Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

OrthoScan AI: A Unified, Interpretable Deep-Learning Pipeline for Scoliosis Severity and Knee Osteoarthritis Grading from Plain Radiographs on Commodity CPUs

Khushikumari Parmanand Poddar
Dept. of Computer Engineering Pune Vidyarthi Griha's College of Engineering and Technology Nashik, Maharashtra, India

Rohan Laxman Shinde
Dept. of Computer Engineering Pune Vidyarthi Griha's College of Engineering and Technology Nashik, Maharashtra, India

Vishakha Sanjay Kamble
Dept. of Computer Engineering Pune Vidyarthi Griha's College of Engineering and Technology Nashik, Maharashtra, India

Yash Jayant Sonar
Dept. of Computer Engineering Pune Vidyarthi Griha's College of Engineering and Technology Nashik, Maharashtra, India

Shubham Bhausaheb Dighe*
Civora Nexus
*Technical guide and supervisor; corresponding author

Pradeep Patil
Dept. of Computer Engineering Pune Vidyarthi Griha's College of Engineering and Technology Nashik, Maharashtra, India

***Abstract*—Scoliosis and knee osteoarthritis (OA) are among the most frequently radiographed musculoskeletal conditions, yet both are still graded by eye: the Cobb angle for spinal curvature and the Kellgren–Lawrence (KL) scale for knee OA carry well-documented inter-observer variability, and specialist readers are scarce outside large hospitals. We present OrthoScan AI, a single screening pipeline that accepts an unlabeled radiograph, decides automatically whether it shows a spine or a knee, and returns a clinically anchored severity output—a Cobb-angle band with an angle estimate, or a KL grade—together with a Grad-CAM explanation, on a CPU-only server (0.3–0.4 s per image without, and 0.7–1.2 s with, the explanation on a 12-core laptop). The two diagnostic branches share one preprocessing chain (border cropping, CLAHE, edge-preserving denoising and letterboxing) and one transfer-learning recipe (frozen-trunk warm-up, discriminative learning rates with cosine decay, class-balanced sampling, label smoothing and early stopping on balanced accuracy). We evaluate on three public datasets. On the Osteoarthritis-Initiative-derived KL-graded knee set of Chen *et al.* (8,260 images, fixed split) a single DenseNet-121 reaches $62.3 \pm 0.6\%$ five-class accuracy, $67.0 \pm 0.9\%$ balanced accuracy and a quadratic-weighted κ of 0.79 over three seeds, with 94% of predictions within one grade and KL-4 recall of 0.88; a pre-declared soft-vote ensemble of the nine trained knee networks raises this to 67.6% accuracy, 71.0% balanced accuracy and $\kappa = 0.84$ (McNemar $p < 10^{-4}$), and an ablation with paired tests shows that balanced sampling alone recovers minority-grade recall while stacking it with loss re-weighting over-corrects. On 338 real clinical spine radiographs the system detects scoliosis with 96.8% sensitivity (pooled five-fold accuracy 95.9%, AUC 0.986). On the 20,000-image Spinal-AI2024 synthetic set the learned angle head estimates the major Cobb angle with a mean absolute error of 2.6° (89% within 5°), whereas the weights-free geometric tracer the system uses as a confidence-gated fallback reaches only 11.6°, so we characterise it as a coarse screening signal rather than a measurement. Robustness to compression, noise, blur and exposure, Grad-CAM plausibility and per-stage latency are reported. Evaluation scripts, trained weights and per-image predictions are available on request. The system is screening support, not a diagnosis, and every output carries that disclaimer.**

***Index Terms*—scoliosis, Cobb angle, knee osteoarthritis, Kellgren–Lawrence grading, transfer learning, convolutional neural networks, Grad-CAM, computer-aided diagnosis, radiography**

I. Introduction

Plain radiography remains the first-line examination for two of the commonest orthopaedic conditions. Adolescent idiopathic scoliosis affects roughly 0.5–5% of adolescents depending on the curve threshold used [1], and knee osteoarthritis accounts for a large and growing share of global musculoskeletal disability [2]. In both cases the radiograph is read by a human and summarised as a single number or grade: the Cobb angle, the angle between the end-plates of the most tilted vertebrae above and below the apex of the curve, and the five-level Kellgren–Lawrence (KL) grade [3], which scores osteophytes, joint-space narrowing and sclerosis. Both summaries are known to vary between readers. Intra- and inter-observer studies of manual Cobb measurement report errors of several degrees [4], [5], and agreement between expert KL readers is only moderate (quadratic-weighted κ of about 0.56–0.67 in the committee study used by Tiulpin *et al.* [6]).

Deep learning has been applied to each task separately with considerable success. For knee OA, convolutional networks trained on the Osteoarthritis Initiative (OAI) reach five-class KL accuracies of roughly 65–77% [6]–[10], close to the human agreement ceiling. For scoliosis, landmark- and segmentation-based networks estimate the Cobb angle with a pooled mean absolute error near 3° in a recent meta-analysis [11]. What is still rare is a *deployable* system that

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

handles both conditions behind one interface, reports results on the clinical scales clinicians already use, explains itself, and runs without a GPU—the configuration a district hospital or screening camp can actually afford.

This paper describes and evaluates such a system. Our contributions are:

1) **A unified two-branch screening pipeline.** A lightweight region router selects the spine or knee branch automatically; both branches share one preprocessing chain and one training recipe, and both output grades on the clinical scale (Cobb band and angle; KL 0–4) with Grad-CAM [12] explanations.
2) **A CPU-feasible transfer-learning recipe with honest reporting.** We report balanced accuracy, macro-F1, quadratic-weighted κ, calibration error and per-grade recall on fixed public splits, with bootstrap confidence intervals, paired McNemar tests, ablations of the class-imbalance components, repeated seeds, and a pre-declared ensemble of the trained networks.
3) **A quantified weights-free Cobb fallback.** The system traces the spinal midline geometrically when no trained weights are available, and uses the same trace to flag disagreement with the learned estimate. We measure this estimator against 4,000 ground-truth angles and show when its self-reported confidence can be trusted.
4) **Robustness and latency characterisation.** We measure degradation under JPEG compression, noise, blur and exposure changes, and per-stage serving latency on a 12-core CPU.

II. Related Work

A. Knee osteoarthritis grading

Antony *et al.* [7] first showed that CNN features outperform hand-crafted descriptors for KL grading. Tiulpin *et al.* [6] trained a Siamese network on the symmetric knee halves and reported 66.7% accuracy and $\kappa = 0.83$ on OAI, together with attention maps. Chen *et al.* [8] combined YOLO-based joint detection with an ordinal loss and released the KL-graded crops we use [13]; on that split they report about 69.7% five-class accuracy. Norman *et al.* [9] applied DenseNet, Tiulpin and Saarakkala [14] graded individual OARSI features, and Pi *et al.* [10] reached 76.9% accuracy with an ensemble on the same 8,260-image split. These works establish both the achievable range and the fact that most residual error lies between adjacent grades.

