ACTA ORTHOPAEDICA et TRAUMATOLOGICA TURCICA
www.aott.org.tr

Research Article

Predicting mechanical complications in adult spinal deformity patients with postoperative proportioned and moderately disproportioned alignment

Baris Balaban[1], Nuri Demirci[2], Caglar Yilgor[3], Altug Yucekul[3], Tais Zulemyan[4], Sleiman Haddad[5], Shahnawaz Haleem[6], Feyzi Kilic[4], Ibrahim Obeid[7], Javier Pizones[8], Frank Kleinstueck[9], Francisco Javier Sanchez Perez[8], Ferran Pellise[5], Ahmet Alanay[3], Cetin Bagci[10], Osman Ugur Sezerman[1], European Spine Study Group[11]

[1]*Department of Biostatistics and Bioinformatics, Institute of Health Sciences, Acibadem Mehmet Ali Aydinlar University, Istanbul, Türkiye*
[2]*Acibadem University School of Medicine, Istanbul, Türkiye*
[3]*Department of Orthopedics and Traumatology, Acibadem University School of Medicine, Istanbul, Türkiye*
[4]*Comprehensive Spine Center, Acibadem Maslak Hospital, Istanbul, Türkiye*
[5]*Spine Surgery Unit, Hospital Universitari Vall d'Hebron, Barcelona, Spain*
[6]*Spine Unit, Royal Orthopaedic Hospital, Birmingham, UK*
[7]*Clinique du Dos, Elsan Jean Villar Private Hospital, Bordeaux, France*
[8]*Spine Surgery Unit, Hospital Universitario La Paz, Madrid, Spain*
[9]*Spine Center Division, Department of Orthopedics and Neurosurgery, Schulthess Klinik, Switzerland*
[10]*Bilmed Computer and Software Company, Istanbul Türkiye*
[11]*Colloborators listed at the end of the manuscript.*

ARTICLE INFO

Article history:
Submitted October 23, 2024
Received in revised form March 18, 2025
Last revision received form June 16, 2025
Accepted June 17, 2025
Publication Date July 18, 2025

Keywords:
Adult spinal deformity surgery
GAP score
Interpretable trees
Machine learning
Random forest

ORCID iDs of the authors:
B.B. 0000-0001-6573-402X;
N.D. 0000-0001-5820-1591;
C.Y. 0000-0002-5372-995X;
A.Y. 0000-0002-1942-2444;
T.Z. 0000-0003-3194-5923;
S.H. 0000-0002-1942-2444;
S.H. 0000-0003-2678-9428;
F.K. 0000-0001-9757-3013;
I.O. 0000-0002-5602-0508;
J.P. 0000-0001-7394-9637;
F.K. 0000-0003-2057-2201;
F.J.S-P. 0000-0003-1112-5995;
F.P. 0000-0002-0644-7757;
A.A. 0000-0003-2252-6647;
C.B. 0009-0001-3997-4246;
O.U.S. 0000-0003-0905-6783.

Corresponding author:
Baris Balaban and Caglar Yilgor
baris.balaban92@gmail.com,
caglaryilgor@gmail.com

Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ABSTRACT

Objective: Mechanical complications are common after adult spinal deformity (ASD) surgery and can significantly impair outcomes. This study aimed to predict such complications in proportioned and moderately disproportioned patients using a machine learning approach, to inform preoperative planning and enable early preventive care.

Methods: Prospectively collected clinical data, including preoperative, intraoperative, and postoperative variables, radiographic parameters, technical details, and patient-reported outcomes, were obtained from a multi-center ASD surgery database. Parameter tuning of a random forest (RF) classifier was performed using 9-times 3-fold cross-validation over 3 rounds of grid search, with the F-score used as the primary optimization metric. The final RF model was used to derive a clinically interpretable rule set using the inTrees framework. Permutation-based feature importance was assessed for F-score, accuracy, and sensitivity.

