A radiation-free screening system for adolescent idiopathic scoliosis using deep learning on 3D back surface point clouds

Jingfan Yang[1,†], Haonan Shi[2,†], Zifang Huang[3,†], Xiaodong Wang[2,†], Wenhui Wang[4], Tianyuan Zhang[1], Jingwen Wang[2], Yifan Zhan[5], Hui Liu[5], Zifang Zhang[1], Jiannan Zhang[5], Zhijun Fei[6], Xiaoling Xuan[6], Yihao Gao[6], Yaolong Deng[1], Long Tian[2], Liming Wang[2], Xiyang Liu[2], Yaqing Zhang[1,*], Lirong Ai[5,*], Junlin Yang[1,*]

[1]Pediatric Orthopedics and Spine Center, Xin Hua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China.
[2]School of Computer Science and Technology, Xidian University, Xi'an, China.
[3]The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China.
[4]Hangzhou Institute of Technology, Xidian University, HangZhou, China
[5]School of Computer Science, Northwestern Polytechnical University, Xi'an, China.
[6]Xinmiao Scoliosis Prevention Center, Guangzhou, China.

[*]Corresponding author(s). yjunlin@126.com (J.L.Y.); ailr@nwpu.edu.cn (L.A.); zhangyaqingmed@163.com (Y.Q.Z.)
[†]These authors contributed equally to this work.

Abstract

Widespread screening for Adolescent Idiopathic Scoliosis (AIS) is critical for timely intervention but is currently constrained by the radiation risks of X-rays and the subjectivity of physical examinations. Here, we present PointScol, a radiation-free triage system leveraging 3D back surface point clouds. To reconcile the conflicting clinical demands for "zero-miss" screening and "fine-grained" severity assessment, we developed a two-stage deep learning framework. First, an automated segmentation module extracts the dorsal region of interest (ROI) to standardize input geometry. Second, the system employs a dual-branch diagnostic strategy: a binary classification network designed for maximal sensitivity to rule out health, and a 5-class grading network designed to stratify severity (0–10°, 11–20°, 21–30°, 31–40°, >40°). Validation on a multi-center dataset (n=128) confirmed the distinct utility of this hierarchical approach. For the scoliosis screening task using a 10° Cobb angle threshold, the binary classification model achieved a sensitivity of 100.00% in the external cohort, ensuring that no cases requiring further clinical attention were missed. While the 5-class grading task inherently faces greater complexity, it successfully achieved an overall accuracy of 84.48% and, crucially, demonstrated a high specificity of 98.42% for severe surgical cases (>40°). This performance profile establishes PointScol as a safe clinical filter: the binary module reliably excludes healthy individuals, while the 5-class module flags high-risk patients for prioritized intervention, collectively offering a non-invasive, resource-efficient paradigm for scoliosis management.

Introduction

Adolescent Idiopathic Scoliosis (AIS) is a complex three-dimensional spinal deformity characterized by a lateral curvature with a Cobb angle greater than 10°. Affecting 3% to 5% of

adolescents, the condition can lead to severe thoracic asymmetry, cardiopulmonary dysfunction, and psychological distress if left untreated.[1,2] Consequently, widespread screening is critical to ensure timely intervention, preventing disease progression and improving long-term quality of life.[3] Optimal screening requires a delicate balance: high sensitivity is critical to identify all potential cases for referral, while high specificity is essential to minimize unnecessary radiation exposure and anxiety for healthy adolescents.[4]

While the radiographic Cobb angle remains the clinical "gold standard" for diagnosing scoliosis, its reliance on ionizing radiation poses significant cumulative health risks, particularly for adolescents requiring serial monitoring.[5] Although radiation-free alternatives such as freehand 3D ultrasound have emerged, their widespread deployment is often constrained by high operator dependency, the need for complex tracking equipment, and low clinical throughput.[6,7] To address these bottlenecks, recent computer vision approaches have leveraged deep learning to analyze 2D back surface photography or synthesize virtual X-rays via generative adversarial networks (GANs).[8,9] However, scoliosis is inherently a complex, three-dimensional volumetric deformity. Purely 2D-based models suffer from an unavoidable "dimensionality gap," failing to capture critical rotational depth information. Furthermore, generative approaches in the absence of geometric constraints are prone to "hallucinations," introducing non-existent anatomical features that compromise diagnostic reliability.[9]

To overcome these planar limitations, research paradigms have pivoted towards capturing volumetric data via markerless surface topography and RGB-D sensing,[10,11] alongside dynamic analysis of gait sequences.[12] The theoretical superiority of direct 3D learning is well-established; point cloud architectures are capable of extracting local geometric features of asymmetry that are unattainable by 2D methods.[13] Despite this technological maturity, a critical gap remains in clinical translation: existing 3D tools primarily focus on shape reconstruction or scalar regression of the Cobb angle.[14] There is a notable lack of dedicated frameworks designed specifically as high-sensitivity triage filters—tools capable of safely "ruling out" healthy individuals.[15] Moreover, current 3D deep learning models often operate as "black boxes," lacking the geometric interpretability necessary to foster clinical trust.[16]

Herein, we present PointScol, a novel radiation-free triage system that reconciles screening safety with diagnostic granularity, with its overall workflow and system architecture illustrated in Fig. 1. Our framework employs a streamlined two-stage, dual-task deep learning architecture. First, to ensure data fidelity, an automated segmentation module isolates the dorsal region of interest (ROI), eliminating environmental noise that plagues traditional non-contact screenings. Second, the system executes a hierarchical diagnostic protocol: (1) A Binary Screening Module functions as a high-sensitivity gatekeeper to identify all potential risk cases (>10°); (2) A Fine-grained Grading Module then stratifies these cases into five severity classes. This design allows the system to achieve "zero-miss" safety for exclusion while providing specific severity contexts (e.g., distinguishing observation candidates from surgical urgencies) to guide referral urgency. Developed and validated on a large-scale, multi-center dataset comprising paired 3D scans and X-ray ground truths, our system demonstrates distinct utility across these tasks. PointScol achieves a sensitivity of 100.00% for the primary screening threshold, effectively ensuring no false negatives. Simultaneously, its severity grading capability offers a reliable "second opinion" for resource allocation, bridging the gap between automated detection and personalized patient management.

The primary contributions of this work include:

1. **Establishment of a Multi-Center 3D Benchmark:** We constructed a large-scale dataset paired with X-ray ground truths, addressing the scarcity of high-quality annotated 3D data.
2. **A Hierarchical "Safety-First" Triage Paradigm:** We propose a novel workflow that combines **high-fidelity ROI extraction** with a **dual-objective diagnosis**. This structure guarantees near-zero false negatives (Sensitivity: 100%) for safe exclusion while offering **clinically actionable severity stratification** (5-class grading) to optimize referral efficiency.
3. **Interpretability and Clinical Trustworthiness:** By incorporating geometric feature visualization, our system bridges the gap between “black box” AI predictions and clinical decision-making.

