eq. 1 utilizes $|\Omega_k|$ to denote local region cardinality and β as the dynamic range factor. Furthermore, 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 introduce DERNet, a heterogeneous ensemble engineered as a high-sensitivity triage gate.

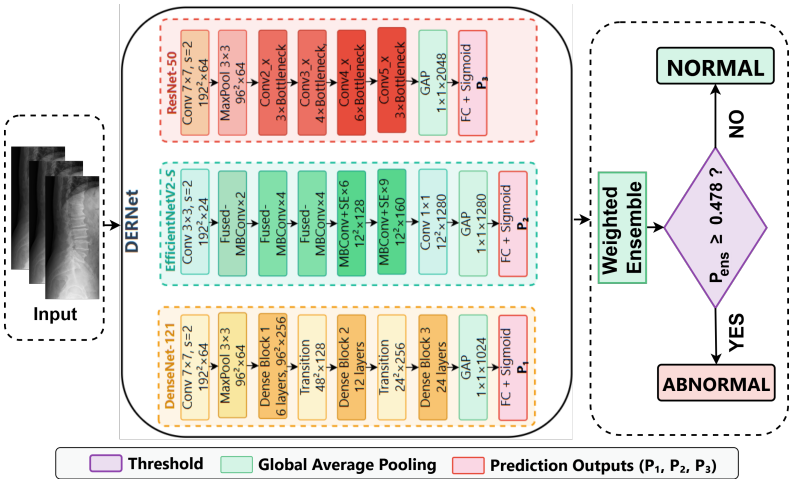


Fig. 3: DERNet probabilistic ensemble architecture for high-sensitivity radiographic triage. Each backbone (DenseNet121, EfficientNetV2-S, ResNet50) produces independent posterior probabilities via Global Average Pooling and Sigmoid activation.

As detailed in Fig. 3, the architecture synthesizes three distinct architectural paradigms to identify heterogeneous spinal pathologies.

(a) **DenseNet121:** This is based on the ensemble's textural sensitivity, which has the highest fusion weight $\omega_1 = 0.42$. Making multi-scale feature reuse, by dense block concatenation, active, it empirically shows that it is essential to the isolation of visually subtle anomalies like marginal osteophytes.
(b) **EfficientNetV2-S:** It is the main specificity driver, where $\omega_2 = 0.32$, of the ensemble, which uses the squeezing-and-excitation (SE) modules to recalibrate channel-wise features dynamically. This mechanism specifically inhibits non-diagnostic background noise and anatomical overlap, providing a clean signal-to-noise ratio to eliminate false positives with mid-scale lesions.
(c) **ResNet50:** This serves as our recall fail-safe with a weight of $\omega_2 = 0.26$. This particular ability makes the ensemble sensitive to gross macroscopic dislocations, including vertebral body collapse or spondylolisthesis, since only the large-scale spatial orientation is vital for diagnosing these conditions.

Following Global Average Pooling and Sigmoid activation, each backbone generates an independent posterior $P(y = c|\mathbf{x}; \theta_m)$. These diverse inductive biases are fused via Weighted Soft-Voting as shown in Eq. 2 to marginalize uncertainty:

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

where weights $\mathbf{\Omega} = [0.42, 0.32, 0.26]$ prioritize DenseNet's textural sensitivity. As shown in Fig. 3, studies scoring $P_{ens} \geq 0.478$ are flagged as Abnormal; others are efficiently filtered as Normal.

3.3 Fine-Grained Localization (YOLO11-L)

Radiographs flagged as pathological are dynamically routed to our localization engine, resolving the sensitivity-specificity trade-off of monolithic architectures [9,10]. As shown in Fig. 2, we engineered an optimized YOLO11-L detector utilizing three integrated blocks to prioritize minute lesions ($< 1\%$ FOV):

(a) **Backbone** (CSP-Darknet with C2PSA): This block utilizes Cross-Stage Partial connections and C2PSA attention to preserve high-frequency anatomical features. By computing scaled dot-product attention in Eq. 3, it enhances pathological discriminability and suppresses background artifacts, securing a state-of-the-art AP of 41.40% for subtle Foraminal Stenosis.
(b) **Neck** (PANet Feature Fusion): The Path Aggregation Network facilitates bidirectional feature flow across multi-scale feature maps ($P1$–$P5$). This block ensures that fine-grained pathology signals are preserved across deep layers, neutralizing the severe class imbalance inherent to spinal benchmarks.
(c) **Head** (Multi-Scale Detection): The localized prediction head employs three distinct spatial resolutions $80 \times 80, 40 \times 40$ and 20×20 coupled with Non-Maximum Suppression. This hierarchical detection strategy optimizes bounding box precision for lesions occupying $< 1\%$ of the field of view.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad \mathcal{L}_{total} = \lambda_{box}\mathcal{L}_{CIoU} + \lambda_{cls}\mathcal{L}_{Focal} + \lambda_{dfl}\mathcal{L}_{DFL} \tag{3}$$

The network minimizes a composite objective $\mathcal{L}_{total}$ as defined in Eq. 3, balancing geometric alignment ($\mathcal{L}_{CIoU}$), class-imbalance mitigation ($\mathcal{L}_{Focal}$), and boundary precision ($\mathcal{L}_{DFL}$). This attention-guided isolation of overarching screening from fine-grained detection explicitly neutralizes false-positive amplification while maintaining parity with production-validation metric.