B. Scoliosis and Cobb angle estimation

The AASCE 2019 challenge dataset introduced with BoostNet [15] made landmark-based Cobb estimation a standard benchmark; vertebra-focused keypoint detection [16] and later contour-based methods such as CurvNet [17] are representative. Zhu *et al.* [11] pooled 17 studies and found a circular mean absolute error of 2.99°, with segmentation-based methods (2.40°) slightly ahead of landmark methods (3.31°), but also noted high heterogeneity and little external validation; a recent multi-centre study [18] confirms that most disagreements with radiologists fall between adjacent severity categories. Classification-only approaches that detect scoliosis without measuring it have been trained on smaller clinical collections [19], [20]. Our spine branch sits between these lines of work: it predicts the severity band and a scalar angle directly from the whole image, which is cheaper than landmark detection but, as our results show, also less precise.

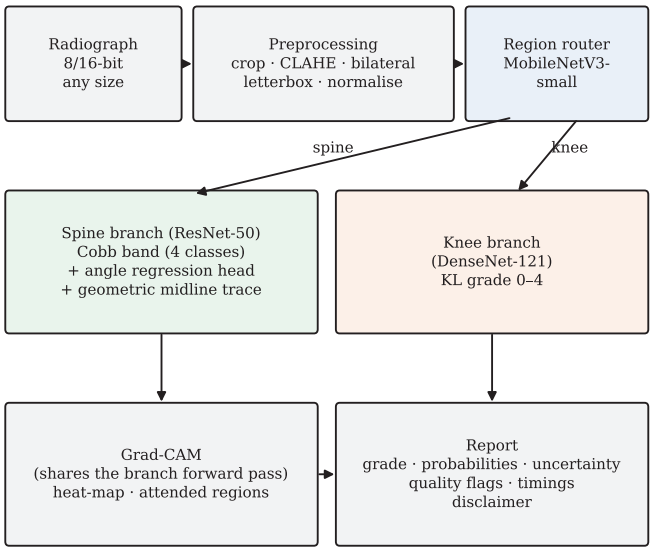


Fig. 1. OrthoScan AI pipeline. One preprocessing chain feeds a region router and two diagnostic branches; Grad-CAM shares the branch's forward pass. The geometric Cobb trace runs in parallel for spine images.

C. Transfer learning and explainability in radiography

Fine-tuning ImageNet-pretrained backbones is the established route for small medical datasets [21], [22]; we use ResNet [23], DenseNet [24], EfficientNet [25] and MobileNetV3 [26] trunks. Grad-CAM [12] remains the most widely used post-hoc explanation in medical imaging [27]; we follow the pointing-game protocol of Zhang *et al.* [28] to check that explanations fall on anatomically plausible regions. To our knowledge no prior work evaluates a single deployable pipeline across both spine and knee radiographs with these outputs.

III. System Design

Fig. 1 summarises the pipeline. All numerical settings below are those of the released code.

A. Preprocessing

Radiographs arrive as 8- or 16-bit, grey or colour files with scanner borders and widely varying exposure. Each image is (i) decoded to a single channel, with 16-bit sources rescaled by their actual intensity range; (ii) cropped to the exposed field found by Otsu thresholding of a blurred copy and the largest external contour, with a 2% margin, unless the detected field is implausibly small ($< 40\%$ of either side); (iii) enhanced by contrast-limited adaptive histogram equalisation (CLAHE,

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

clip limit 2.4, 8×8 tiles) [29]; (iv) denoised with a bilateral filter ($d = 5$, $\sigma_{\text{colour}} = \sigma_{\text{space}} = 45$) [30], which removes sensor grain while keeping the cortical edges that both grades depend on; (v) letterboxed onto a square canvas of side S without changing the aspect ratio, since anisotropic resizing would alter joint-space width and curve geometry; and (vi) standardised per image ($\hat{x} = \text{clip}((x - \mu)/4\sigma + 0.5, 0, 1)$) before ImageNet normalisation. The same code path is used in training, validation and serving, so offline metrics correspond to what the deployed service produces.

B. Region router

A MobileNetV3-small trunk [26] with a two-way head decides whether the image shows a spine or a knee, so that a clinician never has to specify the body part and an image of the wrong region is rejected rather than silently graded. Predictions below a confidence threshold are flagged.

C. Diagnostic branches

Both branches use the same wrapper, *OrthoNet*: a pretrained convolutional trunk with its classifier removed, global average pooling, and a head of dropout (0.3), a 512-unit fully connected layer with batch normalisation and ReLU, dropout (0.18) and a final linear layer. The knee branch has five outputs (KL 0–4). The spine branch has four outputs for the Cobb severity bands used by the Scoliosis Research Society convention—normal ($< 10°$), mild (10–25°), moderate (25–$40°$) and severe ($\geq 40°$)—and an additional regression head (256 units, ReLU, one output with Softplus) that predicts the major Cobb angle in degrees. The last convolutional block of the trunk is the Grad-CAM target layer.

D. Geometric Cobb estimator

Independently of any trained weights, the spine branch traces the spinal midline from the preprocessed image. The image is blurred (Gaussian, $\sigma = 1\%$ of width), the central 60% of columns are kept, and in each of 40 horizontal bands the column centroid of intensities above the band's 75th percentile is taken as the midline point. A fourth-order polynomial $x = f(y)$ is fitted; the tangent angle $\theta(y) = \arctan(f'(y))$ is computed, the top and bottom 10% of samples are discarded, and the angle is $\max\theta - \min\theta$, halved if the two extrema lie within one eighth of the trace of each other (a single inflection rather than a curve). The root-mean-square residual between the raw and fitted midline, scaled by $0.045\,w$, gives a trace confidence in $[0, 1]$. When trained weights exist, the trace is used only to flag disagreements larger than $12°$ at confidence ≥ 0.5; when they do not, it is the only available estimate. Section VII-E quantifies how far it can be trusted.

E. Explanation and reporting

Grad-CAM [12] is computed from the same forward pass that produces the logits, so an explanation costs one backward pass rather than a second evaluation. The heat map is blended over the enhanced radiograph only where it exceeds 0.25 of its peak, and connected regions above 0.55 are reported as boxes.

TABLE I
DATASETS USED. ALL ARE PUBLIC; COUNTS ARE IMAGES.