Results: The model was trained on 295 patients (237 female, 58 male; mean age, 50 ± 19 years) with a minimum 2-year follow-up (mean 53 months, range 24-101). Mechanical complications were observed in 100 patients (34%). A test cohort of 98 patients (33% complication rate) was used for external validation. The RF model achieved 72% accuracy, 91% sensitivity, 64% specificity, and 93% negative predictive value. The derived rule set, comprising 8 rules using 1 to 3 features each, yielded 74% accuracy, 81% sensitivity, 71% specificity, and 83% negative predictive value. The location of the lower instrumented vertebra (LIV) was the most influential predictor.

Conclusion: By excluding patients with severe deformities, as defined by the GAP score, this study focused on the more clinically ambiguous group of proportioned and moderately disproportioned patients. To the authors' knowledge, this is the first study to develop predictive tools specifically for this subgroup to assess the risk of mechanical complications following ASD surgery. These tools may assist in early risk stratification and guide preoperative decision-making to reduce postoperative complications and improve patient outcomes.

Level of Evidence: Level III, Prognostic Study.

Introduction

Since the inception of artificial intelligence in 1950, machine learning has undergone transformational growth, especially over the last 3 decades, propelled by unprecedented advances in computational capabilities and an exponential increase in data generation. This evolution has led to the widespread implementation of machine learning techniques across diverse sectors, prominently in healthcare where its impact is reconstructing numerous aspects of patient care and medical diagnostics. Beyond diagnostics, machine learning's capabilities extend to predicting patient outcomes and stratifying patients. This capability allows healthcare providers to initiate early and personalized treatment interventions, thus shifting from reactive to proactive and preventive care paradigms.

The Global Alignment and Proportion (GAP) score methodology classifies adult spinal deformity (ASD) patients into 3 groups: proportioned, moderately disproportioned, and severely disproportioned.[1] In the original paper, 95% of the severely disproportioned patients suffered from at least 1 mechanical complication. This study attempts to improve the current understanding of the effects of preoperative, operative, and postoperative features on mechanical complications by excluding severely disproportioned patients not experiencing a mechanical complication as their numbers were lower in comparison to the other groups thus preventing any profound conclusions.

Cite this article as: *Balaban B, Demirci N, Yilgor C, et al. Predicting mechanical complications in adult spinal deformity patients with postoperative proportioned and moderately disproportioned alignment. Acta Orthop Traumatol Turc., 2025;59(4):210-221.*

DOI: 10.5152/j.aott.2025.24146

This study aims to predict mechanical complications in ASD patients that are proportioned or moderately disproportioned according to the GAP score methodology and have a high sensitivity rate to ensure that preventive treatments are not missed for such patients. The cost of reoperation is significantly greater than the expense of applying low precision preventive treatments.

Grinsztajn et al's[2] extensive research comparing the effectiveness of deep learning and machine learning models on tabular data concluded that using machine learning models has superior performance metrics and speed for datasets smaller than 10 000 samples. The European Spine Study Group (ESSG) dataset like many other datasets in healthcare has multicollinearity hence the need to select an algorithm not sensitive to multicollinearity. The decision tree algorithm is one such algorithm as it selects only 1 feature when applying a split.[3] Ensemble tree-based models perform better in most cases as they balance out the bias from each tree of an ensemble.[4] Hence, the random forest (RF) algorithm is chosen for this study to model ESSG patients. The drawback of using such ensemble models is making the model so complex that it is not actionable in the clinic. To overcome this problem, a rule set is crafted using the frequent patterns in the RF analysis.

Materials and methods

Dataset

Adult spinal deformity surgeries' clinical data was collected prospectively from 6 operation sites between 2010 and 2019. Institutional review board approval was obtained for patient enrollment and data collection protocols at each site. his study was approved by the Acıbadem University Medical Research Evaluation Board (Acıbadem Üniversitesi Tıbbi Araştırmalar Değerlendirme Kurulu – ATADEK) with the approval reference number 2019-11/34. This study took place at Acıbadem University and ATADEK is the medical board of Acıbadem University. This study took place at Acibadem University, and each patient provided written consent before participation in the study. All the methods were carried out following the Declaration of Helsinki.