3.4 Clinical Deployment

To implement the DERNet and YOLO11-L as a viable triage tool, we bundled the full inference pipeline into a containerized microservice. The system is configured to accept raw hospital data directly; it automatically handles JPEG, PNG or DICOM-native preprocessing and adjusts photometric interpretations in real-time and sends the scans to a PyTorch backend via Gunicorn. We heavily optimized the pipeline to run on a CPU-only system. It produces a complete radiograph and maintains the 91.03% AUROC from the offline validation.

4 Results & Discussions

In this section, we provide a multi-dimensional evaluation of the proposed cascaded framework using the VinDr-SpineXR benchmark. Finally, we conduct a comparative study and provide explainability via LIME, Grad-CAM and Qualitative Visualization to validate the clinical reliability of our findings.

4.1 Experimental Setup

The proposed framework was tested on VinDr-SpineXR benchmark data, after 5-fold stratified cross-validation procedure to guarantee that the results of the research apply to a heterogeneous population with spinal pathologies. This system was run in PyTorch 2.0.1 and on a workstation with an NVIDIA RTX 3050 GPU (8GB). Deterministic algorithms were imposed to ensure reproducibility, as well as the absence of stochastic variation, and a constant initialisation seed of 42.

4.2 Experimental results

Table 1 benchmarks our framework against established baselines [2]. 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 VinD-SpineXR [2] baseline of 88.61% while maintaining robust clinical performance: Sensitivity: 84.91%, Specificity: 81.68%.

Table 1: 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 [2]	88.61	81.06	83.07	79.32
HealNNet [15]	88.84	81.20	-	-
DERNet	**91.03**	**83.09**	**84.91**	**81.68**

Table 2 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 optimized CSP-Darknet backbone's ability to preserve high-frequency anatomical features essential for small-object localization.

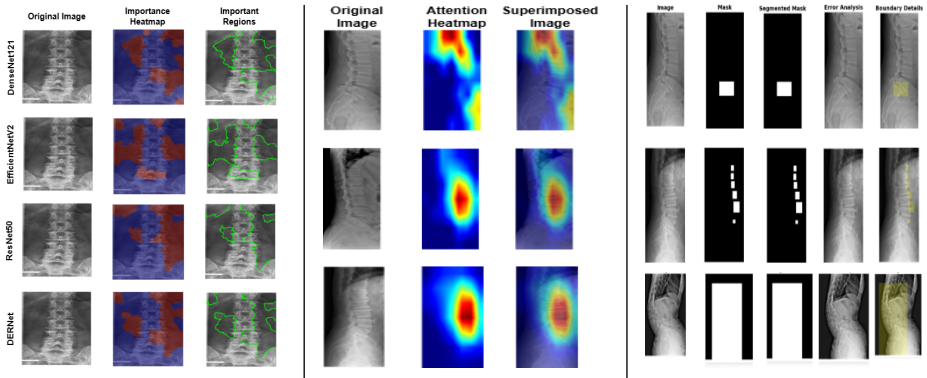
Table 2: Comparison results of different methods. The detection performance for each type of spinal lesion is evaluated by AP (%) at 0.5 IoU threshold. Best results are in **bold**.

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 [20]	14.53	25.35	41.67	32.14	65.49	51.85	5.30	28.09
Faster R-CNN [9]	22.66	35.99	49.24	31.68	65.22	51.68	2.16	31.83
Sparse R-CNN [10]	20.09	32.67	48.16	45.32	72.20	49.30	5.41	33.15
VinDr-SpineXR [2]	21.43	27.36	34.78	41.29	62.53	43.39	**4.16**	33.56
EGCA-Net [18]	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, spondylolysthesis, surgical implant, vertebral collapse and other lesions, respectively, following the same indexing in Table 2.

As illustrated in Fig. 4, the clinical relevance and diagnostic reliability of the architecture are validated through a multi-modal explainability suite:

(a) **LIME:** It testifies that the ensemble is more concerned with the authentic anatomical landmarks than not. The local importance heatmaps support the claim that the predictive weight is directly ascribed to diagnostic areas, e.g., vertebral boundaries.
(b) **Grad-CAM:** It confirms the transparency of the decision-making of the framework. The heatmaps show consistent localization with the pathology lesions, and therefore, the high sensitivity used in triage has a solid foundation in the visual representation.