ID	Source	n	Labels / split
K1	Chen *et al.* [8], [13], OAI	8,260	KL 0–4; 5,778 / 826 / 1,656
S1	Fraiwan *et al.* [19], [34]	338	normal 71, scoliosis 188, spondylolisthesis 79; 5-fold CV
S2	Spinal-AI2024 [17], synthetic	20,000	3 Cobb angles; 11,999 / 4,000 / 4,000

Every response carries the grade, its class probabilities, a normalised entropy and top-2 margin as uncertainty indicators, simple image-quality metrics (Laplacian variance, contrast, exposure clipping), per-stage timings, and a fixed disclaimer that the output is screening support and not a diagnosis.

IV. Training Recipe

All branches are trained with the same two-phase schedule. In phase 1 the trunk is frozen and only the head is trained for two to three epochs at learning rate 10^{-3}. In phase 2 the trunk is unfrozen at 3×10^{-4} (head unchanged) and both decay with a cosine schedule [31] to 10^{-6}. The optimiser is AdamW [32] with weight decay 10^{-4}; gradients are clipped at norm 5. Class imbalance—pronounced for KL 4 (3% of knees) and for the normal and severe spine bands—is addressed twice: the cross-entropy loss is weighted by inverse class frequency, and a weighted random sampler draws balanced mini-batches. Label smoothing of 0.05 [33] regularises the head. Training stops when validation balanced accuracy (mean per-class recall) has not improved for five to eight epochs, and the best-validation checkpoint is evaluated once on the held-out test split. For the spine branch the loss adds $\lambda\,\text{SmoothL1}(\hat{a}, a)$ on images with an angle label ($\lambda = 0.05$).

Augmentation is deliberately anatomy-aware: horizontal flips are applied to knees (left and right knees are both valid) but never to spines (a flip would invert the curve direction); rotations of $\pm 8°$, scaling of 0.94–1.06 and shifts of $\pm 4\%$ model positioning; contrast (0.85–1.18) and brightness (± 18 grey levels) jitter model exposure; Gaussian noise (σ 2–7, probability 0.25) and random erasing (probability 0.2) model grain, markers and implants. No vertical flips are used.

V. Datasets

K1 (knee). The KL-graded dataset of Chen *et al.* [13] consists of 8,260 posterior–anterior knee-joint crops of 224×224 pixels extracted from OAI radiographs, with the authors' own train/validation/test split of 5,778/826/1,656 images. Grades are strongly imbalanced (KL 0: 3,253; KL 1: 1,495; KL 2: 2,175; KL 3: 1,086; KL 4: 251). We use the split unchanged so that our numbers are directly comparable with [8], [10].

S1 (spine, real). The Jordan University of Science and Technology collection [34] holds 338 spinal radiographs from King Abdullah University Hospital labelled normal, scoliosis or spondylolisthesis. It has no angle annotations, so we use it for *detection*. Because the numeric prefix of its file names is

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

not a unique patient key, folds are stratified by class at image level; we use five folds with one held out for testing and one for early stopping in each round, and pool the five test folds for the reported confusion matrix.

S2 (spine, synthetic). Spinal-AI2024 [17] contains 20,000 synthetic AP spine radiographs of 512×512 pixels generated by the CurvNet data engine, each with proximal-thoracic, main-thoracic and thoraco-lumbar Cobb angles. We take the largest of the three as the major curve and derive its severity band. Following the authors, subset 5 is the test set; we use subsets 1–3 for training (11,999 images; one lacks a label) and subset 4 for validation. Bands are skewed towards *mild* (65%), with normal 7%, moderate 17% and severe 11%. We stress that S2 is synthetic: it provides exact angle ground truth at scale, but angle results on it are not comparable with results on real AASCE radiographs, whose public mirror was unavailable at the time of this study.

Router set. All 338 S1 images and 1,000 S2 images form the spine class and an equal number of K1 training crops the knee class, split 70/15/15.

VI. Experimental Setup

Experiments run on a 12-core CPU workstation without a GPU (PyTorch 2.5, CPU build). Knee models use $S = 224$ (the native crop size) and batch size 32; spine models use $S = 320$ and batch size 16. Backbones compared on K1 are DenseNet-121, ResNet-50 and EfficientNet-B0, all ImageNet-pretrained. The reference knee model is repeated with three seeds. Ablations on DenseNet-121 remove, one at a time, the class-weighted loss, the balanced sampler and label smoothing, and all three together. The Spinal-AI2024 branch is trained on a stratified 4,000-image subset of the training split (CPU budget) with ResNet-50 and the angle head. We report accuracy, balanced accuracy, macro-F1, quadratic-weighted Cohen's κ (appropriate for ordinal grades), the fraction of predictions within one grade, expected calibration error (ECE, 15 bins) [35] and per-grade precision/recall; for angles we report MAE, RMSE, the fraction within 5° and 10°, bias and Bland–Altman limits. Robustness is measured by degrading test images *before* preprocessing with JPEG re-compression (quality 90–30), additive Gaussian noise (σ 5–20), Gaussian blur (kernels 3–9), $\pm 25\%$ brightness and $0.7\times$ contrast. Explanations are checked with a pointing-game protocol [28]: for knees the anatomical prior region is the central horizontal band (30–70% of height) containing the tibiofemoral joint; for spines it is a band of $\pm 12\%$ of image width around the traced midline. Since neither dataset has lesion masks, this is a plausibility check, not a localisation benchmark.

VII. Results

All numbers are on held-out test splits that were never used for model selection. Training ran on a single NVIDIA T4 (Google Colab); one knee run took 15–28 min, one spine fold 3–5 min and the Spinal-AI2024 branch 73 min.

TABLE II
Knee KL grading on the K1 test split ($n = 1{,}656$). Mean ± SD over three seeds for DenseNet-121; single seed otherwise. ±1: fraction of predictions within one grade.

Backbone	Acc.	Bal. acc.	Macro-F1	κ_w	±1	ECE
DenseNet-121	.623±.006	.670±.009	.664±.008	.794±.008	.939	.077
ResNet-50	.618	.688	.669	.809	.955	.072
EfficientNet-B0	.569	.668	.637	.779	.941	.095

Published on the same split: Chen *et al.* [8] 69.7% acc.; Pi *et al.* [10] 76.9% acc., F1 0.767 (ensemble).