Patients that were 18 and older, and had at least coronal Cobb angle ≥20°, sagittal vertical axis ≥5 cm, pelvic tilt ≥25° or thoracic kyphosis ≥60° were enrolled to the ESSG database. The inclusion criteria of this study were to have a postoperative GAP score of at most 6 (proportioned and moderately disproportioned patients according to GAP score categories) and have at least 2 years of follow-up. Thus, severely disproportioned patients according to the GAP score were left out.

The dependent variable was the presence of mechanical complication(s), which might be proximal junctional kyphosis and failure, distal junctional kyphosis and failure, rod breakage, and implant-related complications.

Preoperative, operative, and postoperative patient characteristics, radiographic measurements, technical details, and patient-reported outcome measures were included in the analysis. The variables are detailed in Table 1. Out of 166 features that were collected in the ESSG database of ASD patients, 13 features were eliminated which had more than 85% missing data. The number of preoperative, operative, and postoperative features after the elimination was 78, 53, and 22, respectively. The descriptive statistics were reported not for all features, but the ones that were found to be important in the analyses.

About 25% of the data (98 patients) was left for testing using stratified sampling using mechanical complication (dependent variable), preoperative, and postoperative GAP categories for a more accurate sample of ASD surgery patients.

Random forest classification

Random forest is a non-parametric, supervised machine learning method based on an ensemble of decision trees, commonly used for both classification and regression tasks.[5] Each decision tree partitions the data through recursive binary splits, providing an intuitive and interpretable framework. Random forest enhances the predictive power of individual trees by combining many of them through bootstrap aggregation, or "bagging," a technique in which each tree is trained on a randomly selected subset of the data. This approach helps reduce overfitting and the influence of outliers, thereby improving generalization.[6,7] In addition to sampling data, RF introduces randomness in feature selection at each node, further increasing model stability.

In this study, all RF models were developed using R (version 4.3.2) with the "dplyr," "caret," and "randomForest" packages.[8-11] To address the class imbalance—where patients without complications were more prevalent, a 2:1 class weight ratio was applied to penalize false negatives more heavily, thus prioritizing the accurate identification of patients at risk for mechanical complications.

Clinically interpretable rules derived from the model

Although RF models offer strong predictive performance, their ensemble structure, comprising hundreds of decision trees, makes clinical interpretation challenging. Previous efforts to address this issue include the use of linear models (Friedman et al[12]) and association rule mining or classification techniques (Liu et al,[13] Deng et al[14]). Building on these foundations, the InTrees (interpretable trees) framework was developed to extract simplified and meaningful rules from complex tree ensembles.

InTrees operates by identifying, refining, and evaluating patterns within the ensemble. It also incorporates a rule-based learner, known as the simplified tree ensemble learner (STEL), which enhances interpretability by emphasizing frequently interacting variables.[15] In this study, the following inTrees pipeline was applied:

1. Frequent Pattern Extraction: Rules were extracted from the RF model, each annotated with support and confidence values. Here, support represents how often a rule appears in the dataset.
2. Rule Selection and Pruning: A non-redundant set of clinically relevant rules was selected using a guided regularized RF method, which incorporated prior feature importance scores. Rules were then pruned to reduce complexity.[16]
3. Simplified Rule Learning: Using the STEL algorithm, a prioritized rule set was generated to solve the classification task. Rules were added sequentially, beginning with the 1 showing the lowest classification error. Rules with higher support and fewer conditions were preferred, enhancing simplicity and generalizability.

HIGHLIGHTS

- A random forest model and a ruleset using interpretable trees to predict mechanical complications in spinal deformity surgeries.
- Significant factors to predict mechanical complications in postoperatively well-aligned and moderately disproportioned patients using permutation feature importance.
- The location of the lower instrumented vertebra is the most influential variable.

