

SpineTrace: Auditable Video-Based Scoliosis Screening with Temporal Pose Statistics

Teddy Surya Gunawan¹, Siti Aishah Sofea Asri², Mira Kartiwi³, Noreha Abdul Malik⁴,
Ahmad Hud Mahfoodz⁵, Mohd Ghaddafi Wahab⁶

^{1,2,4}Department of Electrical and Computer Engineering, International Islamic University Malaysia

³Department of Information Systems, International Islamic University Malaysia

⁵Department of Orthopaedics, Traumatology and Rehabilitation, Kulliyyah of Medicine,
International Islamic University Malaysia

⁶Department of Surgery, Kulliyyah of Medicine, International Islamic University Malaysia

Article Info

Article histories:

Received Jun 14, 2026

Revised Aug 21, 2026

Accepted Sep 3, 2026

Keywords:

Scoliosis screening

Gait analysis

Temporal pose statistics

Neutral class

Interpretable machine learning

Support Vector Machine

ABSTRACT

Video-derived gait representations may support scalable pre-referral screening for scoliosis, but their evaluation must retain borderline cases, quantify model vulnerability, and distinguish measured performance from clinical or deployment readiness. This study evaluates SpineTrace as an auditable compact engineering baseline rather than a novel learning algorithm or state-of-the-art classifier. The public Scoliosis1K release comprised 1,493 gait sequences reportedly obtained from 1,050 adolescents, including 493 positive, 200 neutral, and 800 negative sequences. The supplied partition of 744 training and 749 test sequences was reproduced exactly. Feature scaling and hyperparameter selection were restricted to five stratified folds within the training partition, with three-class classification designated as the primary task. Among the reproduced RBF-SVM configurations, the highest macro-F1 was obtained using 102 temporal pose features derived from keypoint means and standard deviations. The model achieved 0.737 accuracy, 0.740 balanced accuracy, 0.702 macro-F1, and 0.890 macro one-versus-rest ROC-AUC. Recall was 0.917, 0.619, and 0.685 for negative, neutral, and positive sequences, respectively, demonstrating the greater difficulty of borderline and positive cases. Adding three silhouette and five postural descriptors reduced macro-F1 to 0.656, indicating that semantically plausible handcrafted features did not provide complementary predictive information. Independently retuned PAV-SVM and DTW-*k*NN baselines achieved macro-F1 scores of 0.545 and 0.473, while a compact ST-GCN adaptation achieved 0.645 ± 0.055 across three optimization seeds. Removing keypoint-confidence features from the full model reduced macro-F1 from 0.656 to 0.618, and coordinate jitter equivalent to 2% of torso length reduced it to 0.287, identifying pose quality as a critical vulnerability. Fixed-test uncertainty was estimated using 2,000 class-stratified sequence bootstraps. Because the local release lacks participant mappings and raw RGB videos, participant independence, cross-estimator transfer, end-to-end latency, and cross-dataset validity could not be established. SpineTrace therefore provides a reproducible sequence-level baseline for auditing feature utility and failure boundaries, not a diagnostic or deployment-ready screening system.

Copyright © 2026 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Teddy Surya Gunawan,
Department of Electrical and Computer Engineering,
International Islamic University Malaysia,
Jalan Gombak, 53100 Kuala Lumpur.
Email: tsgunawan@iiu.edu.my

1. INTRODUCTION

Adolescent idiopathic scoliosis (AIS) is a three-dimensional spinal deformity that develops during growth and is conventionally defined by a coronal Cobb angle of at least 10° , accompanied by vertebral rotation. Although prevalence varies across populations and diagnostic thresholds, reported estimates generally range from 0.47% to 5.2% [1]. A large school-based study involving 16,045 adolescents reported an overall prevalence of 2.3%, increasing to 3.1% among girls [2]. Applied to regional or national school populations, even these modest percentages translate into a substantial screening burden. Timely recognition is important because curve progression is associated with skeletal growth, and adolescents identified before maturity may remain eligible for observation, scoliosis-specific exercise, or bracing intended to limit further progression [3].

Standing spinal radiography remains the clinical reference for confirming scoliosis, measuring the Cobb angle, and monitoring curve progression. However, it is unsuitable for indiscriminate use as a first-line population-screening test. Radiography requires specialized equipment and professional interpretation and exposes the individual to ionizing radiation. Although the dose from a single contemporary examination may be low, adolescents undergoing surveillance can accumulate exposure through repeated full-spine examinations. In a historical AIS cohort, patients received an average of 16 radiographs during treatment and follow-up, illustrating the importance of clinically justifying and minimizing cumulative exposure [4]. A non-radiographic method should therefore be positioned as a triage tool preceding clinical assessment, not as a replacement for indicated radiography. Its potential value lies in identifying individuals who warrant professional review while reducing unnecessary radiographic referrals among those at low apparent risk.

Current school-screening practices commonly combine visual inspection, the Adam forward-bend test, and scoliometer measurement. These procedures are inexpensive, but their effectiveness depends on examiner training, measurement thresholds, and referral criteria. The forward-bend test used alone may produce excessive referrals and has limited positive predictive value, whereas adding an objective measurement of trunk rotation can improve screening performance [3, 5]. Population screening therefore involves an important trade-off: inadequate sensitivity may delay recognition of a progressive curve, while poor specificity can create avoidable anxiety, clinical workload, and radiographic examination. This tension motivates objective, repeatable, and non-contact measurements that can supplement established screening pathways.

Back-view gait video is attractive for this purpose because it captures posture and movement using an ordinary camera without ionizing radiation or body-mounted markers. Walking may reveal persistent or cyclic asymmetries involving the shoulders, trunk, pelvis, and limbs that are less evident in a single static image. Video also enables postural behavior to be summarized over time rather than inferred from a single frame. Recent studies have consequently investigated gait silhouettes and skeletal trajectories as privacy-preserving representations for scoliosis screening. Deep silhouette networks, clinically guided pose representations, and graph-based skeletal models have demonstrated that gait contains discriminative information relevant to scoliosis classification [6-8]. Nevertheless, an association between gait features and a screening label does not establish that a model measures spinal curvature or can support diagnosis.

The Scoliosis1K project created an important benchmark for investigating this problem. Its initial release comprises 447,900 silhouette frames from 1,493 walking sequences reportedly collected from 1,050 adolescents in a school-screening setting. The sequences are assigned to positive, neutral, and negative screening categories, making borderline cases explicit rather than removing them from a binary endpoint [7]. The subsequent Scoliosis1K-Pose extension provides 17 keypoints per frame in the MS-COCO format, extracted from the same videos using ViTPose [8]. These representations facilitate reproducible investigation without distributing identifiable raw RGB videos. However, the locally available release does not include participant identifiers linking multiple sequences to the same individual. Consequently, the published sequence partition can be reproduced, but participant-level independence between the training and test sets cannot be verified from the released files.

Pose-based screening also introduces measurement uncertainty. A keypoint coordinate is an output of a computer-vision estimator rather than a direct anatomical measurement. Its position and confidence may be affected by viewpoint, self-occlusion, clothing, image resolution, crop placement, illumination, and the pose-estimation architecture. A classifier may consequently exploit estimator behavior or acquisition artifacts that correlate with the labels without representing scoliosis-related anatomy. Scale normalization can reduce sensitivity to camera distance, but it does not guarantee invariance to camera rotation, perspective distortion, or systematic keypoint error. Evaluation using only unperturbed benchmark features may therefore overstate robustness, especially when the intended implementation employs a pose estimator different from the one used to construct the dataset.

SpineTrace addresses this setting through a compact sequence-level representation followed by a radial-basis-function support vector machine (RBF-SVM). The representation comprises 102 temporal pose statistics derived from the means and standard deviations of 17 keypoints and their associated confidence values,

together with three silhouette descriptors and five scale-normalized postural variables. This design offers several engineering advantages: every input has an identifiable computational source, feature groups can be inspected or removed systematically, and classifier inference is substantially less demanding than end-to-end video processing. These properties make SpineTrace suitable as an auditable engineering baseline. They do not, however, constitute a new learning algorithm or independently establish clinical validity, participant-level generalization, or deployment readiness.

Accordingly, this study evaluates whether a compact, clinically motivated gait representation can provide a reproducible scoliosis-screening baseline under more demanding and transparent evaluation conditions. Three-class discrimination among positive, neutral, and negative sequences is designated as the primary task because it retains the borderline cases most relevant to screening uncertainty. Two secondary tasks are also examined: a pre-referral formulation in which positive and neutral sequences are grouped as requiring review, and the legacy positive-versus-negative classification setting. All preprocessing, feature scaling, hyperparameter selection, and model selection are restricted to the training data. Throughout this paper, screening refers to assigning a research flag for subsequent professional assessment; it does not imply diagnostic confirmation, Cobb-angle estimation, or autonomous referral.

The principal contributions are fourfold. First, the study audits the released sequence partition, label distribution, and pose-estimator provenance, thereby establishing the sequence as the precise evidential unit supported by the available data. Second, it provides independently tuned comparisons across five conventional classifier families, along with PAV-SVM, DTW- (k) NN, and a compact ST-GCN adaptation, all evaluated on the same local partition. Third, it examines model behavior through class-specific analysis of neutral sequences, feature-group removal, repeated model-agnostic attribution, coordinate-noise perturbation, frame-loss testing, and explicit assessment of keypoint-confidence information. Fourth, it provides reproducible code, configurations, sequence-level predictions, fitted models, figures, and file hashes. The study does not seek to outperform higher-capacity attention-based or multistream systems. Instead, it establishes SpineTrace as an auditable sequence-level baseline that identifies which compact gait summaries remain informative, which feature additions fail to provide complementary evidence, and how rapidly performance deteriorates when pose quality is disrupted. Participant linkage, cross-estimator transfer, end-to-end video processing, and external validation remain unresolved requirements for clinical or field evaluation.

2. RELATED WORK

Related research can be organized into four connected areas: clinical and nonradiographic screening, classical temporal modeling, deep gait representation learning, and scoliosis-specific video systems.

2.1. Clinical and Nonradiographic Screening

Clinical examination, surface imaging, ultrasound, and photography provide different types of evidence and should not be treated as interchangeable measurements. The Adam forward-bend test and angle-of-trunk-rotation measurement are intended primarily to support referral decisions, but their reliability depends on examiner technique and the selected threshold. Surface topography and ultrasound can provide radiation-free geometric information, although they require specialized acquisition procedures and do not directly reproduce the radiographic Cobb angle. Conventional photography and markerless pose estimation further reduce the hardware burden, but they measure projected external posture rather than vertebral alignment or rotation. Correlations between selected photographic and radiographic variables support the potential value of camera-based screening, but they do not establish equivalence with radiography [9, 10]. Consequently, gait-derived scores should be interpreted as pre-referral screening signals rather than radiographic measurements or diagnostic outputs.