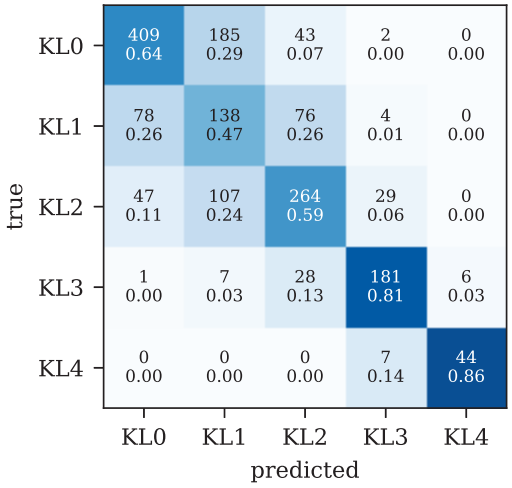


Fig. 2. Knee KL confusion matrix (DenseNet-121, seed 0, test split). Cell text: count and row-normalised rate.

A. Knee osteoarthritis grading (K1)

Table II compares the three backbones on the 1,656-image test split. DenseNet-121 and ResNet-50 are close (62% accuracy, 0.79–0.81 quadratic-weighted κ); EfficientNet-B0 is clearly weaker in accuracy at similar balanced accuracy. Over three seeds DenseNet-121 varies by less than one percentage point in every metric, so the differences between backbones are of the same order as seed noise and we do not claim a winner. For the reference model (seed 0) the bootstrap 95% confidence intervals (2,000 resamples of the test set) are 0.603–0.649 for accuracy, 0.647–0.702 for balanced accuracy and 0.766–0.811 for κ_w. Fig. 2 shows the DenseNet-121 confusion matrix: 94% of all predictions fall within one grade of the truth, and the errors concentrate where readers also disagree—KL 1 ("doubtful") is recalled only 41% of the time, mostly confused with KL 0 and KL 2—whereas the clinically consequential grades are recalled well (KL 3: 0.82, KL 4: 0.88, means over seeds).

In plain accuracy our single models sit below the two published results on this split (Table II). The gap is expected and, we argue, partly a design choice: our recipe optimises balanced accuracy, which deliberately gives up majority-class (KL 0) accuracy for minority-grade recall—KL 4 recall of 0.88 against a prior of 3%—and a single network is compared with an ensemble [10] and an ordinal loss [8]. The ablation below makes the trade-off explicit, and Section VII-B shows how

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

TABLE III
PRE-DECLARED SOFT-VOTE ENSEMBLES OF THE TRAINED KNEE NETWORKS ON THE K1 TEST SPLIT ($n = 1{,}656$). CI: BOOTSTRAP 95% INTERVAL; p: EXACT MCNEMAR TEST AGAINST THE SINGLE REFERENCE MODEL.

Model	Acc. [95% CI]	Bal. acc.	κ_w	±1	ECE	p
DenseNet-121, seed 0 (reference)	.626 [.603, .649]	.674	.789	.937	.106	–
ENS-A: 3 DenseNet-121 seeds	.646 [.622, .668]	.689	.818	.952	**.028**	.020
ENS-B: 5 backbone runs	.649 [.626, .673]	.703	.825	**.955**	.065	.013
ENS-C: all 9 knee networks	**.676** [.655, .699]	**.710**	**.836**	.954	.064	$< 10^{-4}$

much of the gap an ensemble of our own networks recovers.

B. Ensembling and ordinal decoding

Because every knee run saved its test-set probabilities, we can evaluate ensembles without training anything new. To avoid selecting on the test set, the member sets were fixed in advance: ENS-A, the three DenseNet-121 seeds; ENS-B, all five backbone runs of Table II; and ENS-C, all nine trained knee networks (ENS-B plus the four ablation variants). Probabilities are averaged and the argmax taken. Table III shows that every ensemble improves on the single reference model and that the gain grows with diversity: ENS-C reaches 67.6% accuracy (95% CI 0.655–0.699), 71.0% balanced accuracy and $\kappa_w = 0.836$, with 95.4% of predictions within one grade and KL-4 recall 0.90. The improvement over the single model is significant by an exact McNemar test on the paired test images ($p = 0.020$, 0.013 and $< 10^{-4}$ for ENS-A, -B and -C). ENS-C's accuracy is within two points of the ordinal-loss result of Chen *et al.* [8] while keeping the minority-grade recall their accuracy figure does not report; the 76.9% of Pi *et al.* [10] remains out of reach for networks trained with our balanced objective. The seed ensemble ENS-A is also the best-calibrated model we trained (ECE 0.028; Fig. 3).

We also tested *ordinal decoding*—predicting $\mathrm{round}(\sum_k k\, p_k)$ instead of the argmax, a common trick for ordinal scales—on every model and ensemble. It is harmful here: for the class-balanced models accuracy collapses to 0.44–0.49 (e.g. 0.626→0.471 for the reference model, McNemar $p < 10^{-4}$) because re-weighted probability mass on distant grades drags the expectation towards the middle of the scale; it helps only the two models trained without class weights (plain CE: κ_w 0.798→0.821). Expected-value decoding therefore presupposes probabilities that are calibrated to the class frequencies, which balanced training deliberately destroys.

C. Ablation of the imbalance recipe

Table IV removes components of the training recipe one at a time (DenseNet-121, seed 0). The plain cross-entropy model has the highest accuracy (65.1%) but the lowest balanced accuracy (63.9%) and KL 4 recall (0.82), because it favours KL 0. Removing the loss re-weighting while keeping the balanced sampler gives the best run overall (65.0% accuracy, 68.8% balanced accuracy, κ_w 0.817, KL 4 recall 0.96, ECE 0.027): weighting the loss *and* re-sampling the batches both correct

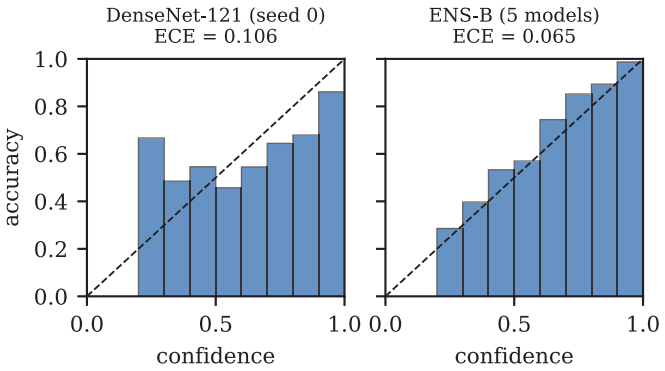


Fig. 3. Reliability diagrams on the K1 test split (10 confidence bins). Left: single DenseNet-121; right: ENS-B. The dashed line is perfect calibration.

TABLE IV
ABLATION ON K1 (DENSENET-121, SEED 0, TEST SPLIT). RECALL IS PER GRADE.