(a) LIME (b) Grad-CAM (c)Qualitative Visu.

Fig. 4: Performance Evaluation

(c) **Qualitative Visualization:** It achieves a high-resolution boundary across complex spinal anatomies, isolating microstructural deformations by preserving high-frequency spatial gradients that are degraded by baselines.

Discussion: Our proposed DERNet filters out normal studies, dynamically routing only the abnormal candidates to a customized YOLO11-L detector for precise lesion detection. The integration of explainability tools further confirms that the framework's decision-making is deeply rooted. Furthermore, optimizing spatial feature preservation proves critical for isolating visually ambiguous, minute pathologies. Future directions include multi-center validation for demographic generalization and integrating longitudinal radiographic analysis to monitor disease progression.

5 Conclusion

This paper helps fill the gap between high-sensitivity screening and accurate localization of automated spinal diagnostics. Our probabilistic DERNet ensemble for initial high-sensitivity screening dynamically routes abnormal cases to an optimized YOLO11-L detector for fine-grained localization. It structurally separates the diagnostic process, eliminating class imbalances and preventing the amplification of false positives. Clinically, this methodology is superior, as it actively suppresses non-diagnostic background noise. Furthermore, integrating explainability tools ensures that the framework's decisions are transparently grounded in genuine anatomical features, establishing the necessary interpretability and trust for real-world clinical deployment. The drawbacks are the use of a single-center dataset and the lack of longitudinal patient history. Future studies will therefore focus more on multicenter validation to facilitate generalization of the findings to a broader population and on longitudinal radiographic surveillance to monitor the disease course proactively.

References

1. Dong Zhang, Bo Chen, and Shuo Li. Sequential conditional reinforcement learning for simultaneous vertebral body detection and segmentation with modeling the spine anatomy. *Medical Image Analysis*, 67:101861, 2021.
2. Hieu T Nguyen, Hieu H Pham, Nghia T Nguyen, Ha Q Nguyen, Thang Q Huynh, Minh Dao, and Van Vu. 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*, pages 291–301. Springer, 2021.
3. Francesco Priolo and Alfonso Cerase. The current role of radiography in the assessment of skeletal tumors and tumor-like lesions. *European Journal of Radiology*, 27:S77–S85, 1998.
4. Elizabeth A Krupinski, Kevin S Berbaum, RT Caldwell, KM Schartz, and J Kim. Long radiology workdays reduce detection and accommodation accuracy. *Journal of the American College of Radiology*, 7(9):698–704, 2010.
5. Antonio Pinto, Daniela Berritto, Anna Russo, Federica Riccitiello, Martina Caruso, Maria Paola Belfiore, Vito Roberto Papapietro, Marina Carotti, Fabio Pinto, Andrea Giovagnoni, 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.
6. Jun Huang, Xinyu Zhu, Zhuo Chen, Guudan Lin, Meiyan Huang, and Qianjin Feng. Pathological priors inspired network for vertebral osteophytes recognition. *IEEE Transactions on Medical Imaging*, 43(7):2522–2536, 2024.
7. Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.
8. Li-Wei Cheng, Hsin-Hung Chou, Yu-Xuan Cai, 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.
9. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
10. Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, 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*, pages 14454–14463, 2021.
11. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
12. Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.
13. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
14. Hieu Huy Pham, Hieu Nguyen Trung, and Ha Quy Nguyen. VinDr-SpineXR: A large annotated medical image dataset for spinal lesions detection and classification from radiographs. *PhysioNet*, August 2021. Version 1.0.0.
15. Visharad Upadhyay. Healnnet-lesions: A deep learning framework for spinal lesion detection demonstrating x-ray images as a screening alternative to mri and ct with deep learning in radiology. 2025.

16. Salih Turk, Ozkan Bingol, Ahmet Coskuncay, and Tolga Aydin. The impact of implementing backbone architectures on fracture segmentation in x-ray images. *Engineering Science and Technology, an International Journal*, 59:101883, 2024.
17. Debopom Sutradhar, Nur Mohammad Fahad, Mohaimenul Azam Khan Raiaan, Mirjam Jonkman, and Sami Azam. Cervical spine fracture detection utilizing yolov8 and deep attention-based vertebrae classification ensuring xai. *Biomedical Signal Processing and Control*, 101:107228, 2025.
18. Lisha Guo, Bo Peng, Jianjun Lei, Xu Zhang, Jun Zhao, and Qingming Huang. Spinal lesion detection in X-Ray images via edge guidance and context aggregation. 74:1–11, 2025.
19. Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Harry Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, pages 1–12, 2023.
20. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.