Methods

Data Cohorts and Preprocessing

This study was approved by the Ethics Committee of Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine and conducted in accordance with the Declaration of Helsinki. Due to the retrospective nature of the development dataset, the requirement for informed consent was waived, whereas written informed consent was obtained from all participants in the external validation cohort. The dataset utilized in this study was collaboratively collected over a comprehensive three-year period, from October 2022 to September 2024, involving a primary development center and two independent validation centers. To simulate a resource-constrained screening environment, all 3D back surface topographical data were acquired using a portable, consumer-grade scanning solution: the 3D Systems iSense scanner (approximate market price ~$499 USD) integrated with an Apple iPad or iPhone interface. This specific hardware configuration was selected to demonstrate the feasibility of low-cost, radiation-free mass screening. To establish the radiological ground truth for spinal deformity, standing full-spine high-resolution X-ray imaging was acquired synchronously with the point cloud scans for all subjects. The Cobb angle, serving as the gold standard for diagnosing and grading scoliosis, was measured and verified by senior radiologists to ensuring the reliability of the labels used for model development.

The study design incorporated two geographically and institutionally distinct cohorts to rigorously evaluate model robustness against real-world variations. The Xinhua Development Cohort served as the foundational dataset for model training and internal parameter optimization. This cohort, affiliated with Shanghai Jiao Tong University School of Medicine, contributed a total of 648 samples collected under controlled clinical conditions. To ensure the model's generalizability and robustness, we established an independent Multi-Center Validation Cohort, comprising 128 samples collected from The Third Affiliated Hospital of Sun Yat-sen University and the Xin Miao Scoliosis Treatment Centre. This multi-center approach was explicitly designed to expose the model to heterogeneity in patient demographics, slight variations in imaging instrumentation, and procedural nuances inherent to different clinical settings. Strict inclusion and exclusion criteria were applied uniformly across all centers: participants were excluded if they had a history of spinal surgery, visible physical defects unrelated to scoliosis that might distort back topology, or incomplete imaging data. To prevent data leakage and ensure an unbiased evaluation, the dataset was partitioned strictly on a per-patient basis.

To facilitate high-throughput mass screening and eliminate inter-operator variability, we implemented a fully automated preprocessing pipeline to normalize raw scan data for the downstream classification network. First, raw 3D data acquired in stereolithography (STL) format were converted into point cloud representations via point sampling. To isolate the Region of Interest (ROI), we employed a deep learning segmentation network to automatically extract the posterior torso (lower neck, trunk, and lower back) while computationally excising the head and limbs. Ground truth for this network was established by clinical experts, who manually annotated the anatomical ROI using the open-source software CloudCompare.[17]

Following segmentation, the raw point clouds exhibited varying densities. To standardize input resolution, all samples were normalized to a fixed resolution of N = 10,000 points. High-density scans were downsampled using Farthest Point Sampling (FPS) to maximally preserve geometric topology, while sparse scans were upsampled via linear interpolation. This normalization ensures that feature density remains invariant across patients of different body sizes, directing the classifier to focus on topographical shape asymmetries indicative of scoliosis rather than artifacts arising from scanning variations.

ROI Extraction via 3D Point Cloud Segmentation

To precisely isolate the dorsal Region of Interest (ROI) from raw back surface scans—a critical step to eliminate environmental noise and clothing artifacts—we employed the Point Transformer network.[18] Unlike projection-based or voxel-based methods that rasterize data into regular grids, often causing a loss of fine-grained topological details, the Point Transformer operates directly on the unordered set of points in their native continuous metric space. This approach is particularly advantageous for preserving the intrinsic surface curvature and subtle asymmetry required for scoliosis assessment.

The network employs a U-Net-style encoder-decoder architecture adapted for dense semantic segmentation. The backbone is built upon Point Transformer Layers utilizing a vector self-attention mechanism. Specifically, for a local neighborhood of points, the layer applies a subtraction-based relation function combined with trainable position encodings. This mechanism allows the model to modulate individual feature channels effectively, capturing complex local geometric relationships such as the scapular prominence and the lateral waistline rather than relying solely on global shape statistics. The encoder progressively reduces point cardinality while increasing feature dimensionality through five stages, utilizing Farthest Point Sampling (FPS) and K-Nearest Neighbor (kNN) grouping to ensure robust feature pooling even across scans with varying point densities. The symmetric decoder then recovers the high-resolution point set via trilinear interpolation, fusing deep semantic features with shallow geometric details to accurately delineate anatomical boundaries along the neck and lower back.

The selection of the Point Transformer offers distinct advantages for processing biological surface topology in a clinical screening context. First, the self-attention operator is permutation-invariant, making it mathematically robust to the irregular and unstructured nature of handheld 3D scans. Second, by employing vector attention rather than scalar attention, the model achieves adaptive modulation of feature channels, which significantly enhances its ability to differentiate smooth skin surfaces from background noise compared to baseline MLPs. Third, the integration of explicit trainable position encoding enables the network to learn relative geometric contexts, which is

critical for recognizing the standardized dorsal orientation required for subsequent deformity analysis.

Model training and inference were conducted using the default configuration parameters described in the original Point Transformer literature to ensure reproducibility. The input to the network was a normalized point cloud tensor of shape (N, 3), representing the 3D coordinates. We utilized a neighborhood size of k = 16 for the k-Nearest Neighbor (k-NN) search in the local attention and downsampling modules, as ablation studies indicate this value provides the optimal balance between local context capture and noise reduction. The encoder comprises five stages with downsampling rates of [1, 4, 4, 4, 4], resulting in point set cardinalities of [N, N/4, N/16, N/64, N/256] respectively. The network was optimized using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 10^{-4}.

PointGST-based Classification Model

To efficiently adapt pre-trained point cloud models to the specific morphological nuances of AIS, we employed the PointGST (Point cloud Graph Spectral Tuning) framework.[19] Unlike standard fully fine-tuning (FFT) strategies that update all model parameters—often leading to overfitting on limited clinical datasets—this approach freezes the pre-trained Transformer backbone (e.g., Point-MAE) and integrates a lightweight, trainable Point Cloud Spectral Adapter (PCSA) into each Transformer layer.[20] We selected this spectral architecture specifically for its ability to decompose complex spinal deformities into distinct geometric frequencies: identifying global curvature trends while isolating subtle local asymmetries (e.g., rib humps) that are often obscured in the spatial domain.