Configuration	Acc.	Bal. acc.	κ_w	R_{KL0}	R_{KL4}	ECE
Full recipe	.626	.674	.789	.640	.863	.106
– class weights	.650	**.688**	**.817**	.776	**.961**	**.027**
– balanced sampler	.640	.663	.810	.709	.804	.041
– label smoothing	.630	.667	**.817**	.673	.824	.126
Plain CE (none of the three)	**.651**	.639	.798	**.887**	.824	.058

for the same imbalance, and together they over-correct. Removing the sampler instead costs balanced accuracy (0.663). Removing label smoothing leaves discrimination unchanged but triples the calibration error (ECE 0.126 vs 0.041–0.058), so its role in this recipe is calibration rather than accuracy. Paired exact McNemar tests against the full recipe on the same 1,656 test images give $p = 0.061$ (plain), 0.053 (no class weights), 0.23 (no sampler) and 0.72 (no label smoothing): the two largest effects are borderline and the others are not significant at a single seed. The differences should therefore be read as indicative; the direction of each effect is consistent with the mechanism described, and a multi-seed repetition is the obvious next step.

D. Scoliosis detection on real radiographs (S1)

Pooling the five test folds (Table V), DenseNet-121 classifies the 338 real spine radiographs with 95.9% accuracy (bootstrap 95% CI 0.938–0.979), 95.3% balanced accuracy (0.926–0.976) and a macro one-vs-rest AUC of 0.986; the per-fold accuracy is 0.959 ± 0.011. Sensitivity for scoliosis—the screening quantity—is 0.968 (182 of 188; CI 0.940–0.990), with three scoliosis cases called normal and three called spondylolisthesis. ResNet-50 is slightly lower (94.1% accuracy, scoliosis sensitivity 0.936). Fig. 4 shows the pooled DenseNet-121 confusion matrix. With 338 images from one hospital these numbers describe in-distribution performance only; we return to this in Section VIII.

E. Cobb severity and angle on Spinal-AI2024 (S2)

Geometric estimator. On the 4,000 test images the weights-free trace of Section III-D achieves a mean absolute error of

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

TABLE V
SPINE DETECTION ON S1, FIVE TEST FOLDS POOLED ($n = 338$).

Backbone	Acc.	Bal. acc.	Macro-F1	AUC	Sens. scol.	Sens. spondy.
DenseNet-121	.959	.953	.951	.986	.968	.975
ResNet-50	.941	.943	.934	.982	.936	.962

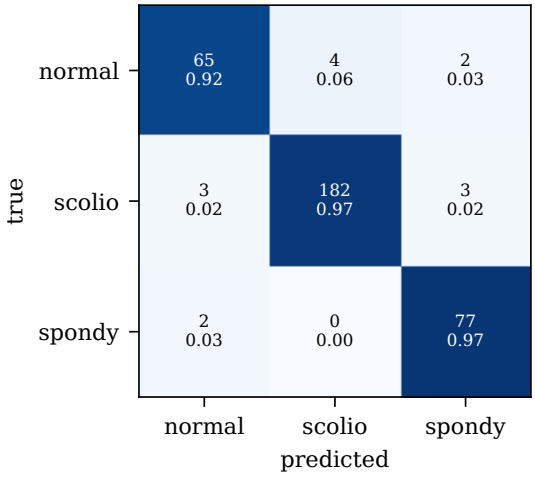


Fig. 4. Pooled five-fold confusion matrix on the real spine set (DenseNet-121).

11.6° (RMSE 15.3°, median 9.1°), with 29.5% of estimates within 5° and 53.7% within 10°, Pearson $r = 0.44$ and a bias of +4.1° (Fig. 5). It over-estimates small curves (bias +16° on normal spines) and under-estimates large ones (−12° on severe), so its four-band accuracy is only 45% ($\kappa_w = 0.32$). Its self-reported trace confidence is nevertheless informative: MAE falls from 14.0° at confidence below 0.4 ($n = 1{,}618$) to 10.8° at 0.4–0.6 and 8.4° at 0.6–0.8 ($n = 741$). We therefore treat it as a coarse, confidence-gated screening signal and never as a measurement. It costs 31 ms per image on CPU.

Learned branch. The ResNet-50 angle head reduces the error by a factor of four: MAE 2.62°, RMSE 4.21°, median 1.81°, 89.4% within 5° and 97.1% within 10°, $r = 0.955$, bias −0.46°. Error grows with curve magnitude (MAE 1.9° for mild, 2.4° moderate, 7.1° severe) and small curves are slightly over-estimated (+2.6° on normal spines), the same pattern reported for real radiographs in the literature [11], [18]. For the four severity bands the two outputs of the same network disagree in an instructive way (Table VI). The class-weighted classification head is tuned for balanced recall and over-predicts *normal* (508 of 2,616 mild spines are called normal), giving 76.7% accuracy but 86.3% balanced accuracy; deriving the band from the regressed angle instead gives 86.6% accuracy, $\kappa_w = 0.877$ and a mild-band recall of 0.914, at the price of lower normal-band recall (0.576), because curves near the 10° threshold are genuinely ambiguous. The two heads agree on 78.8% of images. The served system therefore reports the angle-derived band as the clinically anchored output and flags disagreement with the classification head, exactly as it flags disagreement with the geometric trace. We stress that S2

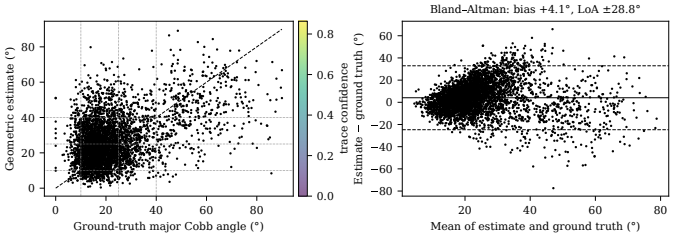


Fig. 5. Geometric Cobb estimator on the Spinal-AI2024 test set ($n = 4{,}000$). Left: estimate versus ground truth, coloured by trace confidence; dotted lines mark the severity-band thresholds. Right: Bland–Altman plot.

TABLE VI
SEVERITY BANDS ON SPINAL-AI2024 TEST ($n = 4{,}000$) FROM THE TWO OUTPUTS OF THE SAME RESNET-50 AND FROM THE GEOMETRIC TRACE.

Band source	Acc.	Bal. acc.	κ_w	Angle MAE
Classification head	.767	**.863**	.825	–
Regressed angle	**.866**	.789	**.877**	**2.6°**
Geometric trace (no weights)	.452	.440	.317	11.6°

is synthetic and these angle errors are not comparable with results on real AASCE radiographs.