Table 1. Variables used in the analyses

Category	Features
Preoperative features	Demographic: Gender, age, height, weight, BMI Medical history: Diagnosis, duration of spine problems, number of prior spine surgeries, timing of last surgery, presence of mechanical complications Radiographic measurements: Pelvic incidence, various kyphosis measurements, Cobb angles, distances to CSVL, coronal balance, trunk shift, clavicle angle, shoulder height, T1-Rib angle, leg length discrepancy, pelvic obliquity Bone quality: BMD, menopause presence and duration, number of pregnancies General health: Occupation, smoking status, ASA score, comorbidities Patient-reported outcomes: NRS scores, COMI score, ODI score, SF-36 MCS and PCS, SRS-22 score
Operative features	Demographic: Surgery site, year of surgery, surgeon experience Surgical efficiency: Surgery duration, EBL, urine output, intraoperative transfusions, ICU and hospital stay Technique: Number of stages, surgical approach, number of levels fused, UIV and LIV locations, sacroiliac fixation, construct location and rigidity, rod details, osteotomies, thoracoplasty, graft materials, antifibrinolytics usage
Postoperative features	Follow-up duration Radiographic measurements: (Same as preoperative) Changes in measurements: Delta values for thoracic kyphosis and lumbar lordosis

ASA, American Society of Anesthesiologists; BMD, bone mineral density; BMI, body mass index; COMI, Core Outcome Measures Index; CSVL, central sacral vertical line; EBL, estimated blood loss; ICU, intensive care unit; LIV, lower instrumented vertebra; MCS, mental component summary; NRS, Numeric Rating Scale; ODI, Oswestry Disability Index; PCS, physical component summary; SF-36, Short Form-36; SRS-22, Scoliosis Research Society-22 Patient Questionnaire; UIV, upper instrumented vertebra.

In this framework, each rule was selected based on its predictive contribution, and once applied, the corresponding subset of data was removed from subsequent selection steps resulting in a compact, interpretable rule set suitable for clinical decision-making.

Data imputation and feature engineering

Among the 153 features included in the analysis, 38 contained missing values. On average, each feature had 19.6 missing entries, ranging from 1 to 34. In the training set, missing values were imputed using the median of the specific data subset used to build the corresponding tree within the RF algorithm. For the test set, median values derived from the overall training data were used for imputation to ensure consistency when calculating model performance metrics.

The STEL algorithm is capable of handling missing values; therefore, imputation was not performed during its construction. However, for evaluation purposes, the imputed version of the test set was used to allow comparison with the RF model using the same data.

Given the bimodal distribution of GAP-related variables, Relative Pelvic Version, Relative Lumbar Lordosis, Lordosis Distribution Index, and Relative Spinopelvic Alignment (RSA), these were replaced with statistical weight values based on subgroup classifications described by Yilgor et al.[1]

To support effective decision tree splits, all features were ordinally encoded. For example, comorbidity data included both the total number of comorbidities per patient and separate binary indicators for specific conditions. Similarly, neurological history was encoded as a binary variable (0=no history, 1=presence of neurological issues).

Optimizing the random forest model using cross-validation

To determine the optimal parameter set for the RF model, a 9-times 3-fold cross-validation procedure was conducted. The parameters tuned included: the number of features considered at each split (mtry), the maximum number of terminal nodes allowed in a tree (maxnodes), the minimum number of observations per terminal node (nodesize), and the total number of trees in the forest. Although increasing the number of trees theoretically improves accuracy, prior work has shown that after a certain threshold, the benefit becomes negligible or even detrimental to performance.[17] Therefore, it is included in the parameter set to analyze the performance metrics.

During each cross-validation round, the training dataset was stratified into training and validation subsets, using the same variables that were employed for the overall training and testing split, namely, mechanical complication status (the dependent variable), along with preoperative and postoperative GAP categories.