The core mechanism we utilized is the PCSA, which operates in parallel with the Feed-Forward Network (FFN) of the frozen Transformer blocks. Following the PointGST protocol, input tokens T_{in} are transformed from the spatial domain to the spectral domain using the Graph Fourier Transform (GFT). Specifically, tokens are first down-projected to a lower dimension r via a learnable matrix W_d to generate a low-rank graph signal T_s, where $r \ll C$. We then apply the GFT by multiplying this signal with pre-computed spectral bases, effectively decomposing the spatial signals into orthogonal frequency components. This step is clinically pivotal, as it effectively decouples the holistic torso shape (low-frequency components) from pathological local protrusions (high-frequency components). In the spectral domain, a shared linear layer equipped with a residual connection and Swish activation function adapts these spectral coefficients. This process allows for targeted tuning by modifying the coefficients within the eigenvector space, effectively de-correlating confused tokens. Finally, an inverse GFT (iGFT) restores the features to the spatial domain, which are then up-projected via W_u, scaled by a factor s, and integrated back into the computation flow.

To enable robust spectral analysis of the dorsal topography, we constructed multi-scale point cloud graphs, comprising one global graph to capture the overall spinal alignment and k local sub-graphs to resolve fine-grained paraspinal details. The adjacency matrix $\mathcal{W}$ for these graphs was defined using a data-dependent scaling strategy, where edge weights $w_{i,j}$ are inversely related to the Euclidean distance between key points. We computed the Laplacian matrix $L = D - \mathcal{W}$ and performed eigenvalue decomposition to obtain orthogonal eigenvector matrices ($U_{\mathcal{G}}$ and $U_{\mathcal{L}}$). These eigenvectors served as the spectral bases for the GFT and were shared across all Transformer layers,

ensuring that the intrinsic geometric information of our specific point clouds was explicitly incorporated into the fine-tuning process.

We implemented the fine-tuning pipeline using the PyTorch framework on a single 12 GB NVIDIA GPU. The pre-trained backbone remained frozen throughout the training process. For optimization, we employed the AdamW optimizer with a weight decay of 0.05, coupled with a cosine learning rate scheduler that featured a 10-epoch warm-up phase. The initial learning rate was set to 5×10^{-4}, and the models were trained for 300 epochs with a batch size of 32. To balance performance with parameter efficiency, the adapter rank was set to 36. regarding data processing and local aggregation, we sampled 1,024 points per input object and configured the model with a group size of 32 and 64 groups. Additionally, gradient clipping with a threshold of 10 was applied to ensure training stability.

Training Strategy and Ensemble Inference

To ensure robust model performance and mitigate overfitting, distinct training strategies were adopted for the segmentation and classification tasks. For the segmentation network, a fixed data partitioning scheme was applied, dividing the dataset into a training set, an internal held-out test set, and an independent external multi-center test set. To establish high-quality ground truth, all segmentation masks underwent a secondary review by senior radiologists. This process involved manual refinement of 19 samples in which automated thresholding failed due to factors such as hair occlusion or clothing artifacts, ensuring anatomically accurate boundary annotations for model learning.

For the classification task, an ensemble learning strategy based on random sampling with replacement was employed to enhance model robustness. The development cohort was used to generate seven distinct expert models, each trained using a random split of 60% for training and 40% for internal validation. The number of expert models was chosen to ensure comprehensive data coverage—statistical estimation indicates that approximately six iterations are sufficient to sample the entire dataset—while the use of an odd number (n = 7) avoids tie outcomes during ensemble decision-making. Importantly, the multi-center validation cohort was kept completely isolated throughout training and hyperparameter optimization and was accessed only once for final performance evaluation, thereby providing a stringent assessment on previously unseen populations.

During the inference stage, a voting-based ensemble strategy was applied. Each expert model independently performed diagnostic classification, assigning the class with the highest predicted probability as its individual output. The final diagnosis was determined by majority voting across the seven models, with the category receiving the highest consensus designated as the final result. This ensemble approach effectively reduces prediction variance from individual models and yields a more stable and reliable screening decision.

Deciphering morphological signatures

To provide clinical interpretability and visualize the morphological features driving the AI's decision, we generated class activation heatmaps based on the PointGPT (Transformer-based) architecture. Unlike standard point-wise networks, our model processes the input point cloud by grouping local points into patches and utilizing a self-attention mechanism. To create the visualization, we extracted the self-attention maps from the final layer of the Transformer encoder.

Specifically, we calculated the attention weights between the global class token ([CLS]) and each local patch token. These weights quantify the contribution of each local geometric structure to the global shape descriptor. Since the model operates on grouped points, the patch-level attention scores were interpolated and mapped back to the original 3D spatial coordinates of the constituent points.

Statistical Analysis

Patient-level screening assessments were derived from the aggregated output of the ensemble deep learning framework. To ensure robust classification, we integrated predictions from seven independent expert models. The final risk probability for each participant was determined by the consensus voting strategy, which stabilizes the decision boundary against data heterogeneity. Participants were classified as positive for scoliosis when the aggregated probability exceeded the optimal decision threshold derived from the internal development set.

In evaluating the system's overall performance, we calculated a comprehensive set of diagnostic metrics, including Accuracy, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), F1-score, and the Matthews Correlation Coefficient (MCC). The discriminative capability of the model was further assessed using the Area Under the Receiver Operating Characteristic Curve (AUC) and Average Precision (AP). Additionally, for the automated ROI segmentation module, geometric fidelity was quantified using Point Accuracy (PA) and mean Intersection over Union (mIoU).

To estimate the statistical reliability of these estimates, 95% confidence intervals (CI) for all performance metrics were calculated using bootstrap resampling with 1,000 iterations, maintaining the original cohort sizes. All statistical analyses and metric calculations were performed using Python (version 3.8) with the scikit-learn (v1.0) and SciPy (v1.16.3) libraries.

Results

Overview of datasets

To develop and validate the radiation-free scoliosis screening system, we retrospectively analyzed data from 648 participants who underwent high-resolution 3D back surface topological scanning and matched full-spine radiography at Xinhua Hospital (Shanghai Jiao Tong University School of Medicine) between 2022 and 2024.After applying strict inclusion criteria—specifically age 8–18 years, no prior spinal surgery, and same-day acquisition of 3D back surface data and radiographic imaging—the internal development cohort comprised 648 distinct patients. Each patient contributed a single data sample to ensure independence and prevent data leakage. The study population reflected the natural epidemiology of adolescent idiopathic scoliosis (AIS), with a mean age of 13.55 ± 3.04 years and a female predominance (78.6%). Anthropometric measurements showed a mean height of 159.32 ± 10.74 cm and weight of 47.01 ± 10.42 kg. The cohort covered a broad spectrum of spinal deformities and Body Mass Index (BMI) values (range: 11.05–34.69 kg/m^2), which fell within normal variations, enabling the model to learn features robust to soft tissue differences. We randomly allocated 80% of these cases to the training set (n = 532) for model optimization and the remaining 20% to the internal testing set (n = 116) for performance evaluation.