F. Region router

The MobileNetV3-small router separates spine from knee radiographs with 99.5% accuracy on its held-out split (2 errors in 398, both spine images called knee); it converged in its first epoch and costs 38 ms on CPU. Misrouted images are caught downstream by the branch's low-confidence flag, and the clinician can force the region.

G. Robustness to image quality

Fig. 6 summarises balanced accuracy under degradation. The knee model (651 test images) is insensitive to JPEG re-compression down to quality 50 (−0.012) and to mild blur or contrast loss, loses 0.02–0.05 at JPEG quality 30, 9×9 blur or +25% brightness, and fails only under heavy additive noise ($\sigma = 20$ grey levels: 0.665 → 0.397). The spine detector (fold 0, 68 images) is more tolerant: it keeps ≥ 0.90 balanced accuracy under every degradation except $\sigma = 20$ noise (0.819). Gaussian noise is thus the only degradation a deployment must guard against, which the quality report's sharpness and contrast flags already surface.

H. Explanation plausibility

With the anatomical-prior regions of Section VI, the Grad-CAM maximum falls inside the knee joint band for 93.8% of 291 test crops (chance, by area, 39.9%), and 60.4% of the map's energy lies in that band; the hit rate rises with grade (0.90 for KL 0 to 1.00 for KL 4). For the spine detector the maximum lies within the traced spinal column for 80.9% of 68 images (chance 65.4%) with 73.2% of the energy inside; normal spines are attended almost entirely along the column (hit rate 1.00) whereas scoliosis and spondylolisthesis cases draw attention partly outside it (0.78 and 0.69), plausibly to the pelvis and rib cage, which carry information about curvature

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

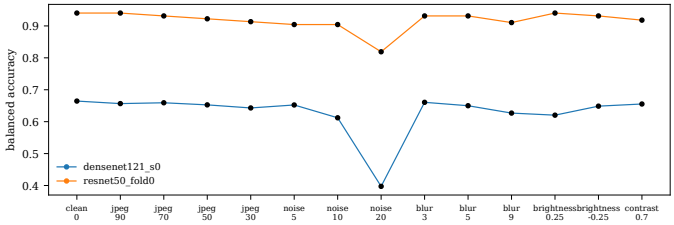


Fig. 6. Balanced accuracy under image degradation applied before preprocessing (knee: DenseNet-121, $n = 651$; spine: ResNet-50 fold 0, $n = 68$).

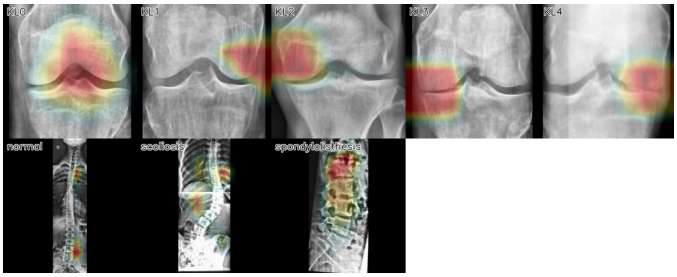


Fig. 7. Grad-CAM overlays for correctly graded test images: knee KL 0–4 (top) and spine normal / scoliosis / spondylolisthesis (bottom). Only regions above 25% of the map's peak are tinted.

and listhesis. Fig. 7 shows one correctly classified example per class. This is evidence of anatomical plausibility, not of lesion-level localisation, for which no annotations exist in these datasets.

I. Serving latency

Table VII reports per-stage medians over 100 images on a 12-core laptop CPU (12 threads, PyTorch 2.5 CPU build, no GPU), measured by calling the serving modules directly. A knee radiograph is graded in 0.27 s and a spine radiograph in 0.43 s; the Grad-CAM backward pass is the single most expensive stage and roughly doubles the end-to-end time to 0.67 s and 1.20 s respectively. The geometric Cobb trace adds 11 ms. The 95th percentiles are 0.85 s (knee) and 2.1 s (spine). These figures are compatible with interactive use on a clinic workstation; disabling the explanation when it is not needed halves the response time.

VIII. Discussion

Where the knee errors are. Nearly all residual knee error is between adjacent grades, and more than half of it involves KL 1. This mirrors the human literature: KL 1 is defined by *doubtful* narrowing and *possible* osteophytes, and inter-reader κ on the full scale is only moderate [6]. A screening tool should therefore report the probability over grades and the top-two margin, as ours does, rather than a single hard label; a KL 0/1 split with a thin margin is a legitimate "read again" signal, while a confident KL 3/4 is actionable. Our accuracy gap to the best published single-split results [8], [10] is partly the balanced-accuracy objective (Table IV) and partly architectural: a pre-declared ensemble of our own networks recovers five accuracy points and 0.05 of κ_w at

TABLE VII
Median per-stage latency (ms) on a 12-core laptop CPU, 100 images each.

Branch	Prep.	Route	Class.	Grad-CAM	Cobb	Render	Total
Knee (DenseNet-121, 224 px)	8	38	228	387	–	5	671
Spine (ResNet-50, 320 px)	30	66	330	702	11	16	1,196

no training cost (Table III), closing most of the distance to the ordinal-loss result of [8], and a distance-weighted ordinal penalty in the loss is the natural next step—whereas ordinal *decoding* of balanced-trained probabilities should be avoided, as Section VII-B shows. The ablation also carries a practical lesson: of the three imbalance measures that are routinely stacked, the balanced sampler does the work, loss re-weighting on top of it hurts, and label smoothing earns its place through calibration—with the caveat that, at one seed, only the first two effects approach significance.

Spine: real detection, synthetic measurement. The real-image result is strong but narrow—one hospital, 338 images, detection labels only—so it shows that the preprocessing and recipe transfer to clinical spine films, not that the model generalises across sites. The angle result is the mirror image: large and precisely labelled, but synthetic. Two observations nevertheless carry over. First, the regressed angle is a better basis for the severity band than a separately trained class head, and the two can be cross-checked for free. Second, the weights-free geometric trace, although useful as a sanity check and as the only estimate before training, is far too imprecise (11.6°) to grade severity; its confidence score should gate whether it is shown at all. A paper reporting a geometric Cobb method without such a comparison would overstate it.

What a clinician sees. Every output carries the grade, the class probabilities, the uncertainty indicators, the image-quality flags and, for spines, the agreement between three independent estimates (class head, angle head, geometric trace). The robustness study tells us which flag matters most—noise—and the Grad-CAM check shows that the attended regions are anatomically sensible in more than nine of ten knee cases. None of this replaces a radiologist; it is designed to decide which films a radiologist should look at first.