Pose estimation introduces an additional measurement layer between the recorded movement and the classifier. ViTPose uses a high-capacity vision-transformer backbone, whereas BlazePose was developed for efficient, real-time tracking on mobile devices [11, 12]. These estimators differ in architecture, landmark definitions, training data, confidence calibration, and error distribution. Their outputs therefore cannot be assumed to be interchangeable without paired evaluation on the same videos. This distinction is particularly important in the present study: the released Scoliosis1K-Pose tensors were generated using ViTPose, while the raw RGB videos required for an independent MediaPipe or BlazePose rerun were unavailable. Consequently, the present results characterize classification from ViTPose-derived sequences and do not demonstrate transfer to another pose estimator.

2.2. Classical and Deep Temporal Gait Modeling

Dynamic time warping (DTW) and hidden Markov models (HMMs) provide established approaches to temporal sequence analysis. DTW aligns two sequences under monotonic temporal deformation and is

therefore suitable when walking speed, sequence length, or gait phase varies. HMMs instead model observations through latent states and class-conditional transition probabilities [13, 14]. Their relative transparency does not eliminate the need for leakage-safe validation. DTW templates, neighborhood size, HMM topology, state number, and observation distributions must all be selected using training data alone. Their computational requirements must also be measured rather than inferred from their age or conceptual simplicity.

Modern gait-recognition architectures provide stronger representation-learning mechanisms, although most were developed for biometric identification or action recognition rather than clinical screening. GaitSet treats silhouette frames as an unordered set and therefore emphasizes sequence-level appearance while discarding temporal order. GaitGL combines global and local spatial features with short-range temporal aggregation. ST-GCN represents skeletal sequences as graphs whose spatial edges encode anatomical connections and whose temporal edges connect corresponding joints across frames [15-17]. These approaches motivate useful architectural controls, but their published identity- or action-recognition results cannot be interpreted as evidence for scoliosis screening. Clinical comparison requires the same input release, label definitions, partition, tuning procedure, and evaluation metrics. Accordingly, the present study evaluates a compact ST-GCN adaptation on the same three-class sequence partition but treats GaitSet and GaitGL only as architectural context.

2.3. Scoliosis-specific gait systems

ScoNet and ScoNet-MT introduced Scoliosis1K and established large-scale silhouette-based gait classification for positive, neutral, and negative screening categories. ScoNet-MT combines scoliosis classification with gait-identification supervision, using set-based silhouette processing to learn discriminative representations [7]. The later Scoliosis1K-Pose extension provides 17-keypoint ViTPose sequences from the same recordings. Its Dual Representation Framework (DRF) combines skeleton maps with a 24-element Postural Asymmetry Vector (PAV), while PAV-guided attention directs representation learning toward clinically motivated asymmetries. DRF reported 86.0% accuracy and 80.8% macro-F1 under its three-class protocol, demonstrating the value of PAV as an attention prior rather than proving that PAV alone is sufficient [8]. The present study therefore evaluates PAV-SVM as a separate handcrafted baseline; it does not describe this baseline as a reproduction of DRF.

Subsequent systems have emphasized temporal alignment, borderline cases, and multimodal fusion. TG-MILNet uses DTW-based clustering to organize gait phases, inter-bag temporal attention to identify informative intervals, and a boundary-aware component to improve neutral-case recognition [18]. ScoliGaitX combines silhouettes, optical flow, and gait-energy-image differences through gated multistream fusion [19]. These models illustrate how temporal and motion representations can improve scoliosis-specific classification, but they are higher-capacity systems, and their published results remain dependent on their respective protocols.

A separate 2026 preprint introduced ScoliGait, comprising radiographically annotated gait videos and a subject-independent test cohort. It combines video, clinical text, and a kinematic knowledge map through latent attention pooling [20]. This work is methodologically important because it connects gait labels to radiographic information and explicitly addresses repeated-subject leakage. However, ScoliGait was unavailable for the present experiments. Its results therefore establish a stronger external-validation target, not a directly comparable benchmark.

2.4. Research Gap and Study Positioning

The research gap does not lie in using temporal means, standard deviations, or an SVM, all of which are established techniques. Existing scoliosis-specific gait studies have primarily advanced learned representations and predictive performance. Less attention has been given to auditing what a compact, inspectable pose representation supports when neutral cases are retained, model selection is isolated from testing, pose quality and temporal coverage are disrupted, and feature contributions are evaluated using model-appropriate methods.

As summarized in Table 1, general gait and skeleton models provide architectural context, whereas published scoliosis-specific systems use higher-capacity set-based, attention-guided, or multistream representations. Their reported results are treated as contextual evidence because differences in inputs, labels, partitions, and validation units prevent direct algorithmic ranking. Only methods reproduced using the local sequence partition are included as empirical comparators.

SpineTrace addresses this evidential gap as an auditable engineering baseline rather than a state-of-the-art alternative to these learned systems. Three-class discrimination is treated as the primary task. Conventional classifiers, PAV-SVM, DTW- k NN, and a compact ST-GCN adaptation are tuned

independently using training data. Feature removal, confidence suppression, coordinate perturbation, frame loss, and repeated permutation importance examine predictive dependence and operating boundaries.

The evaluation deliberately retains counter-results: temporal pose dispersion improves performance, but additional silhouette and postural descriptors do not, while modest coordinate perturbation causes severe degradation. SpineTrace is therefore positioned as a transparent sequence-level reference for auditing feature utility, neutral-case behavior, estimator dependence, and robustness—not as a deployment-ready or state-of-the-art classifier.

Table 1. Research positioning and rules governing comparison with previous methods

Method	Input	Task and temporal mechanism	Neutral treatment	Interpretability	Role in this study
GaitSet [15]	Silhouette	Identity recognition; unordered set pooling	Not clinical	Set-level pooling	Architectural context only
GaitGL [16]	Silhouette	Identity recognition; global–local temporal aggregation	Not clinical	Learned spatial regions	Architectural context only
ST-GCN [17]	Skeleton	Action recognition; spatial–temporal graph convolution	Not clinical	Anatomical graph structure	Compact three-class adaptation reproduced
ScoNet-MT [7]	Silhouette	Three-class classification; set-based multitask learning	Explicit class	Learned features	Published Scoliosis1K context
DRF/PAV [8]	Skeleton map and PAV	Three-class classification; prior-guided attention	Explicit class	PAV-guided heatmaps	PAV-SVM reproduced; DRF contextual
TG-MILNet [18]	Video and text	DTW bags, temporal attention, and boundary-aware learning	Boundary-aware	Text and attention	Published context
ScoliGaitX [19]	Silhouette, flow, and GEI difference	Three-class multistream fusion	Explicit class	Gated stream contributions	Published context
ScoliGait [20]	Video, kinematic map, and text	Multimodal latent-attention pooling	Separate benchmark	Structured kinematic map	External-validation context
SpineTrace	ViTPose sequences and silhouettes	Three-class primary task; temporal moments	Explicit class	Named features, removal tests, and permutation importance	Reproduced on the local sequence split

3. PROPOSED SPINETRACE FRAMEWORK

SpineTrace transforms each 300-frame back-view gait sequence into a compact sequence-level representation and evaluates whether that representation can distinguish negative, neutral, and positive scoliosis-screening labels. As illustrated in Figure 1, the framework comprises data auditing, pose normalization, feature construction, training-only model selection, locked test evaluation, and post hoc robustness and attribution analyses. The workflow also establishes an explicit evidence boundary: the released data support sequence-level reproduction, but the absence of participant mappings and raw RGB videos prevents verification of participant independence, transfer between pose estimators, and end-to-end video-processing performance.

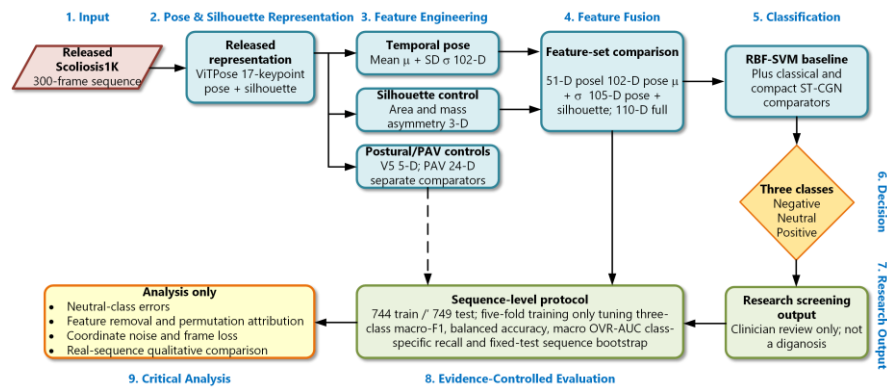


Figure 1. Evidence-controlled SpineTrace workflow. Released ViTPose and silhouette sequences undergo auditing, normalization, feature construction, training-only model selection, locked test evaluation, and robustness and attribution analyses. Unavailable participant mappings and RGB videos limit participant-level and end-to-end claims.

3.1. Data Audit and Sequence Partition

The local Scoliosis1K cache contained all 1,493 released gait sequences. Each sequence comprised 300 frames, with 17 anatomical keypoints represented in the MS-COCO order by horizontal coordinate, vertical coordinate, and estimator confidence. Every pose sequence was accompanied by a 64×64 binary silhouette sequence [8, 21]. The audit found no duplicate sequence identifiers, dropped clips, nonfinite coordinates, or unmatched pose–silhouette pairs. The complete dataset contained 493 positive, 200 neutral, and 800 negative sequences.

The supplied `Scoliosis1K_118.json` file allocated 744 sequences to training and 749 to testing. These counts differ by one sequence in each partition from the 745/748 totals reported in the associated MICCAI papers. The experiments used the supplied file without modifying its membership, and its SHA-256 hash was stored with the experimental records. This decision preserves exact local reproducibility and avoids undocumented reassignment of sequences to reproduce a published total.