Moreover, to assess the system's robustness against scanner variability and population shifts, we established an independent external validation cohort comprising 128 patients recruited from two

distinct institutions (The Third Affiliated Hospital of Sun Yat-sen University and Xin Miao Scoliosis Treatment Centre) between October 2022 and September 2024. This cohort followed the same inclusion criteria but utilized different operator teams and environmental setups. The final data division statistics and severity distribution across the training, internal testing, and external testing sets were rigorously managed to ensure balanced representation.

Automated ROI Segmentation and Preprocessing

A critical prerequisite for high-throughput, reproducible screening is the fully automated isolation of the dorsal Region-of-Interest (ROI) from raw 3D scans. Raw acquisition data invariably contain significant environmental noise (e.g., background artifacts) and non-relevant biological structures that can confound downstream diagnostic algorithms. To address this, we implemented a segmentation framework, an attention-based network selected for its ability to capture local geometric details within global shape contexts.

Performance evaluation was stratified across an internal test set and an independent external test set to verify algorithmic robustness. On the internal test set, the model demonstrated high segmentation fidelity, achieving a Point Accuracy (PA) of 92.57% (95% CI: 91.48–93.66%) and a dorsal Intersection over Union (IOU) of 89.70% (95% CI: 88.17–91.23%). Crucially, to assess clinical generalizability, we validated the pipeline on the external test set. The model maintained consistent performance with a PA of 91.83% (95% CI: 91.02–92.64%) and a dorsal IOU of 87.89% (95% CI: 87.20–88.58%), indicating minimal performance degradation across distinct data sources.

To validate the clinical applicability of these quantitative results, we conducted a qualitative assessment of segmentation fidelity. Examination of the IOU distribution and error patterns revealed that the pipeline precisely delineates anatomical boundaries—including the shoulders, waistline, and iliac crests. The method demonstrated high stability across subjects with varying morphological characteristics, effectively handling sparse point density and surface curvature largely reducing the need for manual intervention. This high-fidelity segmentation ensures that the subsequent diagnostic module receives standardized, noise-free topological input, decoupling diagnostic accuracy from variations in patient positioning or background clutter.

Binary Classification Performance and Clinical Generalizability

To operationalize the screening task, we defined the binary classification criteria as scoliosis (Cobb angle >10°) versus non-scoliosis (0°–10°). Upon internal validation, the deep learning framework demonstrated high discriminative fidelity. As illustrated in the normalized confusion matrices (Fig. 2), the ensemble voting strategy correctly identified all positive and negative samples, resulting in an accuracy of 100% and a corresponding Area Under the Curve (AUC) of 1.00 (Fig. 3a). This performance indicates that the model successfully captured the distinct morphological features of the spinal point clouds without overfitting, establishing a robust baseline for subsequent external testing.

To rigorously evaluate the model's utility in a real-world setting, we deployed the trained ensemble on an independent external validation set. The model exhibited strong generalization capabilities, maintaining a high classification accuracy of 99.24% (95% CI: 97.66%–100.00%). The Receiver Operating Characteristic (ROC) analysis further confirmed this robustness, yielding an AUC of 99.54% (95% CI: 98.32%–100.00%) (Fig. 3b). Notably, the ensemble voting strategy—which

aggregates predictions from seven independent expert models—proved superior to individual models, effectively mitigating variance and stabilizing the decision boundary against data heterogeneity inherent in multi-center datasets (Fig. 4 and Fig. 3b).

In the external validation cohort, the model achieved both a sensitivity (recall) and a Negative Predictive Value (NPV) of 100.00%, identifying all patients with a Cobb angle >10°. Accompanying this sensitivity, the model maintained a specificity of 98.93% (95% CI: 96.33%–100.00%) and a Positive Predictive Value (PPV) of 97.40%. Error analysis via confusion matrices (Fig. 4) confirmed only a single false-positive instance across the entire external dataset.

The comprehensive statistical evaluation underscores the reliability of the proposed method. The F1-score, which harmonizes precision and recall, reached 98.67% for the external set, while the Matthews Correlation Coefficient (MCC) was 98.16%, indicating a balanced prediction performance even if class imbalances were to exist. The consistency observed between the individual expert models and the final voting outcome (Figs. 2, 3 and 4) confirms that the proposed point cloud analysis pipeline provides a stable, reproducible, and highly accurate solution for automated scoliosis detection. However, detection represents only the first stage of the triage pipeline; effective clinical management necessitates not only identification but also precise quantification of the deformity to guide specific treatment protocols.

Fine-grained Severity Stratification and Clinical Grading

To advance the framework's utility from initial screening to precise clinical grading, we extended the evaluation to a fine-grained five-grade stratification task. This progression confirmed that the model maintains high discriminative fidelity across severity gradients, crucially preserving the robust identification of non-scoliosis cases (0–10°) established in the binary baseline even under increased classification complexity.

When evaluated on the external dataset, the model maintained robust performance with a macro-average AUC of 94.86%, demonstrating effective generalization to unseen data distributions. Analysis of the prediction errors revealed that misclassifications were predominantly confined to immediately adjacent severity grades. For instance, while the sensitivity for the 21–30° range was 53.17%, misclassified cases were primarily predicted as 11–20° or 31–40°, rather than being categorized as extreme outliers. This suggests that the model effectively captures the continuous morphological progression of the spine, with errors likely reflecting the inherent ambiguity often found near the boundaries of Cobb angle definitions.

Despite the challenges inherent in multi-class classification on limited datasets, the model showed promising specificity in ruling out severe progression. For the most severe category (>40°), which typically necessitates surgical consideration, the model achieved a specificity of 98.42% (95% CI: [96.00%–100.00%]) in the external cohort. The One-vs-Rest ROC analysis further supported this, with the AUC for the >40° class reaching 98.39%. The ensemble voting strategy played a pivotal role here; by aggregating predictions across folds, it smoothed out the decision boundaries, resulting in a more stable grading system that minimized catastrophic errors compared to individual model runs.

To further validate the model's utility in clinical decision-making, we derived binary performance metrics based on the 5-class probability distributions at two critical intervention thresholds: Cobb

angles of 20° (observation vs. bracing) and 40° (bracing vs. surgery). By aggregating the posterior probabilities of classes above these thresholds, the 5-class model achieved high diagnostic precision for these derived tasks. For the 20° threshold on the external test set, the system demonstrated a sensitivity of 62.73% (95% CI: 53.28%–72.19%) and a specificity of 94.20% (95% CI: 92.43%–95.98%); similarly, for the critical 40° surgical threshold, the derived classification yielded an impressive AUC of 92.89% (95% CI: 85.33%–100.00%). These results indicate that the fine-grained training supervision provides a rich feature representation that can be flexibly adapted to various clinical cutoff requirements without retraining.