A. Limitations

The spine angle results are on synthetic radiographs, and the real spine set is small and has detection labels only; the synthetic set carries no explicit licence and is cited as its authors request. Folds on S1 are stratified at image rather than patient level. The knee crops come from a single cohort (OAI) and a single pre-processing pipeline, so cross-institutional generalisation [18] is untested, and even our best ensemble trails the best published single-split result [10] by nine accuracy points. Ablation differences are single-seed and only borderline significant; the ensembles, although pre-declared, were evaluated once on the same test split as their members. Serving the nine-network ensemble would multiply the classification latency of Table VII by about nine. Grad-CAM plausibility is checked against anatomical priors, not lesion annotations.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Latency was measured on one laptop CPU. Finally, no reader study was performed; agreement with clinicians on new cases remains future work.

IX. Conclusion

We described and evaluated OrthoScan AI, a single CPU-deployable pipeline that routes a plain radiograph to a spine or knee branch and returns a severity output on the clinical scale together with an explanation and explicit uncertainty. On public data a single knee network grades KL severity with 0.79 quadratic-weighted κ and 94% adjacent-grade agreement, favouring minority-grade recall by design, and a pre-declared ensemble of the nine trained networks reaches 67.6% accuracy and $\kappa_w = 0.84$; the spine branch detects scoliosis on real films with 96.8% sensitivity and, on synthetic films with exact labels, estimates the major Cobb angle to within 2.6° on average. Four findings are of general use beyond this system: stacking loss re-weighting on a balanced sampler over-corrects class imbalance; expected-value (ordinal) decoding is harmful for probabilities produced by balanced training; a regressed angle is a better basis for the severity band than a parallel classification head; and the weights-free geometric Cobb trace, a tempting shortcut for curvature screening, is an order of magnitude less precise than a learned estimator and should be used only as a confidence-gated fallback. The next steps are external validation on multi-site spine and knee radiographs, an ordinal penalty in the knee loss, multi-seed ablations, and a reader study measuring whether the uncertainty and explanation outputs change how clinicians triage.

Code and Data Availability

All three datasets are public (Table I). The training, evaluation and figure scripts, the fixed split definitions, every trained checkpoint, the per-image test predictions used for the ensembles and statistics, and the Colab notebook that reproduces the GPU runs are available from the corresponding author on request. All numbers in this paper are generated by those scripts from the saved outputs.

Acknowledgment

The authors used Claude (Anthropic) to assist with drafting and editing the text of this manuscript and with writing the evaluation scripts; all experiments were run, verified and interpreted by the authors, who take full responsibility for the content.

References

[1] M. R. Konieczny, H. Senyurt, and R. Krauspe, "Epidemiology of adolescent idiopathic scoliosis," *Journal of Children's Orthopaedics*, vol. 7, no. 1, p. 3–9, Feb. 2013. [Online]. Available: http://dx.doi.org/10.1007/s11832-012-0457-4

[2] J. D. Steinmetz, G. T. Culbreth, L. M. Haile, Q. Rafferty, J. Lo, K. G. Fukutaki, J. A. Cruz, A. E. Smith, S. E. Vollset, P. M. Brooks, M. Cross, A. D. Woolf, H. Hagins, M. Abbasi-Kangevari, A. Abedi, I. N. Ackerman, H. Amu, B. Antony, J. Arabloo, A. Y. Aravkin, A. M. Argaw, A. A. Artamonov, T. Ashraf, A. Barrow, L. M. Bearne, I. M. Bensenor, A. Y. Berhie, N. Bhardwaj, P. Bhardwaj, V. S. Bhojaraja, A. Bijani, P. S. Briant, A. M. Briggs, N. S. Butt, J. Charan, V. K. Chattu, F. M. Cicuttini, K. Coberly, O. Dadras, X. Dai, L. Dandona, R. Dandona, K. de Luca, E. Denova-Gutiérrez, S. D. Dharmaratne, M. Dhimal, M. Dianatinasab, K. E. Dreinhoefer, M. Elhadi, U. Farooque, H. R. Farpour, I. Filip, F. Fischer, M. Freitas, B. Ganesan, B. N. B. Gemeda, T. Getachew, S.-H. Ghamari, A. Ghashghaee, T. K. Gill, M. Golechha, D. Golinelli, B. Gupta, V. B. Gupta, V. K. Gupta, R. Haddadi, N. Hafezi-Nejad, R. Halwani, S. Hamidi, A. Hanif, N. I. Harlianto, J. M. Haro, J. Hartvigsen, S. I. Hay, J. J. Hebert, G. Heidari, M.-S. Hosseini, M. Hosseinzadeh, A. K. Hsiao, I. M. Ilic, M. D. Ilic, L. Jacob, R. Jayawardena, R. P. Jha, J. B. Jonas, N. Joseph, H. Kandel, I. M. Karaye, M. J. Khan, Y. J. Kim, A.-A. Kolahi, O. Korzh, R. Koteeswaran, V. Krishnamoorthy, G. A. Kumar, N. Kumar, S.-w. Lee, S. S. Lim, S. W. Lobo, G. Lucchetti, M.-R. Malekpour, A. A. Malik, L. G. G. Mandarano-Filho, S. Martini, A.-F. A. Mentis, M. K. Mesregah, T. Mestrovic, E. M. Mirrakhimov, A. Misganaw, R. Mohammadpourhodki, A. H. Mokdad, S. Momtazmanesh, S. D. Morrison, C. J. L. Murray, H. Nassereldine, H. B. Netsere, S. Neupane Kandel, M. O. Owolabi, S. Panda-Jonas, A. Pandey, S. Pawar, P. Pedersini, J. Pereira, A. Radfar, M.-M. Rashidi, D. L. Rawaf, S. Rawaf, R. Rawassizadeh, S.-M. Rayegani, D. Ribeiro, L. Roever, B. Saddik, A. Sahebkar, S. Salehi, L. Sanchez Riera, F. Sanmarchi, M. M. Santric-Milicevic, S. Shahabi, M. A. Shaikh, E. Shaker, M. Shannawaz, R. Sharma, S. Sharma, J. K. Shetty, R. Shiri, P. Shobeiri, D. A. S. Silva, A. Singh, J. A. Singh, S. Singh, S. T. Skou, H. Slater, M. S. Soltani-Zangbar, A. V. Starodubova, A. Tehrani-Banihashemi, S. Valadan Tahbaz, P. R. Valdez, B. Vo, L. G. Vu, Y.-P. Wang, S. H. Yahyazadeh Jabbari, N. Yonemoto, I. Yunusa, L. M. March, K. L. Ong, T. Vos, and J. A. Kopec, "Global, regional, and national burden of osteoarthritis, 1990–2020 and projections to 2050: a systematic analysis for the global burden of disease study 2021," *The Lancet Rheumatology*, vol. 5, no. 9, p. e508–e522, Sep. 2023. [Online]. Available: http://dx.doi.org/10.1016/S2665-9913(23)00163-7