Table 2. Audited class distribution in the supplied sequence partition

Partition	Positive	Neutral	Negative	Total
Training	74	74	596	744
Test	419	126	204	749
All sequences	493	200	800	1,493

The training partition has an approximately (1:1:8) positive-to-neutral-to-negative ratio, whereas the test partition contains a substantially larger proportion of positive sequences. This label-distribution shift was retained because it is part of the released benchmark. Overall accuracy was therefore supplemented by balanced accuracy, macro-F1, class-specific metrics, and macro one-versus-rest area under the receiver operating characteristic curve (ROC-AUC).

Although Scoliosis1K was reportedly obtained from 1,050 adolescents, the local release does not map its 1,493 sequence identifiers to individual participants. It was possible to confirm that the training and test sequence lists were disjoint and collectively covered the cache, but not whether different recordings from the same participant occurred in both partitions. Consequently, validation and uncertainty analyses were performed at the sequence level. The terms participant-grouped validation, participant-independent evaluation, and participant-level confidence interval are not used for the present evidence.

3.2. Pose Provenance and Normalization

The released keypoints were generated using ViTPose rather than MediaPipe or BlazePose [8, 12]. Raw RGB videos were unavailable, so the same frames could not be reprocessed with another estimator. The reported results therefore apply specifically to the released ViTPose-derived poses and silhouettes. We retained confidence as the third keypoint channel because it may carry information about visibility and estimation reliability; we subsequently examined its predictive contribution through removal and missing-joint experiments.

For frame t and joint j , the two-dimensional position was defined as

$$\mathbf{p}_{t,j} = [x_{t,j} \quad y_{t,j}]^T. \quad (1)$$

The hip and shoulder midpoints were calculated from their respective bilateral landmarks:

$$\mathbf{h}_t = \frac{\mathbf{p}_{t,LHip} + \mathbf{p}_{t,RHip}}{2}, \quad \mathbf{s}_t = \frac{\mathbf{p}_{t,LShoulder} + \mathbf{p}_{t,RShoulder}}{2}. \quad (2)$$

Torso length was defined as $d_t = \|\mathbf{s}_t - \mathbf{h}_t\|_2$. Coordinates were translated to the hip midpoint and divided by torso length:

$$\tilde{\mathbf{p}}_{t,j} = \frac{\mathbf{p}_{t,j} - \mathbf{h}_t}{\max(d_t, \epsilon)}, \quad \epsilon = 10^{-6}. \quad (3)$$

Confidence $c_{t,j}$ was not rescaled during coordinate normalization. Eq. (3) reduces sensitivity to image position and apparent body scale, but it does not ensure invariance to camera rotation, perspective, occlusion, clothing, or systematic pose-estimation error. No near-zero torso length produced nonfinite normalized coordinates.

3.3. SpineTrace Feature Representations

Figure 2 illustrates how the released pose and silhouette sequences are converted into the candidate SpineTrace representation. The principal pose descriptor summarizes average alignment and temporal dispersion. For joint j and channel $q \in [\tilde{x}, \tilde{y}, c]$, the sequence mean, and population standard deviation were calculated as

$$\mu_{j,q} = \frac{1}{T} \sum_{t=1}^T z_{t,j,q}, \quad (4)$$

$$\sigma_{j,q} = \sqrt{\frac{1}{T} \sum_{t=1}^T (z_{t,j,q} - \mu_{j,q})^2}, \quad T = 300. \quad (5)$$

The means and standard deviations of 17 joints across three channels produced $17 \times 3 \times 2 = 102$ variables. The means describe average pose and confidence, whereas the standard deviations quantify temporal variability. This representation is compact and traceable to named landmarks, but it is invariant to frame order. It therefore does not encode the sequence or duration of individual gait phases.

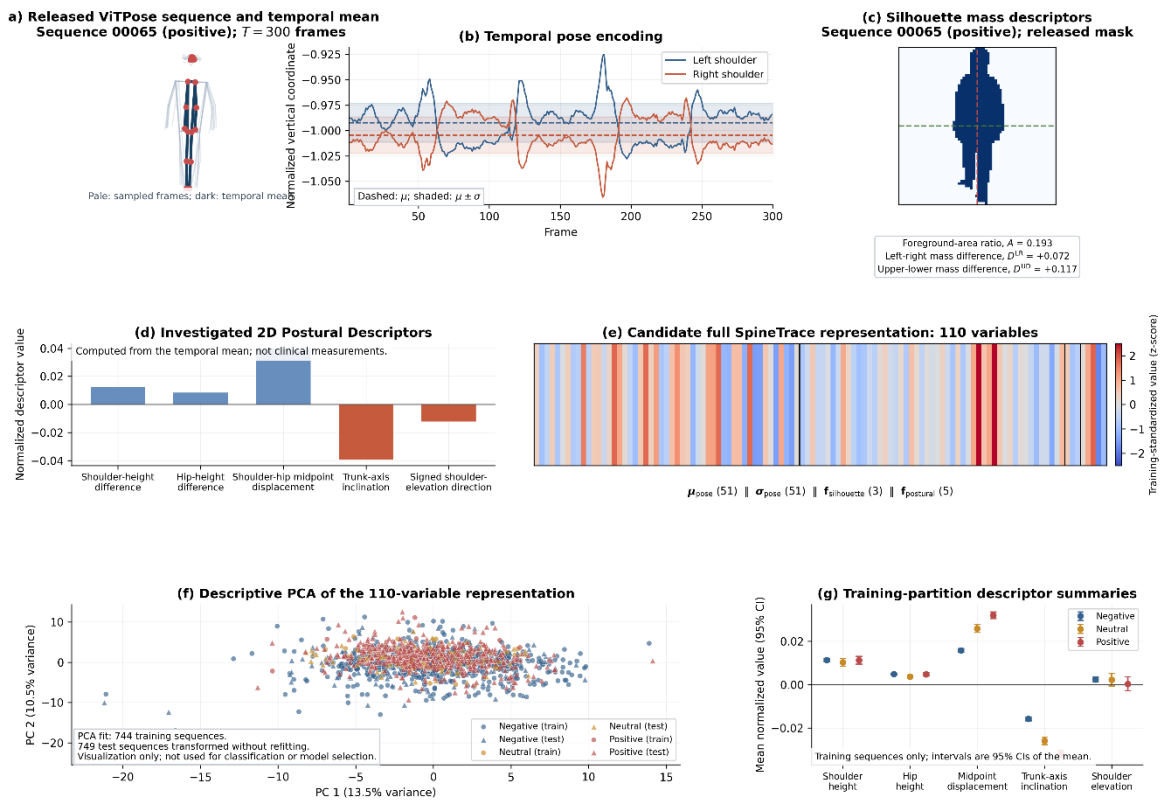


Figure 2. Construction of the SpineTrace representation from Scoliosis1K sequences. The figure shows the temporal mean pose, keypoint variability, silhouette mass descriptors, five 2D postural descriptors, the candidate 110-dimensional feature vector, and descriptive three-class comparisons. These external postural descriptors are not radiographic measurements.

Three coarse silhouette descriptors were reconstructed. For binary mask $M_t(u, v)$, the foreground-area ratio was

$$A_t = \frac{1}{HW} \sum_{u=1}^H \sum_{v=1}^W M_t(u, v). \quad (6)$$

Left-right and upper-lower mass differences were calculated as

$$D_t^{LR} = \frac{m_t^L - m_t^R}{m_t^L + m_t^R + \varepsilon}, \quad D_t^{UD} = \frac{m_t^U - m_t^D}{m_t^U + m_t^D + \varepsilon}. \quad (7)$$

Each descriptor was averaged across the 300 frames. These variables characterize projected foreground area and mass distribution, but they remain sensitive to segmentation quality and crop position. They are not interpreted as anatomical measurements.

Five postural descriptors were calculated from the temporal mean skeleton: absolute shoulder-height difference, absolute hip-height difference, signed shoulder-hip midpoint displacement, trunk-axis inclination, and signed shoulder-elevation direction. Representative definitions are:

$$\Delta_{\text{shoulder}} = \frac{|\bar{y}_{L\text{Shoulder}} - \bar{y}_{R\text{Shoulder}}|}{\bar{d}}, \quad \Delta_{\text{hip}} = \frac{|\bar{y}_{L\text{Hip}} - \bar{y}_{R\text{Hip}}|}{\bar{d}}, \quad (8)$$

$$I_{\text{trunk}} = \frac{\bar{x}_{\text{shoulder}} - \bar{x}_{\text{hip}}}{|\bar{y}_{\text{shoulder}} - \bar{y}_{\text{hip}}| + \varepsilon}. \quad (9)$$

These features describe external two-dimensional posture and are not Cobb-angle proxies, measurements of vertebral rotation, or diagnostic clinical variables.

The four feature blocks were concatenated to form the candidate full SpineTrace representation:

$$\mathbf{f}_{\text{full}} = \mu_{\text{pose}} \parallel \sigma_{\text{pose}} \parallel \mathbf{f}_{\text{silhouette}} \parallel \mathbf{f}_{\text{postural}} \in R^{110}. \quad (10)$$

The term candidate is used deliberately. Concatenation does not establish that every block provides complementary predictive information. The 102-dimensional pose representation, 105-dimensional pose-and-silhouette representation, and candidate full 110-dimensional representation were therefore evaluated and tuned separately. Their relative utility was determined empirically rather than assumed from feature dimensionality or clinical plausibility.

A 24-dimensional Postural Asymmetry Vector (PAV) was implemented as the closest clinically structured comparator. Eight bilateral pairs—eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles—each contributed absolute vertical deviation, bilateral-midpoint deviation from the hip axis, and bilateral segment angle. Values outside 1.5 interquartile ranges were removed independently over time, valid observations were averaged, and scaling was fitted within each training fold. PAV-SVM evaluates whether PAV alone provides sufficient discrimination; it is not presented as a reproduction of the complete DRF architecture, in which PAV guides attention over learned skeleton maps.

Table 3. Evaluated representations and analytical purpose

Representation	Dimension	Purpose and principal limitation
Pose mean	51	Average pose and confidence; no dispersion or phase order
Pose mean and SD	102	Average pose and temporal dispersion; no phase order
Silhouette	3	Coarse shape and framing control
Postural	5	Explicit two-dimensional postural descriptors
Pose and silhouette	105	Reconstruction of the earlier combined representation
Full SpineTrace	110	Candidate pose, silhouette, and postural representation
PAV	24	Bilateral postural-asymmetry comparator
DTW trajectories	25×4	Ordered shoulder and hip signals

3.4. Endpoints and Training-only Model Selection

Three-class prediction was the primary endpoint, with negative, neutral, and positive encoded as 0, 1, and 2. Macro-F1 was the principal selection metric because it weights all three classes equally and highlights failures in the neutral category. Accuracy, balanced accuracy, macro one-versus-rest ROC-AUC, confusion matrices, and class-specific precision, recall, and F1 were also reported.