Three grid searches were conducted to tune the model, each progressively narrowing the search space and expanding the number of parameter combinations explored. For all grid searches, the mean F-score across validation folds served as the primary evaluation metric. In the initial search, default values were used for maxnodes and nodesize to establish a baseline. The second and third rounds refined the search based on prior results. The F-score, calculated as the harmonic mean of precision and recall, was prioritized over area under the receiver operating characteristic curve (AUC) due to the class imbalance in the dataset. This metric better reflected the model's ability to accurately identify patients likely to develop mechanical complications.

Model outputs and performance metrics

During each round of cross-validation, the mean and SD of several key performance metrics were calculated, including the F-score, AUC, sensitivity, specificity, area under the precision-recall curve, and precision. Scatter plots were generated for the second and third grid searches to visualize how performance varied across parameter settings (e.g., mean F-score vs. parameter values).

In the results section, the out-of-bag error curve is presented as a function of the number of trees, as well as performance metrics for the final RF model and the simplified decision rule set derived from it. Model performance was evaluated on both the training and held-out test datasets.

Feature importance for the RF model was first assessed using the built-in method, which calculates the average decrease in Gini impurity across all trees.[11] However, prior studies have noted that this method may introduce bias, particularly favoring categorical variables with many levels. To address this limitation and allow consistent comparison between the RF and STEL models, permutation-based feature importance analysis was also performed using F-score, accuracy, and sensitivity as evaluation metrics.[18]

To address multicollinearity, features were grouped using hierarchical clustering based on the dissimilarity matrix derived from 1 minus the absolute value of correlation coefficients.[19] Complete linkage was used to form the clusters. Variable groups were defined by cutting the resulting dendrogram at a height of 0.5, using the "cutree" function from the R stats package.[8]

Next, these clusters, along with any remaining ungrouped features, were permuted 10 000 times, each time keeping all other feature values unchanged. For each permutation, model performance metrics

were recalculated. The decreases of the F-score, accuracy, and sensitivity metrics were presented using error rates (($e_{permuted}$ – $e_{original}$) -1) for each feature in the results section.

Results

The training set of the models consisted of 295 patients (237 Female, 58 Male), and 100 (34%) of them had suffered from at least 1 mechanical complication in the follow-up period which was a minimum of 2 years. 25% of the whole dataset (393 patient records) was held out with stratified sampling containing 98 patients and 32 (33%) of them had suffered from at least 1 mechanical complication.

An RF model with 500 trees was built, and for simplification, a rule-based method was created to predict the possibility of mechanical complications from this RF model with a simplified tree ensemble learner resulting 8 rules containing 1 to 3 features in each.

Descriptive statistics of important features

Descriptive statistics of the features that were the top 10 most significant features found by permutation feature importance in all metrics and/or present in the rule-based decision approach were presented in 2 tables. In Table 2, the preoperative features are presented and in Table 3, the operative and the postoperative features are presented. Both tables include the Pearson correlation test *P*-values with the response variable. In Supplementary Table 1, 30 other randomly selected features' descriptive statistics are presented to avoid reporting bias.

Random forest cross-validation results

Means (SDs) of 9-times 3-fold cross-validation performance metrics are given in Supplementary Table 2 for each grid search, along with the tuned parameters.

In the first grid search, 42 parameter combinations were evaluated. This included 7 different values for the number of features considered at each split (mtry: 13 to 150) and 6 different settings for the number of trees (10 to 1000). During this initial search, the values for nodesize and maxnodes were kept at their defaults to establish a performance baseline.

The second grid search tested 840 combinations, expanding the search space to include 7 mtry values (13 to 150), 5 tree sizes (10 to 500), 6 nodesizes (1 to 45), and 4 maxnodes values (4 to 16).

In the third and most refined search, 2,916 parameter combinations were assessed. This included 9 values for mtry (30 to 70), 9 tree sizes (100 to 500), 6 nodesizes (16 to 31), and 6 maxnodes values (5 to 10).