The 5-class stratification results highlight the framework's potential as an automated grading assistant. While the exact binning of intermediate cases remains a challenging task due to the continuous nature of spinal deformation, the high specificity across all classes and the robust identification of the non-scoliosis baseline confirm the system's reliability. The overall accuracy of 84.48% on the external dataset, combined with a weighted F1-score of 83.75%, suggests that the model provides a reliable "second opinion" for quantifying severity, effectively flagging cases that require prioritized review by specialists.

Interpretability and Anatomical Alignment

To foster clinical trust and validate that the classification module leverages biologically relevant features rather than spurious correlations (e.g., background artifacts or clothing), we visualized the model's decision-making process using point-wise saliency mapping. As illustrated in Fig. 5, we present a comparative visualization of the input 3D point cloud back surface alongside the corresponding ground-truth X-ray radiograph.

The attention heatmaps reveal that the model consistently assigns high importance weights to paravertebral regions exhibiting pathological asymmetry, effectively filtering out irrelevant peripheral areas. Specifically, the model’s focus is concentrated on key orthopedic markers: the scapular prominence, the lateral waistline asymmetry, and the paraspinal rib hump. Overlaying these attention maps with the spinal radiographs demonstrates a high degree of spatial alignment between the AI-highlighted surface regions and the underlying skeletal deformity. The "hotspots" on the 3D surface correspond intimately with the spinal curvature and the convex side of the rib cage visible on the X-ray.

This qualitative alignment is clinically significant, as the rib hump is the direct topological manifestation of vertebral rotation—a critical biomechanical driver of scoliosis progression. By focusing on these anatomical indicators rather than global image statistics, the classification module provides clinicians with a visual rationale alongside the risk prediction, thereby supporting informed referral decisions based on established markers of deformity.

Discussion

The clinical management of adolescent idiopathic scoliosis (AIS) has historically been trapped in a reactionary cycle, paralyzed by a critical conflict: the imperative for early detection versus the carcinogenic risks of cumulative radiation exposure. Current screening pathways are bottlenecked by reliance on subjective physical examinations (e.g., the Adam’s Forward Bend Test), which suffer from high inter-rater variability and frequently result in missed early-stage diagnoses.[21] Conversely,

definitive radiographic confirmation imposes unnecessary radiation burdens on healthy adolescents.[22] To resolve this historic impasse, we present a radiation-free, fully automated screening framework that transforms 3D back surface point clouds into objective diagnostic biomarkers. By bridging the gap between safety and accuracy, our work shifts the clinical paradigm from intermittent, reactive diagnosis based on visible deformity to continuous, proactive surveillance.

Our study makes three distinct contributions to the field of orthopedic AI. Our findings primarily establish the system's utility as a high-reliability clinical gatekeeper. Aligning with the urgent need for high-sensitivity triage, PointScol achieved 100% sensitivity at the primary screening threshold of 10° across multi-center cohorts. This operationalizes a "near-zero false-negative" strategy, ensuring that healthy individuals are safely excluded from the referral pipeline without radiographic confirmation. In healthcare systems where cost-efficiency is paramount, this automated triage significantly reduces the economic burden of unnecessary X-rays and specialist consultations, while simultaneously alleviating the substantial psychological distress patients and families experience during "false alarm" referrals.

Methodologically, our approach addresses the limitations of previous computer vision techniques by prioritizing geometric fidelity over texture. Unlike 2D photographic models that suffer from the 'dimensionality gap' and are easily confounded by lighting or skin texture, our direct 3D processing captures the intrinsic topological asymmetry essential for volumetric analysis. This strategy dovetails with a growing consensus in high-precision computational medicine: that direct geometric processing—operating on raw point clouds or surface meshes—far surpasses voxel-based or multi-view approaches in resolving complex topologies. A prime example lies in digital dentistry, where geometric deep learning networks (e.g., MeshSegNet) have successfully replaced manual segmentation by directly analyzing the curvature and normal vectors of intraoral scan data.[23] Similarly, in vascular diagnostics, point-based architectures like PointNet++ are now deployed to map coronary centerlines with sub-millimeter precision, bypassing the resolution loss inherent in volumetric grids.[24] By translating these geometric principles to the spinal dorsal surface, our PointScol framework confirms that raw coordinate data contains the richest diagnostic signal, uncorrupted by projection artifacts.

We address the "black box" skepticism that limits the adoption of AI in specialized diagnostics.[25] Prioritizing clinical interpretability alongside accuracy, our system incorporates geometric feature visualization to reveal the specific regions of asymmetry driving the diagnosis. This transparency fosters the clinical trust necessary for medical integration. Additionally, we validated our model on a large-scale, multi-center benchmark of paired 3D and X-ray data, confirming that clinically actionable insights can be derived from non-invasive morphological data alone, offering a richer information profile than previous alternatives.

Despite these advancements, several limitations warrant careful consideration. While the proposed five-category classification framework is intended to provide granular risk stratification, its effectiveness is inherently dependent on adequate and balanced data coverage across the full Cobb angle spectrum. In the present dataset, uneven representation within certain angle intervals restricts the model's ability to achieve uniformly high performance across all categories. Nevertheless, from a clinical standpoint, the primary objective remains the reliable binary discrimination between

normal and pathological cases, a task for which our model demonstrated consistently strong performance. Building on this foundation, future work will prioritize expanding and balancing the training dataset across underrepresented Cobb angle ranges to further improve fine-grained multi-category classification.

A key limitation is that the dataset was derived exclusively from a Chinese population, which may limit generalizability given potential ethnic differences in BMI distribution and trunk morphology. Moreover, despite the use of a multicenter cohort, the overall sample size remains limited, and further expansion across additional centers and populations will be necessary to more rigorously assess the robustness and generalizability of the proposed approach in large-scale prospective studies.

The radiation-free design allows safe longitudinal disease monitoring, addressing the clinical need for frequent assessments during growth-related progression in adolescent scoliosis. By mitigating the risks associated with cumulative radiographic exposure, this approach not only enhances patient safety but also supports a paradigm shift toward more accessible, non-invasive screening strategies, with profound implications for health equity in scoliosis management.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Research on prevention and treatment of common frequently-occurring diseases / 2023YFC2507702，2023YFC2507706)