Two secondary endpoints were evaluated. The pre-referral formulation grouped positive and neutral sequences as requiring professional review and compared them with negative sequences. The legacy formulation excluded neutral sequences and compared positive with negative. For binary tasks, ranking metrics were calculated from the untransformed decision function. An F1-maximizing threshold and an exploratory threshold satisfying at least 0.90 training sensitivity were selected exclusively from training out-of-fold scores. SVM decision scores are not described as calibrated probabilities.

All conventional classifiers were implemented as scikit-learn pipelines so that scaling and model fitting occurred independently within each of five stratified training folds. The RBF-SVM search included $C \in [0.01, 0.1, 1, 10, 100, 1000]$ and $\gamma \in [\text{scale}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$, with balanced class weights [22]. Logistic regression, linear SVM, random forest, and histogram gradient boosting used prespecified grids stored with the code. Every classifier, representation, and feature-removal model underwent independent tuning. Hyperparameters or thresholds were not transferred between feature sets.

After selection, each pipeline was refitted using all 744 training sequences and evaluated once on the fixed 749-sequence test partition. Test uncertainty was summarized using 2,000 class-stratified sequence bootstraps. These intervals quantify sensitivity to resampling the available test sequences; they are not

participant-clustered confidence intervals. Participant-level hypothesis tests were not performed because participant independence could not be verified.

3.5. Classical and deep temporal comparators

DTW- k NN provided a classical comparator that retained temporal order. Shoulder-height difference, hip-height difference, shoulder-midpoint horizontal position, and hip-midpoint horizontal position were uniformly resampled to 25 points and standardized from training data. A training-only carve-out selected the warping window from [2,4] and k from [1,3,5,7]. Templates were balanced by class and drawn exclusively from the fitting partition. An HMM was considered but not reported because a validated implementation with adequate stability and convergence audit was unavailable.

A compact ST-GCN adaptation provided the deep temporal comparator [17]. The graph comprised 17 anatomical joints, with spatial edges following skeletal connectivity and temporal edges joining corresponding joints across frames. Normalized x , normalized y , and confidence formed the three input channels. The network used channel widths of 32, 64, and 128; class-weighted cross-entropy; AdamW optimization; cosine scheduling; validation macro-F1 early stopping; and three prespecified seeds. The resulting model contained 215,945 trainable parameters. It is described as an adaptation to the three-class clinical task rather than a reproduction of the original action-recognition benchmark.

3.6. Robustness and attribution

After the full 110-dimensional RBF-SVM was locked, normalized test coordinates received Gaussian noise with standard deviations of 0.5%, 1%, 2%, and 5% of torso length. Frame-loss experiments removed 10%, 25%, or 50% of frames either randomly or as a centered contiguous block. Further tests applied temporal smoothing or suppressed estimator confidence, facial and ear landmarks, or shoulder and hip landmarks. Models were not retuned for corrupted test data.

Because an RBF-SVM has no directly interpretable linear coefficients, feature influence was examined using repeated permutation importance. Each feature or feature group was permuted across test sequences 30 times, and importance was defined as the resulting decrease in macro-F1. Confirmatory removal models repeated the complete training-only selection procedure. These analyses quantify predictive dependence rather than causal anatomical importance because correlated variables may share or transfer information.

3.7. Reproducibility and computational scope

Experiments used Python 3.10, NumPy 2.2.6, SciPy 1.15.3, scikit-learn 1.7.2, and PyTorch 2.5.1 with CUDA 12.1. Conventional models ran on the CPU, whereas ST-GCN training used an NVIDIA GeForce RTX 4090. Seed 20260825 controlled the conventional experiments, and ST-GCN used seeds 0, 1, and 2. We preserved configurations, fitted models, feature names, sequence-level predictions, figures, manifests, and source hashes so every aggregate result could be independently recomputed. The source code and reproduction instructions are publicly available at <https://github.com/tsgunawan/SpineTrace>.

Runtime measurements used 20 warm-up passes followed by 200 timed repetitions. They cover feature construction from cached normalized poses and silhouettes, feature scaling, and SVM scoring. Video acquisition, decoding, ViTPose inference, and silhouette extraction were excluded because the raw videos were unavailable. Reported latency is therefore descriptor-plus-classifier latency, not end-to-end screening time.

4. RESULTS

The evaluation considers predictive performance together with class imbalance, model selection, feature dependence, controlled perturbations, and alternative task definitions. This broader analysis is central to the study's contribution: rather than presenting one favorable classification result, it establishes what the compact SpineTrace representation can reproduce, which components contribute useful information, and where its apparent performance becomes unreliable.

4.1. Data Quality and Partition Shift

The data audit identified no missing silhouettes, loading failures, nonfinite pose coordinates, duplicate sequence identifiers, or unmatched pose-silhouette pairs. ViTPose confidence values ranged from 0.014 to 1.064. Values slightly greater than 1 were retained because they occurred in the released tensors and confidence clipping had not been prespecified. Mean confidence across joints and frames was similar for negative, neutral, and positive sequences at 0.901, 0.902, and 0.900, respectively. The proportions of joint observations below 0.50 were also low at 0.97%, 0.73%, and 1.12%. These global similarities indicate no obvious class-wide

difference in pose confidence, although they do not exclude joint-specific or acquisition-related associations with the labels.

The supplied partition presents a substantial label-distribution shift. The training set contains only 74 positive and 74 neutral sequences but 596 negative sequences, whereas the test set contains 419 positive, 126 neutral, and 204 negative sequences. A classifier biased toward the training majority may therefore perform well during internal validation but fail when the test distribution becomes predominantly positive. Conversely, overall test accuracy can conceal poor neutral performance because each class contributes differently across the two partitions. Consequently, balanced accuracy, macro-F1, macro one-versus-rest area under the receiver operating characteristic curve (ROC-AUC), and class-specific recall are emphasized alongside accuracy.

This partition also limits the stability of model selection. In five-fold stratified validation, each fitting fold contains approximately 59 positive and 59 neutral sequences. A relatively small number of borderline cases can therefore alter the selected hyperparameters or decision boundary. The findings should be interpreted as a reproducible evaluation of the supplied sequence partition rather than a stable estimate of participant-level or population-level performance.

4.2. Primary Three-class Performance

Table 4 reports the locked-test results for the primary positive–neutral–negative task. The pose mean-and-standard-deviation RBF-SVM achieved the highest observed macro-F1 among the models reproduced on the local 744/749 sequence partition. It attained 0.737 accuracy, 0.740 balanced accuracy, 0.702 macro-F1, and 0.890 macro one-versus-rest ROC-AUC. This qualification limits the comparison to models evaluated using the same local data and protocol; it does not imply superiority over published systems evaluated using different implementations, modalities, or partitions.

Table 4. Locked-test results for the primary three-class task.

Feature or input	Model	Accuracy	Balanced accuracy	Macro-F1	Macro OVR-AUC
Pose mean, 51-D	RBF-SVM	0.702	0.687	0.658	0.869
Pose mean and SD, 102-D	RBF-SVM	0.737	0.740	0.702	0.890
Silhouette, 3-D	RBF-SVM	0.368	0.377	0.352	0.547
Postural, 5-D	RBF-SVM	0.602	0.550	0.540	0.750
Pose and silhouette, 105-D	RBF-SVM	0.721	0.732	0.691	0.893
Full SpineTrace, 110-D	RBF-SVM	0.680	0.693	0.656	0.899
PAV, 24-D	RBF-SVM	0.591	0.575	0.545	0.760
Four 25-step trajectories	DTW-kkNN	0.502	0.506	0.473	0.657
17-joint sequence	ST-GCN adaptation	0.710±0.068	0.660±0.055	0.645±0.055	0.851±0.072

Bold values indicate the highest observed result within the locally reproduced comparisons. ST-GCN values are means ± optimization-seed standard deviations, not independent-cohort confidence intervals.

Adding temporal standard deviations to the pose means increased macro-F1 from 0.658 to 0.702 and balanced accuracy from 0.687 to 0.740. Neutral recall increased by 0.151. These changes support the premise that temporal dispersion contains useful screening information beyond average posture. They do not demonstrate that ordered gait dynamics were learned because the means and standard deviations remain unchanged when the frames are permuted. The improvement should therefore be attributed to distributional variability rather than gait-phase modeling.

The DTW comparator retained temporal order but achieved a macro-F1 of 0.473. This result does not establish that gait-phase order is uninformative because the implementation was restricted to four shoulder and hip trajectories resampled to 25 time points. It shows only that this compact ordered representation did not outperform the broader pose moments under the present protocol. The compact ST-GCN adaptation achieved (0.645±0.055) macro-F1 across three optimization seeds. Its variability indicates that a deeper temporal model was not automatically more stable or accurate with the limited number of positive and neutral training sequences.

The principal counter-result is that expanding the handcrafted representation did not improve the primary endpoint. Adding the three silhouette descriptors to the 102 pose variables reduced macro-F1 slightly to 0.691. Adding the five postural descriptors produced the candidate full 110-dimensional representation but reduced macro-F1 further to 0.656. Silhouettes alone provided little discrimination, reaching 0.352 macro-F1 and 0.547 ROC-AUC. The five postural descriptors alone attained 0.540 macro-F1, while PAV-SVM attained 0.545.

These findings demonstrate that semantic plausibility and additional dimensionality do not guarantee complementary predictive information. They also justify referring to the 110-dimensional vector as the candidate full representation rather than the definitive or optimized SpineTrace model. For the primary three-class endpoint, the more compact 102-dimensional pose representation was more effective.

The PAV-SVM result does not contradict DRF's higher published performance. PAV-SVM evaluates the handcrafted vector as a standalone classifier input, whereas DRF uses PAV to guide attention over learned

skeleton maps. The comparison therefore indicates that PAV alone was insufficient under the local protocol, not that clinically guided asymmetry features lack value within higher-capacity learned systems.

The candidate full representation achieved the highest observed macro ROC-AUC among the reproduced RBF-SVM feature configurations (0.899), despite its lower macro-F1. This divergence indicates that ranking performance and discrete classification should not be interpreted interchangeably. The additional features may preserve some class-ranking information while shifting class decision boundaries unfavorably. Reporting only ROC-AUC would consequently conceal important classification errors.