[3] J. Kellgren and J. Lawrence, "Radiological assessment of osteo-arthrosis," *Annals of the Rheumatic Diseases*, vol. 16, no. 4, p. 494–502, Dec. 1957. [Online]. Available: http://dx.doi.org/10.1136/ard.16.4.494

[4] R. T. Morrissy, G. S. Goldsmith, E. C. Hall, D. Kehl, and G. H. Cowie, "Measurement of the cobb angle on radiographs of patients who have scoliosis. evaluation of intrinsic error." *The Journal of Bone & Joint Surgery*, vol. 72, no. 3, p. 320–327, Mar. 1990. [Online]. Available: http://dx.doi.org/10.2106/00004623-199072030-00002

[5] D. L. Carman, R. H. Browne, and J. G. Birch, "Measurement of scoliosis and kyphosis radiographs. intraobserver and interobserver variation." *The Journal of Bone & Joint Surgery*, vol. 72, no. 3, p. 328–333, Mar. 1990. [Online]. Available: http://dx.doi.org/10.2106/00004623-199072030-00003

[6] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach," *Scientific Reports*, vol. 8, no. 1, Jan. 2018. [Online]. Available: http://dx.doi.org/10.1038/s41598-018-20132-7

[7] J. Antony, K. McGuinness, N. E. O'Connor, and K. Moran, "Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, Dec. 2016, p. 1195–1200. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2016.7899799

[8] P. Chen, L. Gao, X. Shi, K. Allen, and L. Yang, "Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss," *Computerized Medical Imaging and Graphics*, vol. 75, p. 84–92, Jul. 2019. [Online]. Available: http://dx.doi.org/10.1016/j.compmedimag.2019.06.002

[9] B. Norman, V. Pedoia, A. Noworolski, T. M. Link, and S. Majumdar, "Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs," *Journal of Digital Imaging*, vol. 32, no. 3, p. 471–477, Oct. 2018. [Online]. Available: http://dx.doi.org/10.1007/s10278-018-0098-3

[10] S.-W. Pi, B.-D. Lee, M. S. Lee, and H. J. Lee, "Ensemble deep-learning networks for automated osteoarthritis grading in knee x-ray images,"

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Scientific Reports, vol. 13, no. 1, Dec. 2023. [Online]. Available: http://dx.doi.org/10.1038/s41598-023-50210-4

[11] Y. Zhu, X. Yin, Z. Chen, H. Zhang, K. Xu, J. Zhang, and N. Wu, "Deep learning in cobb angle automated measurement on x-rays: a systematic review and meta-analysis," *Spine Deformity*, vol. 13, no. 1, p. 19–27, Sep. 2024. [Online]. Available: http://dx.doi.org/10.1007/s43390-024-00954-4

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017, p. 618–626. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2017.74

[13] P. Chen, "Knee osteoarthritis severity grading dataset," 2018. [Online]. Available: https://data.mendeley.com/datasets/56rmx5bjcr/1

[14] A. Tiulpin and S. Saarakkala, "Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks," *Diagnostics*, vol. 10, no. 11, p. 932, Nov. 2020. [Online]. Available: http://dx.doi.org/10.3390/diagnostics10110932

[15] H. Wu, C. Bailey, P. Rasoulinejad, and S. Li, *Automatic Landmark Estimation for Adolescent Idiopathic Scoliosis Assessment Using BoostNet*. Springer International Publishing, 2017, p. 127–135. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-66182-7_15

[16] J. Yi, P. Wu, Q. Huang, H. Qu, and D. N. Metaxas, "Vertebra-focused landmark detection for scoliosis assessment," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, Apr. 2020, p. 736–740. [Online]. Available: http://dx.doi.org/10.1109/ISBI45749.2020.9098675

[17] Z. Shao, Y. Yuan, L. Ma, and X. Zhu, "Curvnet: Latent contour representation and iterative data engine for curvature angle estimation," arXiv:2411.12604, 2024. [Online]. Available: https://arxiv.org/abs/2411.12604

[18] Šimon Kubov, S. Klíčník, J. Dandár, Z. Straka, K. Kvaková, and D. Kvak, "Multi-centre validation of a deep learning model for scoliosis assessment," arXiv:2507.14093, 2025. [Online]. Available: https://arxiv.org/abs/2507.14093

[19] M. Fraiwan, Z. Audat, L. Fraiwan, and T. Manasreh, "Using deep transfer learning to detect scoliosis and spondylolisthesis from x-ray images," *PLOS ONE*, vol. 17, no. 5, p. e0267851, May 2022. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0267851

[20] Y. H. Nguyen, "Scoliosis detection using deep neural network," arXiv:2210.17269, 2022. [Online]. Available: https://arxiv.org/abs/2210.17269

[21] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, p. 1299–1312, May 2016. [Online]. Available: http://dx.doi.org/10.1109/TMI.2016.2535302

[22] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, p. 60–88, Dec. 2017. [Online]. Available: http://dx.doi.org/10.1016/j.media.2017.07.005

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2016, p. 770–778. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.90

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jul. 2017, p. 2261–2269. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2017.243

[25] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," arXiv:1905.11946, 2019. [Online]. Available: https://arxiv.org/abs/1905.11946

[26] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for mobilenetv3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019, p. 1314–1324. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2019.00140

[27] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102470, Jul. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.media.2022.102470

[28] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, p. 1084–1102, Dec. 2017. [Online]. Available: http://dx.doi.org/10.1007/s11263-017-1059-x

[29] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, p. 355–368, Sep. 1987. [Online]. Available: http://dx.doi.org/10.1016/S0734-189X(87)80186-X

[30] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, ser. ICCV-98. Narosa Publishing House, p. 839–846. [Online]. Available: http://dx.doi.org/10.1109/ICCV.1998.710815

[31] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," arXiv:1608.03983, 2016. [Online]. Available: https://arxiv.org/abs/1608.03983

[32] ——, "Decoupled weight decay regularization," arXiv:1711.05101, 2017. [Online]. Available: https://arxiv.org/abs/1711.05101

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2016, p. 2818–2826. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.308

[34] Mohammad Fraiwan, "A dataset of scoliosis, spondylolisthesis, and normal vertebrae x-ray images," 2022. [Online]. Available: https://data.mendeley.com/datasets/xkt857dsxk/1

[35] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," arXiv:1706.04599, 2017. [Online]. Available: https://arxiv.org/abs/1706.04599

(This work is licensed under a Creative Commons Attribution 4.0 International License.)