Author contributions

J.L.Y., L.A., and Y.Q.Z. conceived and supervised the project. J.F.Y., Z.H., T.Z., Z.Z., Z.F., X.X., Y.G., and Y.D. collected the original 3D imaging data and clinical records. J.F.Y. and Z.H. annotated the data. H.S., X.W., W.W., J.W., Y.Z., H.L., J.Z., and L.T. developed the algorithms and wrote the code. H.S. and X.W. performed the validation and statistical analysis. J.L.Y. and Y.Q.Z. provided clinical guidance and result calibration. L.W. and X.L. provided technical guidance. J.F.Y., H.S., Z.H., and X.W. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Zhang, Z. *et al.* Intelligent Scoliosis Screening and Diagnosis: A Survey.
2. Kim, H., Chang, B.-S. & Chang, S. Y. Current issues in the treatment of adolescent idiopathic scoliosis: a comprehensive narrative review. *Asian Spine J.* **18**, 731–742 (2024).
3. Li, M., Nie, Q., Liu, J. & Jiang, Z. Prevalence of scoliosis in children and adolescents: a systematic review and meta-analysis. *Front. Pediatr.* **12**, 1399049 (2024).
4. Kim, D. J. *et al.* The diagnostic accuracy of community spine radiology for adolescent idiopathic scoliosis brace candidates. *Eur. Spine J.* **33**, 3776–3783 (2024).
5. Goldberg, C. J., Moore, D. P., Fogarty, E. E. & Dowling, F. E. Scoliosis: a review. *Pediatr. Surg. Int.* **24**, 129–144 (2008).
6. Lai, K. K.-L. *et al.* Monitoring of Curve Progression in Patients with Adolescent Idiopathic Scoliosis Using 3-D Ultrasound. *Ultrasound Med. Biol.* **50**, 384–393 (2024).

7. Chen, H. *et al.* Development of Automatic Assessment Framework for Spine Deformity Using Freehand 3-D Ultrasound Imaging System. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **71**, 408–422 (2024).
8. Zhu, X. *et al.* MGScoliosis: Multi-grained scoliosis detection with joint ordinal regression from natural image. *Alex. Eng. J.* **111**, 329–340 (2025).
9. He, Z. *et al.* Conditional generative adversarial network-assisted system for radiation-free evaluation of scoliosis using a single smartphone photograph: a model development and validation study. *eClinicalMedicine* **75**, 102779 (2024).
10. Ćuković, S. *et al.* Advancing 3D idiopathic scoliosis assessment through optical scans and web-based integration - Validation of the ScolioSIM system. *Comput. Biol. Med.* **196**, 110735 (2025).
11. Mohamed, N. *et al.* Three-dimensional markerless surface topography approach with convolutional neural networks for adolescent idiopathic scoliosis screening. *Sci. Rep.* **15**, 8728 (2025).
12. Zhou, Z. *et al.* Gait Patterns as Biomarkers: A Video-Based Approach for Classifying Scoliosis. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* (eds Linguraru, M. G. et al.) vol. 15005 284–294 (Springer Nature Switzerland, Cham, 2024).
13. Dong, N., Zhang, X., Liu, X., Guo, W. & Wang, F. Key Points Positioning: A Two-Stage Algorithm For Single-view Point Cloud of Human Back Based on Point-wise Network. in *Proceedings of the 2022 11th International Conference on Computing and Pattern Recognition* 264–272 (ACM, Beijing China, 2022). doi:10.1145/3581807.3581846.

14. Rothstock, S., González-Ruiz, J. M., Weiss, H.-R., Turnbull, D. & Krueger, D. Severity and Cobb angle of scoliosis patients quantified with markerless trunk surface topography using k-NN search and multivariate regression analysis. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **11**, 2433–2439 (2023).

15. Goldman, S. N. *et al.* Applications of artificial intelligence for adolescent idiopathic scoliosis: mapping the evidence. *Spine Deform.* **12**, 1545–1570 (2024).

16. Amirian, S. *et al.* Explainable AI in Orthopedics: Challenges, Opportunities, and Prospects. Preprint at https://doi.org/10.48550/ARXIV.2308.04696 (2023).

17. CloudCompare. CloudCompare. (2024).

18. Wu, X., Lao, Y., Jiang, L., Liu, X. & Zhao, H. Point Transformer V2: Grouped Vector Attention and Partition-based Pooling. Preprint at https://doi.org/10.48550/arXiv.2210.05666 (2022).

19. Liang, D. *et al.* Parameter-Efficient Fine-Tuning in Spectral Domain for Point Cloud Learning. Preprint at https://doi.org/10.48550/arXiv.2410.08114 (2025).

20. Pang, Y. *et al.* Masked Autoencoders for Point Cloud Self-supervised Learning. Preprint at https://doi.org/10.48550/arXiv.2203.06604 (2022).

21. Kadhim, M. *et al.* Current status of scoliosis school screening: targeted screening of underserved populations may be the solution. *Public Health* **178**, 72–77 (2020).

22. Luan, F.-J., Wan, Y., Mak, K.-C., Ma, C.-J. & Wang, H.-Q. Cancer and mortality risks of patients with scoliosis from radiation exposure: a systematic review and meta-analysis. *Eur. Spine J.* **29**, 3123–3134 (2020).

23. Lian, C. *et al.* Deep Multi-Scale Mesh Feature Learning for Automated Labeling of

Raw Dental Surfaces From 3D Intraoral Scanners. *IEEE Trans. Med. Imaging* **39**, 2440–2450 (2020).

24. Banerjee, A. *et al.* Point-Cloud Method for Automated 3D Coronary Tree Reconstruction From Multiple Non-Simultaneous Angiographic Projections. *IEEE Trans. Med. Imaging* **39**, 1278–1290 (2020).

25. Hettikankanamage, N. *et al.* eXplainable Artificial Intelligence (XAI): A Systematic Review for Unveiling the Black Box Models and Their Relevance to Biomedical Imaging and Sensing. *Sensors* **25**, 6649 (2025).

Figures and Legends

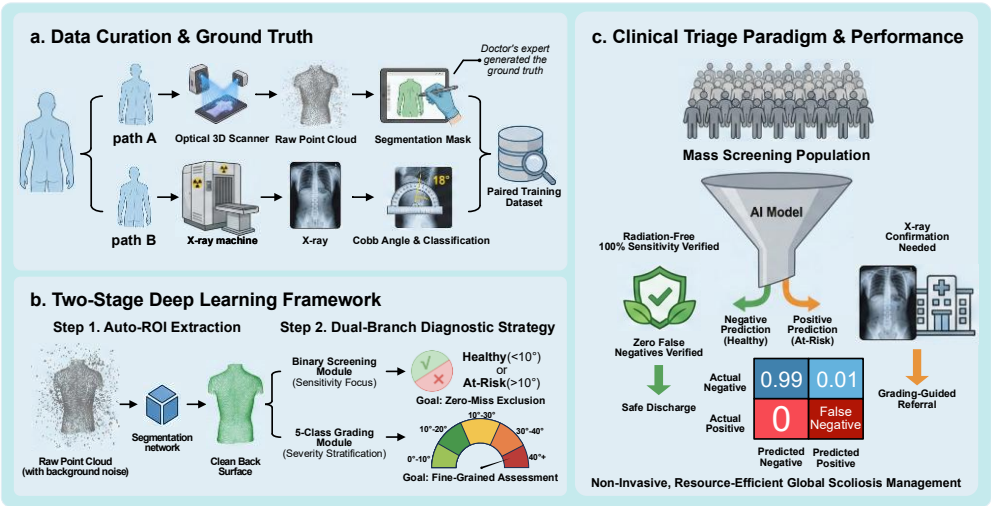

Figure 1: **Overview of the radiation-free AIS screening framework based on 3D point clouds. (a)** Construction of a multi-center dataset pairing radiation-free 3D point clouds with X-ray ground truth labels. **(b)** The two-stage architecture: An automated ROI extraction module standardizes the input, followed by a dual-branch strategy featuring a Binary Screening Module for safety and a 5-Class Grading Module for severity assessment. **(c)** The hierarchical clinical triage workflow. The system first acts as a "zero-miss" filter to safely exclude healthy individuals (100% sensitivity), then stratifies risk patients to guide referral urgency (e.g., identifying surgical candidates), maximizing resource efficiency.