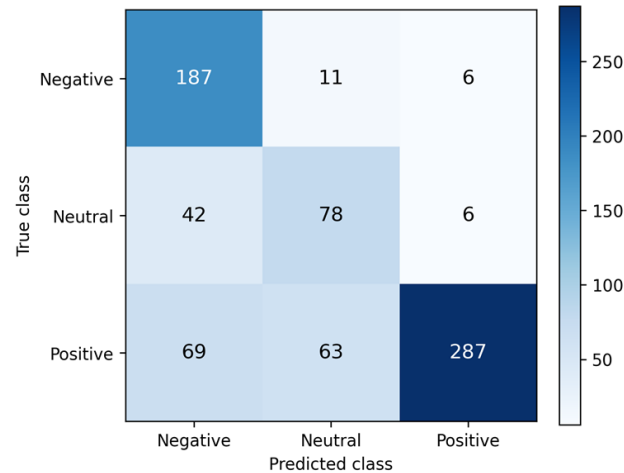


Figure 3. Confusion matrix for the selected pose mean-and-standard-deviation RBF-SVM on the three-class test partition.

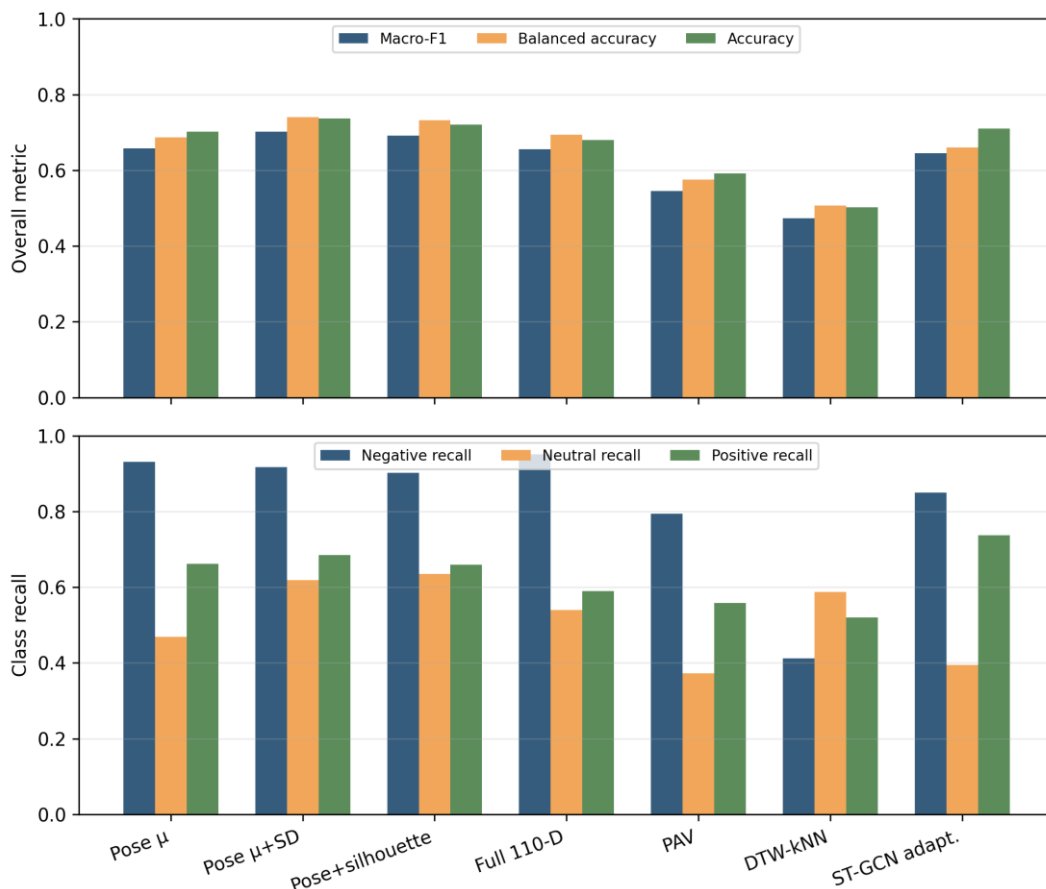


Figure 4. Locked-test aggregate performance and class-specific recall for the reproduced models.

The class-stratified sequence bootstrap placed macro-F1 for the pose mean-and-standard-deviation model between 0.668 and 0.735 and macro ROC-AUC between 0.870 and 0.908. For the candidate full representation, the corresponding intervals were 0.621–0.690 and 0.880–0.917. Their overlap cautions against treating the observed difference as a definitive population ranking. These intervals quantify sensitivity to resampling the fixed test sequences; they do not incorporate participant clustering, alternative data partitions, or model-selection variability.

Classifier selection also affected the findings. With the candidate full 110-dimensional representation, logistic regression attained 0.693 macro-F1 and 0.738 accuracy, exceeding the RBF-SVM's 0.656 macro-F1. Linear SVM reached 0.677 macro-F1, while random forest and histogram gradient boosting reached 0.593 and 0.595. The assumption that RBF-SVM is uniformly preferable therefore did not survive independently tuned three-class evaluation. Figure 3 presents the confusion matrix for the selected pose mean-and-standard-deviation RBF-SVM. Figure 4 compares aggregate performance and class-specific recall across the locally reproduced models.

4.3. Neutral Cases and Class-specific Behavior

Neutral remained the most difficult class. As shown in Table 5, the selected pose model correctly classified 78 of 126 neutral sequences, resulting in 0.619 recall and 0.561 F1. It assigned 42 neutral sequences to negative and six to positive. Neutral predictions also included 63 positive and 11 negative sequences, explaining why neutral F1 remained lower than its recall.

Table 5. Class-specific recall and F1 on the primary test partition.

Model	Negative recall/F1	Neutral recall/F1	Positive recall/F1
Pose mean and SD RBF-SVM	0.917/0.745	0.619/0.561	0.685/0.799
Full 110-D RBF-SVM	0.951/0.661	0.540/0.574	0.589/0.733
PAV-RBF-SVM	0.794/0.599	0.373/0.373	0.558/0.664

The full model achieved higher negative recall (0.951), but this apparent advantage came with lower positive and neutral recall. The model therefore shifted decisions toward negative rather than improving all three classes. PAV-RBF performed worse across the classes, with neutral recall falling to 0.373. These results confirm that borderline cases cannot be treated as a minor extension of binary classification. Retaining neutral changes the decision boundary, the error profile, and the clinical interpretation of performance.

Misclassification costs are asymmetric. Assigning a positive sequence to negative may delay assessment; assigning a negative sequence to positive may increase anxiety and referral burden; and assigning a neutral sequence to either extreme may erase useful uncertainty. Macro-F1 gives the classes equal importance because no clinically validated error-cost matrix was available. Future deployment studies should prespecify the action associated with each output and the acceptable referral burden before selecting thresholds.

4.4. Feature Removal and Attribution

Feature removal provides a more direct test of contribution than semantic plausibility. Every configuration in Table 6 was independently retuned within the training folds.

Table 6. Independently retuned feature-removal results.

Configuration	Dimension	Macro-F1	Balanced accuracy	Accuracy
Full reference	110	0.656	0.693	0.680
Without confidence	76	0.618	0.665	0.653
Without face and ears	80	0.635	0.678	0.666
Without silhouette	107	0.648	0.684	0.676
Without postural descriptors	105	0.691	0.732	0.721

Removing the five postural descriptors increased macro-F1 from 0.656 to 0.691, confirming that their clinical interpretability did not produce complementary predictive value. Removing silhouettes changed macro-F1 only to 0.648, consistent with their weak standalone performance. In contrast, removing confidence reduced macro-F1 to 0.618, while removing face and ear features reduced it to 0.635.

Reliance on confidence and posterior-view facial landmarks is an important warning. These variables may aid classification, but their usefulness cannot automatically be attributed to scoliosis-related anatomy. They may capture head visibility, crop alignment, occlusion, motion blur, or estimator-specific behavior. Their contribution may improve performance on this release while reducing transferability to different cameras or pose estimators. Grouped permutation importance produced the same broad pattern. As shown in Table 7, coordinates, pose means, temporal standard deviations, confidence, and face and ear features produced the largest decreases in macro-F1.

Table 7. Grouped permutation importance for the locked full RBF-SVM

Permuted group	Mean macro-F1 decrease	SD
Coordinates	0.277	0.015
Pose means	0.256	0.014
Pose temporal SDs	0.195	0.013
Confidence	0.174	0.014
Face and ears	0.158	0.015
Postural descriptors	0.048	0.009
Silhouette	0.018	0.007

These groups overlap and should not be summed as independent contributions. For example, the coordinate group includes coordinate means and standard deviations, while the pose-mean group includes confidence means. The strongest individual variables included ankle and knee positions, signed shoulder–hip displacement, trunk-axis inclination, and selected confidence channels. This mixture combines plausible gait geometry with acquisition-sensitive information. Permutation importance therefore identifies predictive dependence rather than anatomical causality. Figure 5 presents grouped and individual importance across 30 repeated permutations.

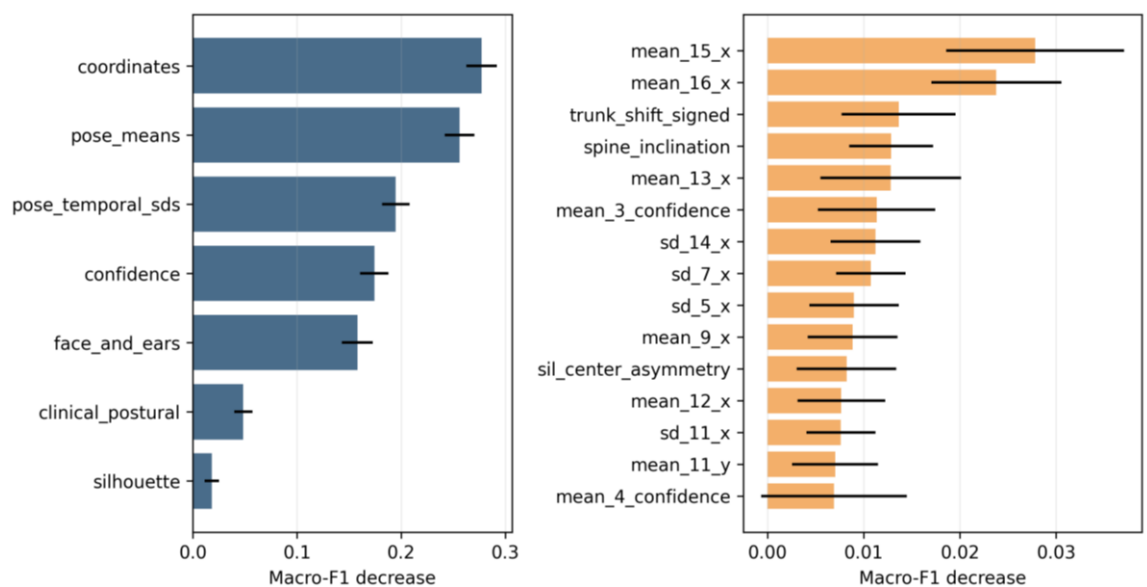


Figure 5. Repeated permutation importance for the locked full RBF-SVM, reported as the reduction in macro-F1.