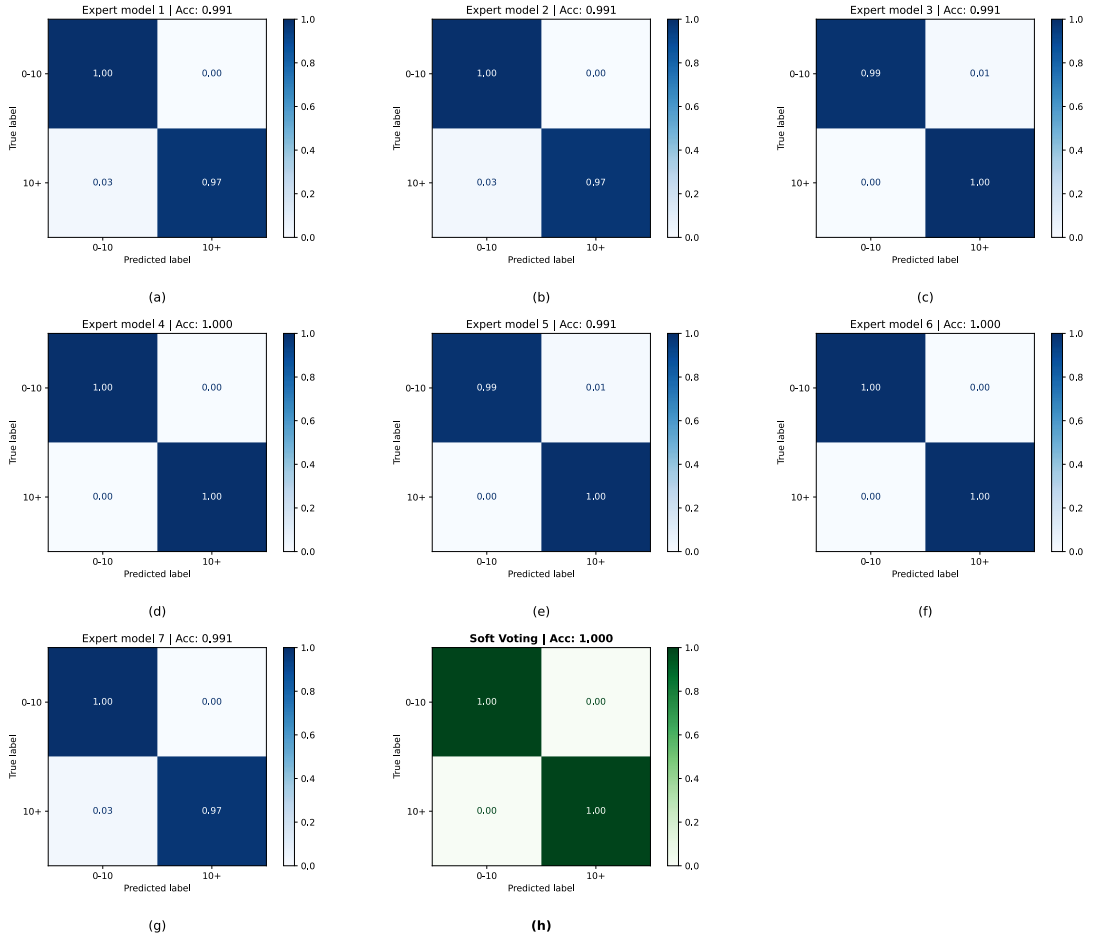

Figure 2: **Normalized confusion matrices evaluating classification accuracy on the internal dataset.** Panels (a–g) display the performance of the seven individual expert models (blue color scale), while panel (h) illustrates the final majority voting integration (green color scale) for the binary classes '0–10' and '10+'. The values within each cell represent the proportion of true labels classified into each predicted category, where diagonal elements indicate correct predictions. Overall accuracy (Acc) is denoted above each subplot. Notably, the majority voting mechanism (h) aggregates the experts' decisions to mitigate individual misclassifications, achieving perfect accuracy (Acc: 1.000).

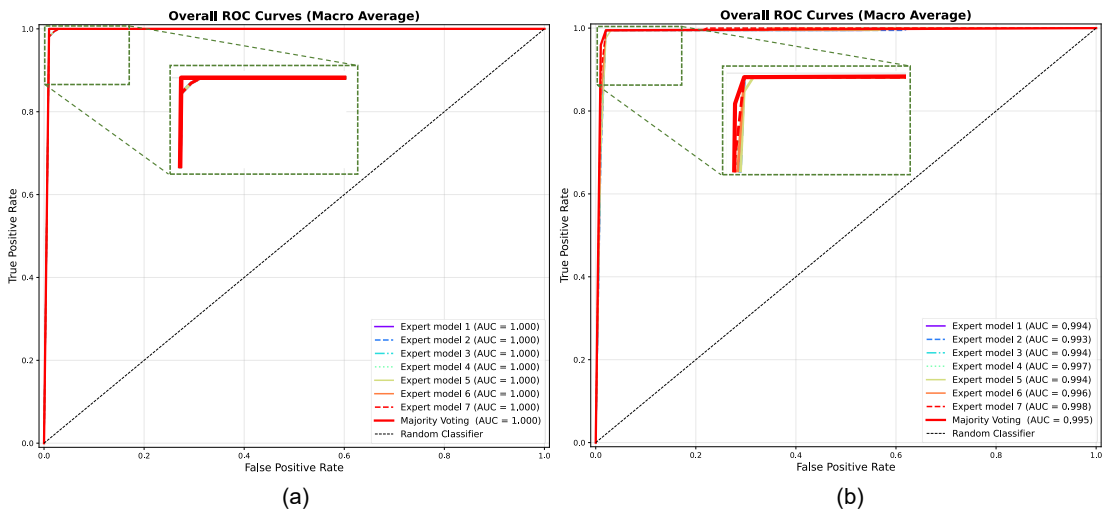

Figure 3: **Performance evaluation of expert models and ensemble voting strategy in binary classification.** a, b, Macro-average Receiver Operating Characteristic (ROC) curves derived from the internal testing set (a) and the external testing set (b). The plots compare the classification performance of seven individual expert models (colored dashed lines) against the final majority voting ensemble model (solid red line). The insets (green dashed boxes) provide magnified views of the top-left corner to highlight performance distinctions in the high-sensitivity and high-specificity regions. Area Under the Curve (AUC) values for each model are listed in the legend, demonstrating the robustness of the ensemble approach across both datasets.

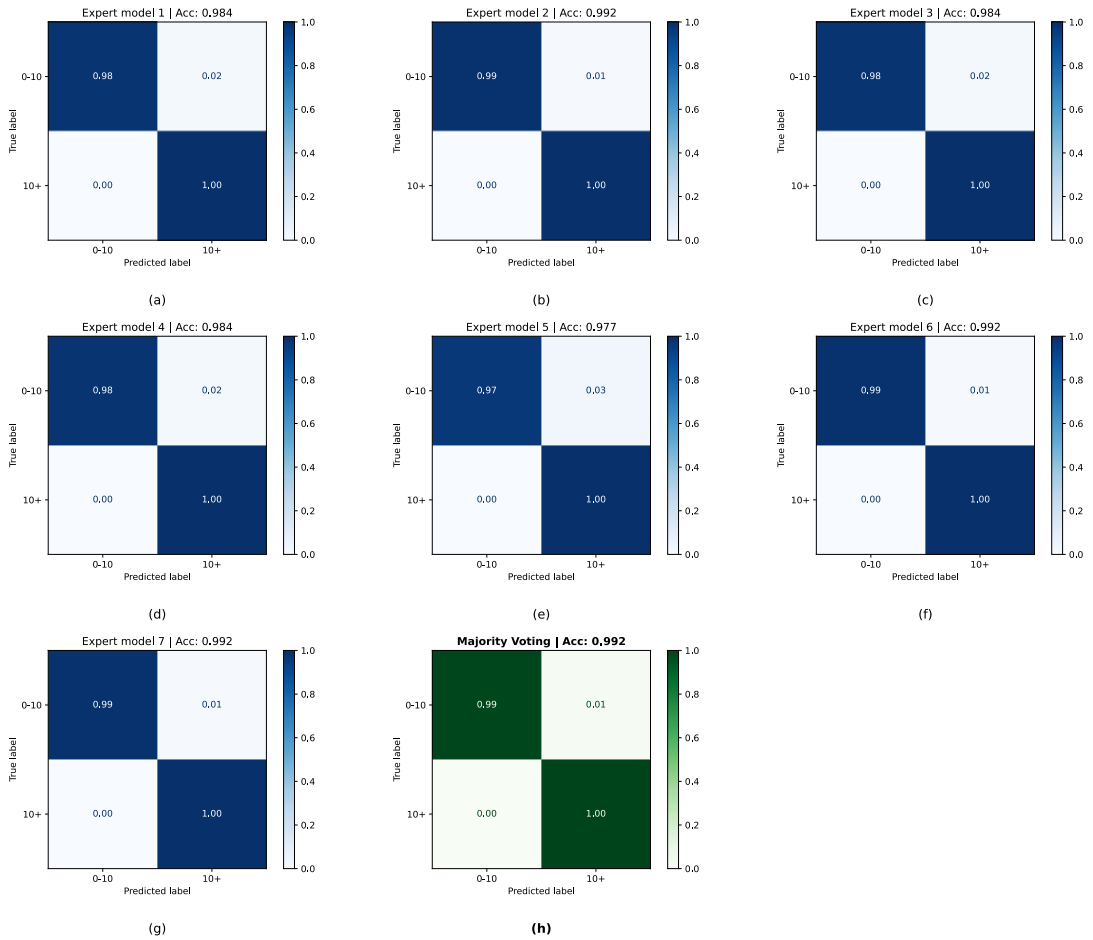

Figure 4: **Assessment of generalization capability on an independent external test cohort.** Confusion matrices illustrate the predictive performance of (a–g) the seven individual expert models and (h) the majority voting ensemble on unseen data. While individual experts exhibit variability in sensitivity or specificity due to data distribution shifts, the majority voting strategy leverages consensus to maintain high diagnostic stability and accuracy.

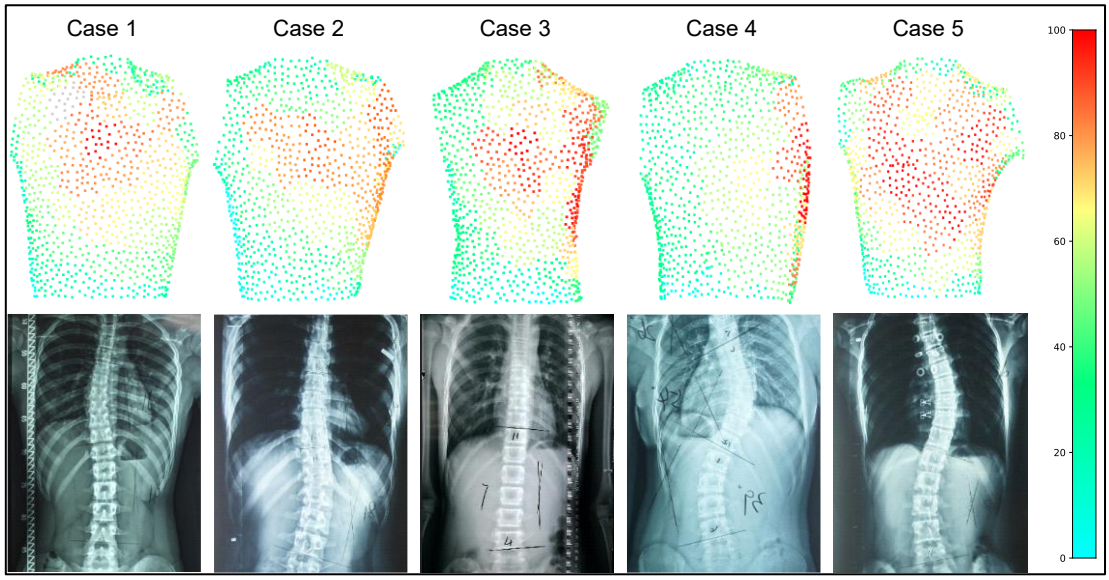

Figure 5: **Visualization of model attention regions compared to anatomical ground truth.** Heatmap visualizations of attention regions for the classification model are presented. The first row displays the point cloud attention maps generated by the model, where warmer colors (red) indicate higher model attention. The second row shows the corresponding X-ray images with clinically annotated scoliosis regions.