4.5. Controlled Robustness

Table 8 shows that the locked full model was stable under very small coordinate perturbations but failed sharply beyond a narrow range.

Table 8. Robustness of the locked full model without corruption-specific retuning

Condition	Macro-F1	Balanced accuracy	Accuracy
No coordinate jitter	0.656	0.693	0.680
Jitter: 0.5% torso length	0.664	0.701	0.692
Jitter: 1% torso length	0.646	0.684	0.678
Jitter: 2% torso length	0.287	0.407	0.383
Jitter: 5% torso length	0.143	0.333	0.272
Random frame removal: 10%	0.652	0.690	0.680
Random frame removal: 25%	0.650	0.685	0.681
Random frame removal: 50%	0.643	0.677	0.676
Contiguous frame removal: 10%	0.622	0.656	0.668
Contiguous frame removal: 25%	0.516	0.570	0.539
Contiguous frame removal: 50%	0.368	0.458	0.426
Savitzky–Golay smoothing	0.664	0.696	0.690
Face and ear masking	0.143	0.333	0.272
Shoulder and hip masking	0.143	0.333	0.272

Macro-F1 remained approximately stable at 0.5% and 1% jitter but collapsed to 0.287 at 2% and 0.143 at 5%. Because the representation includes temporal standard deviations, independently sampled coordinate noise directly inflates the variables intended to represent sway. A pose estimator with moderate frame-to-frame localization error may therefore generate the same dispersion signal that the classifier interprets as gait variability. This is one of the study's most important negative findings: temporal dispersion improves unperturbed classification while creating direct sensitivity to pose instability.

Random frame removal was less damaging. Removing up to 50% of randomly selected frames produced only a modest decline because the remaining frames were distributed across the sequence and still approximated its mean and standard deviation. Contiguous removal was substantially more harmful, reducing macro-F1 to 0.622, 0.516, and 0.368 at 10%, 25%, and 50% removal. This contrast suggests that coverage across the gait sequence matters more than the absolute number of retained frames.

Savitzky–Golay smoothing increased macro-F1 modestly to 0.664. Because smoothing was not selected prospectively, interpret this result as a sensitivity analysis rather than an optimized improvement. Masking face and ear joints or shoulder and hip joints reduced performance to near-single-class behavior. These severe tests confirm that the locked representation cannot tolerate complete loss of important landmark groups.

Figure 6 summarizes the effects of coordinate jitter and random or contiguous frame loss. These perturbations characterize controlled failure mechanisms rather than field robustness. They do not reproduce changes in camera angle, illumination, clothing, segmentation quality, or pose-estimator architecture.

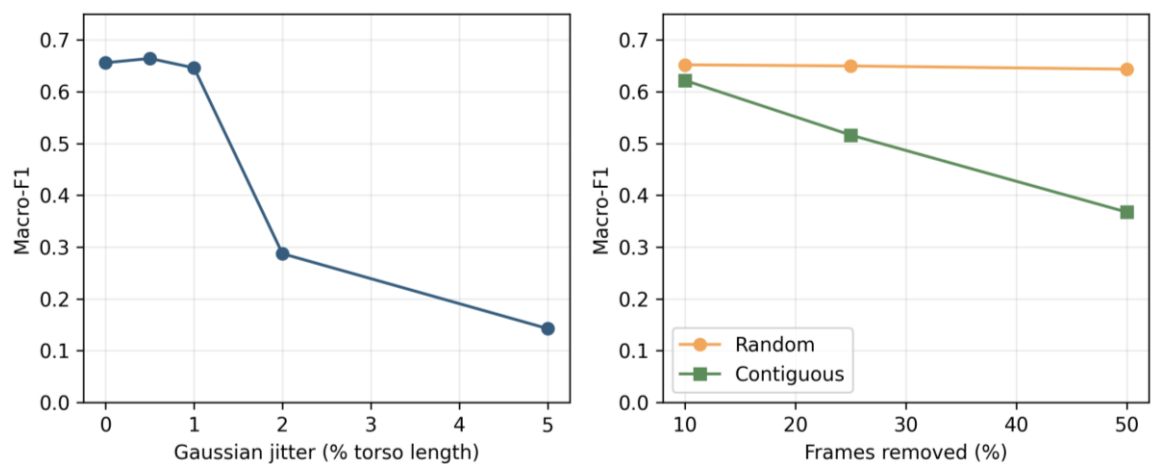


Figure 6. Robustness of the locked full model under coordinate jitter and random or contiguous frame removal.

4.6. Qualitative Model Disagreement

Figure 7 shows one negative, one neutral, and one positive test sequence using the released silhouettes, normalized pose traces, and model-specific class support. The cases were selected to expose disagreement while requiring the pose mean-and-standard-deviation model to identify the displayed reference class correctly. They illustrate computations and failure modes rather than their prevalence.

The neutral sequence receives predominant neutral support from the pose-statistics and PAV models, whereas the full descriptor shifts toward negative and ST-GCN supports positive. For the positive sequence, the pose-statistics model and ST-GCN support positive, while the full descriptor favors negative and PAV favors neutral. These examples demonstrate that interpretable features do not guarantee stable decisions and that agreement among models derived from the same estimated poses does not constitute independent clinical confirmation. The bars contain softmax-normalized class scores rather than calibrated probabilities. Interpret their magnitudes only within each model and sequence; do not compare them quantitatively across the RBF-SVM, PAV, and ST-GCN models.

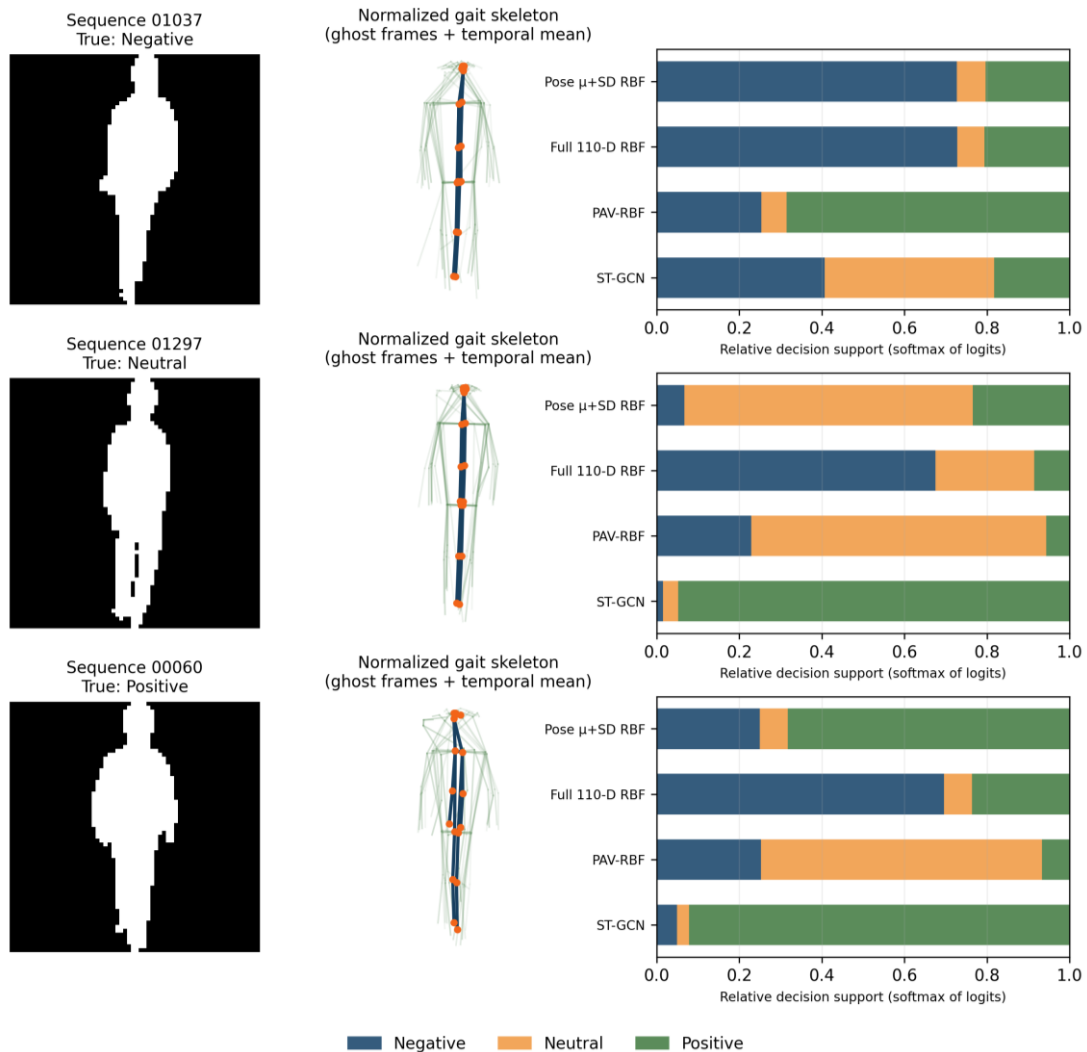


Figure 7. Illustrative test sequences showing silhouettes, normalized poses, and relative class support from four reproduced models.

4.7. Secondary Endpoints, Contextual Evidence, and Runtime

The secondary binary endpoints demonstrate how task definition changes apparent model performance. As shown in Table 9, the pre-referral formulation combined positive and neutral sequences into a single “flag for professional review” category. Under this formulation, the full 110-dimensional SpineTrace representation attained 0.810 accuracy, 0.845 balanced accuracy, 0.790 macro-F1, and 0.939 ROC-AUC. PAV-SVM reached 0.727 macro-F1 and 0.824 ROC-AUC. Grouping positive and neutral may better align with a triage pathway because borderline sequences are not forced into diagnostic-like categories. However, this endpoint does not distinguish confirmed positive from uncertain neutral cases and must not be reported as three-class performance.

Table 9. Secondary binary endpoints at training-derived thresholds

Task	Representation	Accuracy	Balanced accuracy	Macro-F1	ROC-AUC
Positive or neutral vs. negative	Full 110-D	0.810	0.845	0.790	0.939
Positive or neutral vs. negative	PAV, 24-D	0.760	0.760	0.727	0.824
Positive vs. negative	Full 110-D	0.854	0.880	0.846	0.960
Positive vs. negative	PAV, 24-D	0.666	0.724	0.664	0.858

The legacy positive-versus-negative task produced the strongest numerical results, with the full SpineTrace representation attaining 0.854 accuracy, 0.846 macro-F1, and 0.960 ROC-AUC. These values do not supersede the primary three-class findings. Excluding neutral sequences removes the most ambiguous screening boundary and produces a simpler decision problem with a different clinical meaning. The present ROC-AUC is also lower than the previously reported value of 0.986 because the revised evaluation uses audited normalization, the supplied sequence partition, training-only model selection, untransformed decision scores,

and an independently selected operating threshold. The difference illustrates why binary performance should not be generalized to a screening population that contains borderline cases.

Published Scoliosis1K systems report higher predictive performance than SpineTrace. ScoNet-MT uses learned set-based silhouette representations, DRF combines skeleton maps with PAV-guided attention, TG-MILNet models temporal alignment and borderline cases, and ScoliGaitX integrates multiple appearance and motion streams [7, 8, 18, 19]. Their reported values are not directly comparable with the present results because their input modalities, implementations, partition counts, and model-selection procedures differ. Nevertheless, the performance gap indicates that SpineTrace should not be presented as a state-of-the-art classifier. Its compact temporal moments discard gait-phase order and short-duration events that higher-capacity models may retain.

SpineTrace instead offers a different contribution: a transparent and reproducible baseline through which feature composition, neutral-case behavior, estimator dependence, and controlled failure modes can be examined directly. The results show that temporal pose dispersion improves classification over average posture alone, but also creates sensitivity to coordinate instability. They further demonstrate that adding semantically appealing postural descriptors does not necessarily improve predictive performance. These findings provide evidence about the limits of compact gait summaries that would be obscured by reporting only the highest binary score.

The computational measurements establish low overhead after pose and silhouette extraction. Feature construction from cached inputs required a median of 0.063 ms per sequence, while feature scaling and SVM scoring required 0.216 ms. The fitted full-model file occupied 354,709 bytes. The compact ST-GCN contained 215,945 trainable parameters and required approximately 0.74 ms for batch-one inference on an NVIDIA GeForce RTX 4090. These measurements exclude video acquisition, decoding, ViTPose inference, and silhouette extraction because the raw RGB videos were unavailable. They therefore characterize descriptor and classifier efficiency, not end-to-end screening latency or CPU-based video deployment.

Therefore, the secondary analyses reinforce rather than replace the primary three-class result. SpineTrace provides transparency and low downstream computational cost rather than maximum predictive performance. Its principal value lies in identifying which compact gait summaries remain informative, which feature additions fail to provide complementary evidence, and how rapidly performance deteriorates when pose quality is disrupted.

5. DISCUSSION

This discussion establishes what SpineTrace contributes as an auditable gait-screening baseline, where its evidence remains bounded, and what validation is required before clinical evaluation.

5.1. Contribution of the Compact Representation

The principal contribution of SpineTrace is an auditable sequence-level baseline, not a new algorithm or state-of-the-art classifier. On the local 744/749 sequence partition, the pose mean-and-standard-deviation RBF-SVM achieved the highest observed macro-F1 among the reproduced models. Its 0.702 macro-F1 exceeded the compact ST-GCN adaptation, PAV-SVM, and DTW- k NN under independently tuned comparisons. This result provides a meaningful reference for subsequent research: higher-capacity systems should be compared not only with untuned classifiers but also with a transparent representation optimized under the same training-only protocol.

The improvement from 0.658 to 0.702 macro-F1 after adding temporal standard deviations shows that within-sequence dispersion contains useful information beyond average posture. The interpretation must nevertheless remain precise. Standard deviations are invariant to frame order, so the improvement demonstrates the value of temporal aggregation rather than gait-phase modeling. The lower and more variable performance of the compact ST-GCN does not establish that ordered modeling is unhelpful; it indicates that greater architectural capacity did not overcome the limited positive and neutral training sample in this implementation.

The results also favor representational restraint. The candidate full 110-dimensional SpineTrace representation did not outperform its 102-dimensional pose component. Adding silhouette and postural descriptors reduced macro-F1, despite their intuitive relevance to body symmetry. The postural variables are derived from estimated two-dimensional landmarks, contain partly redundant signed and absolute measurements, and may become unstable under unfavorable geometry. Their semantic appeal therefore did not support complementary predictive value. Feature removal confirmed that the more compact pose representation was preferable for the primary endpoint. This counter-result strengthens the paper by showing that each feature block was tested rather than justified retrospectively.

5.2. Neutral Cases and Task Definition

Retaining neutral sequences materially changed the problem. The selected pose model correctly identified 61.9% of neutral sequences, but neutral F1 remained 0.561 because many positive sequences were assigned to the middle class. Neutral therefore represents a difficult decision region rather than a minor extension of negative classification. Removing these sequences produced substantially stronger positive-versus-negative results, including 0.960 ROC-AUC, but also removed the ambiguity most relevant to pre-referral screening.

The pre-referral endpoint, which groups positive and neutral sequences as requiring professional review, may be more compatible with a triage pathway. However, its clinical value depends on referral capacity, the cost of missed cases, and the relationship between the released labels and radiographic or specialist assessment. Macro-F1 assigns equal statistical importance to the three categories; it is not a clinical utility function. A future study must prespecify the action associated with each output, the acceptable referral rate, and the relative cost of false-negative and false-positive decisions before selecting an operating threshold.

These findings clarify SpineTrace's intended role. Its outputs should be interpreted as research screening signals for subsequent professional assessment, not as diagnostic categories. A positive or neutral flag could prompt review, whereas a negative output should not override symptoms, physical examination, or clinical judgment. The model does not estimate Cobb angle, determine curve type, select treatment, or provide autonomous reassurance.

5.3. Pose Quality and Interpretation of Model Behavior

Estimator provenance is central to interpreting the results. All reported evidence is based on released ViTPose coordinates. Although a lightweight MediaPipe or BlazePose implementation may be attractive for future use, the current data do not establish that its landmark distribution, confidence scale, or temporal stability would match the training representation. Cross-estimator transfer should therefore be treated as a new validation problem rather than an implementation substitution.

The robustness experiments identify a specific mechanism underlying this concern. Coordinate jitter of 1% of torso length caused little change, but performance collapsed at 2%. Because temporal standard deviations are included as predictors, frame-to-frame localization noise can inflate the same variables intended to represent gait sway. Temporal dispersion is therefore both informative and vulnerable: it improves unperturbed classification but can confuse estimator instability with human movement. The greater effect of contiguous frame removal further suggests that coverage across the walking sequence matters more than the total number of retained frames.

Attribution results reinforce this interpretation. Pose coordinates, temporal standard deviations, confidence, and face and ear variables contributed materially to prediction, whereas silhouette and postural blocks contributed little. Some important variables, such as ankle and knee positions, are consistent with gait geometry. Confidence and posterior-view facial landmarks, however, may reflect visibility, crop alignment, occlusion, or estimator behavior. Repeated permutation importance and independently retuned removal models provide complementary evidence of predictive dependence, but neither establishes a biological mechanism. The model remains interpretable in the computational sense, in which its inputs and dependencies can be traced, without implying that every influential feature has clinical meaning.

5.4. Validity and Next-stage Evaluation

The unavailable sequence-to-participant mapping is the principal internal-validity constraint. Unique sequence identifiers establish file-level separation, but they cannot determine whether multiple recordings from one adolescent occur across training and testing. The reported bootstrap intervals therefore quantify uncertainty across test sequences rather than independent participants. This limitation does not negate the reproducibility of the local evaluation, but it defines its evidential unit and prevents participant-level generalization claims.

External validity also remains unestablished. Recent work, including ScolioGait [20], demonstrates that radiographically labeled and subject-independent gait evaluation is feasible, but compatible external data were unavailable for locked testing of SpineTrace. Controlled coordinate and frame-loss experiments characterize failure mechanisms; they do not reproduce variation across schools, cameras, clothing, body habitus, demographic groups, or clinical workflows. Likewise, the reported computational measurements cover cached feature construction and classification, not video acquisition, pose estimation, silhouette extraction, or complete end-to-end latency.

The next validation study should retain the present model as a locked baseline and obtain verified participant identifiers for participant-grouped nested evaluation. A prospectively collected second-site cohort should preserve borderline cases and use radiographic or specialist reference labels. Raw videos would enable paired ViTPose–MediaPipe analysis, camera and crop perturbations, full-pipeline latency measurement, and

review of model failures. Calibration, decision-curve analysis, subgroup performance, and expected referral workload should then be evaluated against prespecified operating requirements.

Under these boundaries, SpineTrace offers a useful foundation rather than a completed clinical system. Its value lies in establishing a reproducible compact baseline, demonstrating the contribution and vulnerability of temporal pose dispersion, revealing that additional handcrafted descriptors can be counterproductive, and defining the evidence required for participant-independent and cross-site validation.

6. CONCLUSION

This study establishes SpineTrace as a reproducible and auditable sequence-level baseline for video-derived scoliosis screening rather than a novel or state-of-the-art classification algorithm. Using the supplied 744/749 Scoliosis1K sequence partition and training-only model selection, the 102-dimensional pose mean-and-standard-deviation RBF-SVM achieved the highest observed macro-F1 among the locally reproduced models, with 0.737 accuracy, 0.740 balanced accuracy, 0.702 macro-F1, 0.890 macro one-versus-rest ROC-AUC, and 0.619 neutral recall. These findings demonstrate that temporal pose dispersion provides useful information beyond average posture, while also clarifying that statistical moments do not model gait-phase order. The critical analyses further show that additional dimensionality and semantic plausibility do not guarantee complementary predictive value: adding silhouette and postural descriptors reduced primary-task performance, whereas removing the five postural variables improved macro-F1. PAV-SVM, DTW- $\lambda(k)$ NN, and the compact ST-GCN adaptation did not exceed the pose-statistics baseline under the same local evaluation conditions. Keypoint confidence contributed materially to classification but may also encode acquisition or estimator behavior, and the severe degradation under coordinate jitter equivalent to 2% of torso length identified pose stability as a critical operating constraint. Collectively, these results replace an overly favorable binary interpretation with stronger evidence about neutral-case difficulty, feature utility, estimator dependence, and robustness boundaries. SpineTrace should therefore be used as a transparent research reference for evaluating compact gait representations, not as a diagnostic or deployment-ready system. Future work will conduct participant-grouped validation on radiographically labeled external cohorts, compare pose estimators on identical raw videos, and establish clinically prespecified thresholds, calibration, subgroup performance, referral workload, and end-to-end processing reliability.

ACKNOWLEDGMENTS

This research was supported by the SASMEC @ IIUM Research Grant (Project ID: SRG25-098-0098). Ethical approval for this study, including the collection of additional scoliosis X-ray and ultrasound images, was granted by the IIUM Research Ethics Committee (IREC) under application number IREC-2025-3. The Scoliosis1K dataset analyzed in this study is publicly available, and the authors did not collect any patient-identifiable data for the experiments reported in this paper.

REFERENCES

- [1] M. R. Konieczny, H. Senyurt, and R. Krauspe, "Epidemiology of adolescent idiopathic scoliosis," *Journal of Children's Orthopaedics*, vol. 7, no. 1, pp. 3-9, 2013, doi: 10.1007/s11832-012-0457-4.
- [2] H. Yılmaz, C. Zateri, A. Kusvuran Ozkan, G. Kayalar, and H. Berk, "Prevalence of adolescent idiopathic scoliosis in Turkey: An epidemiological study," *The Spine Journal*, vol. 20, no. 6, pp. 947-955, 2020, doi: 10.1016/j.spinee.2020.01.008.
- [3] S. Negrini *et al.*, "2016 SOSORT guidelines: Orthopaedic and rehabilitation treatment of idiopathic scoliosis during growth," *Scoliosis and Spinal Disorders*, vol. 13, p. 3, 2018, doi: 10.1186/s13013-017-0145-8.
- [4] A. Simony, E. J. Hansen, S. B. Christensen, L. Y. Carreon, and M. O. Andersen, "Incidence of cancer in adolescent idiopathic scoliosis patients treated 25 years previously," *European Spine Journal*, vol. 25, no. 10, pp. 3366-3370, 2016, doi: 10.1007/s00586-016-4747-2.
- [5] D. Y. T. Fong *et al.*, "A meta-analysis of the clinical effectiveness of school scoliosis screening," *Spine*, vol. 35, no. 10, pp. 1061-1071, 2010, doi: 10.1097/BRS.0b013e3181bcc835.
- [6] Z. Peng *et al.*, "Graph convolutional networks for 3D skeleton-based scoliosis screening using gait sequences," *The Visual Computer*, vol. 41, no. 9, pp. 6823-6835, 2025, doi: 10.1007/s00371-025-03983-w.
- [7] Z. Zhou, J. Liang, Z. Peng, C. Fan, F. An, and S. Yu, "Gait Patterns as Biomarkers: A Video-Based Approach for Classifying Scoliosis," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*: Springer, 2024, pp. 284-294.
- [8] Z. Zhou, Z. Peng, D. Jin, C. Fan, F. An, and S. Yu, "Pose as Clinical Prior: Learning Dual Representations for Scoliosis Screening," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*: Springer, 2026, pp. 464-474.
- [9] R. M. C. Aroeira, E. B. de Las Casas, A. E. d. M. Pertence, M. Greco, and J. M. R. S. Tavares, "Non-invasive methods of computer vision in the posture evaluation of adolescent idiopathic scoliosis: A systematic review,"

- Journal of Bodywork and Movement Therapies*, vol. 20, no. 4, pp. 832-843, 2016, doi: 10.1016/j.jbmt.2016.02.004.
- [10] G. Goto, K. Ariga, N. Tanaka, K. Oda, H. Haro, and T. Ohba, "Clinical significance of pose estimation methods compared with radiographic parameters in adolescent patients with idiopathic scoliosis," *Spine Surgery and Related Research*, vol. 8, no. 5, pp. 485-493, 2024, doi: 10.22603/ssrr.2023-0269.
 - [11] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," *arXiv preprint arXiv:2006.10204*, 2020, doi: 10.48550/arXiv.2006.10204.
 - [12] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems*, 2022, vol. 35, pp. 38571-38584.
 - [13] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989, doi: 10.1109/5.18626.
 - [14] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43-49, 1978, doi: 10.1109/TASSP.1978.1163055.
 - [15] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, 1 ed., pp. 8126-8133, doi: 10.1609/aaai.v33i01.33018126.
 - [16] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14648-14656, doi: 10.1109/ICCV48922.2021.01440.
 - [17] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32, 1 ed., pp. 7444-7452, doi: 10.1609/aaai.v32i1.12328.
 - [18] H. Li *et al.*, "Text-guided multi-instance learning for scoliosis screening via gait video analysis," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*: Springer Nature Switzerland, 2025, pp. 649-659.
 - [19] K. Vishwakarma and A. Nigam, "ScoliGaitX: A deep multi-modal fusion network for scoliosis assessment via gait video analysis," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.
 - [20] D. Chen *et al.*, "Clinical-prior guided multi-modal learning with latent attention pooling for gait-based scoliosis screening," *arXiv preprint arXiv:2602.06743*, 2026, doi: 10.48550/arXiv.2602.06743.
 - [21] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Computer Vision – ECCV 2014*: Springer, 2014, pp. 740-755.
 - [22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273-297, 1995, doi: 10.1007/BF00994018.

BIOGRAPHY OF AUTHORS



Teddy Surya Gunawan is a Professor in the Department of Electrical and Computer Engineering at the International Islamic University Malaysia (IIUM). He obtained his B.Eng. with cum laude honors in Electrical Engineering from Institut Teknologi Bandung (ITB), Indonesia, in 1998; his M.Eng. from Nanyang Technological University, Singapore, in 2001; and his Ph.D. from the University of New South Wales (UNSW), Australia, in 2007. Prof. Gunawan has held esteemed roles, including Visiting Research Fellow at UNSW (2010–2021) and Adjunct Professor at Telkom University (2022–2023), and he has been a Senior Member of IEEE since 2012, previously chairing the IEEE Instrumentation and Measurement Society – Malaysia Section. Recognized for his contributions to speech and audio processing, biomedical signal processing, image and video processing, and parallel computing, he received IIUM's Best Researcher Award in 2018 and is listed among the World's Top 2% Scientists in Artificial Intelligence and Image Processing by Elsevier for 2023, 2024, and 2025. In addition to his academic and research accomplishments, Prof. Gunawan holds multiple professional engineering certifications, including Fellow Member and Chartered Professional Engineer (CPEng) of Engineers Australia; Insinyur Profesional Utama (PII, Indonesia); ASEAN Engineer; ASEAN Chartered Professional Engineer; and APEC Engineer, reflecting his commitment to professional excellence. Within IIUM, he has also served as Head of Department (2015–2016) and Head of Programme Accreditation and Quality Assurance (2017–2018) at the Faculty of Engineering, reinforcing his leadership and expertise.



Siti Aishah Sofea binti Asri is an undergraduate researcher at the Kulliyah of Engineering, International Islamic University Malaysia (IIUM), where she is completing her Bachelor of Electrical and Electronics Engineering with Honours. Backed by international study exposure at the University of Limoges, France, her research interests include speech and image processing, advanced computer systems, and digital signal processing. Her final-year project focuses on automated scoliosis detection using advanced image analysis techniques.



Mira Kartiwi is a Professor in the Department of Information Systems, Kulliyah of Information and Communication Technology, and currently Director of the Data Management and Data Analytics Section of the Office of Strategic Institution Transformation, International Islamic University Malaysia (IIUM). She was one of the recipients of the Australia Postgraduate Award (APA) in 2004 and received the Higher Degree Research Award for Excellence in 2007 in recognition of her research achievements. She is also listed among the World's Top 5% Scientists by SciRank Global. In recognition of her expertise, she has been appointed to the Editorial Board of several local and international journals. Dr. Kartiwi is also an experienced consultant specializing in the health, financial, and manufacturing sectors. Her areas of expertise include health informatics, e-commerce, data mining, data analytics, information systems strategy, business process improvement, product development, marketing, delivery strategy, workshop facilitation, training, and communications.



Noreha Abdul Malik is an Associate Professor in the Department of Electrical and Computer Engineering at the Kulliyah of Engineering, International Islamic University Malaysia (IIUM), Gombak Campus. She obtained her Bachelor of Engineering in Electrical-Electronics Engineering (Medical Electronics) from Universiti Teknologi Malaysia (UTM), her Master of Engineering (Communication and Computer Engineering) from Universiti Kebangsaan Malaysia (UKM), and her Ph.D. in Electronics and Electrical Engineering from the University of Southampton, UK. Dr. Noreha has over two decades of academic experience, coordinating core courses in programming, circuit analysis, and integrated design. Her interdisciplinary research portfolio spans biomedical electronics, biosignal processing (including EEG, PPG, and respiratory acoustic analysis), non-invasive healthcare monitoring systems, and Internet of Things (IoT) applications.



Ahmad Hud Bin Mahfoodz serves as a faculty member and clinical specialist at the Kulliyah of Medicine, International Islamic University Malaysia (IIUM) Kuantan. Specializing in medical and health sciences, he has clinical expertise in surgery and orthopedic care. Dr. Ahmad Hud completed his specialized medical training with a Master's in Orthopedic Surgery from IIUM, where he now contributes to undergraduate and postgraduate clinical education and surgical research.



Mohd Ghaddafi Bin Wahab is a neurosurgeon and academic faculty member at the Kulliyah of Medicine, IIUM Kuantan Campus. He completed his postgraduate clinical specialization in neurosurgery at Universiti Sains Malaysia (USM). His academic contributions include extensive teaching in both general and specialized surgical postings, alongside active research in neuro-oncology and neurotrauma. His recently published work addresses diagnostic and management challenges in central nervous system tumors and intracranial co-infections. He currently co-supervises advanced clinical research tracking prognostic biomarkers for traumatic brain injuries and intracranial hemorrhages.