Final random forest model

The final RF model was built using the entire training dataset and the optimized parameters identified in the third grid search: 500 trees, mtry=55, nodesize=19, and maxnodes=6. Of the 153 total input features, 24 (16%) were not used by the model. The out-of-bag error curve in relation to the number of trees is shown in Figure 1, and performance metrics for the training, out-of-bag, and test sets are summarized in Table 4.

In the test set, the final model achieved a sensitivity of 91%, which was a critical outcome given the study's goal of enabling early preventive intervention for patients at high risk of mechanical complications. This result indicates that only 9% of patients who eventually developed complications were not correctly identified by the model.

As expected, the training set metrics were higher than those of the out-of-bag and test sets. This discrepancy is attributable to the fact that approximately two-thirds of the training data were used in building each individual tree. Therefore, the out-of-bag and test results provide a more reliable estimate of the model's generalization performance.

Feature importance was evaluated based on the average decrease in Gini impurity across all trees, and the top 10 features are presented in Figure 2.[11] The results are shown in Figure 2, which lists the top 10 features. Feature names are prefixed to indicate their source: "pre" for preoperative, "op" for operative, and "post" for postoperative variables.

According to the importance ranking, the 5 most influential features were: postoperative GAP score, construct location, lower instrumented vertebra (LIV) location, estimated blood loss, and patient age.[11]

Rule set for clinical use derived from random forest

Frequent patterns were extracted using the default parameter settings of the "getFreqPattern" function, and the resulting patterns are presented in Supplementary Table 3. The table was sorted based on support values, which represent how frequently each rule appeared in the dataset. No frequent patterns involving 2 or more features exceeded a support threshold of 0.01.

A total of 2796 rules were derived from 500 decision trees trained on oversampled data. Since only 100 out of 295 patients in the training set experienced mechanical complications, patients in this group were sampled twice to correct class imbalance prior to rule extraction. The generated rules were then pruned using a decay parameter (maxDecay) set at 0.03. A subset of relevant, non-redundant rules was selected using regularized RFs via the "selectRuleRRF" function.

Next, the STEL algorithm was applied to the selected rules. It used a minimum support frequency of 0.1 to construct a reduced, clinically interpretable rule set for predicting mechanical complications. The final rule set is presented in Table 5.

When making a prediction for a new patient, the rules were applied sequentially from top to bottom. The first rule that matched the patient's profile was used to determine the predicted outcome. The overall performance metrics of this rule-based classifier are reported in Table 6.

Key predictors of mechanical complications identified by feature importance

Before determining the most impactful predictors, hierarchical clustering was used to group highly correlated features. Out of 129 features used in the final RF model, 80 clusters were created. The number of features per cluster ranged from 1 to 6. The clusters ranked as the top 10 most influential across all performance metrics are presented in Supplementary Table 4.

Permutation feature importance analysis was conducted by measuring how much model error increased when the values within each cluster were randomly shuffled. Results for the test set are shown in

Table 2. Descriptive statistics of the preoperative features

Features	Mechanical complications - present (n=100)	Mechanical complications - absent (n=195)	Total (n=295)	*P*
Age (years)				<.001
Mean (SD)	58.0 (15.4)	45.2 (18.9)	49.6 (18.8)	
Median (Q1, Q3)	63.4 (51.9, 67.3)	44.1 (26.1, 62.0)	54.1 (31.9, 65.4)	
Min - Max	18.6 - 82.7	17.8 - 82.4	17.8 - 82.7	
Missing	0	0	0	
BMI (kg/m²)				<.001
Mean (SD)	26.2 (3.9)	24.1 (4.3)	24.8 (4.3)	
Median (Q1, Q3)	26.0 (23.5, 28.8)	23.6 (20.7, 26.7)	24.1 (21.7, 27.4)	
Min - Max	17.7 - 36.0	16.1 - 38.6	16.1 - 38.6	
Missing	0	0	0	
Pre GAP score				<.001
Mean (SD)	7.7 (4.1)	5.0 (4.0)	5.9 (4.2)	
Median (Q1, Q3)	9.0 (4.0, 12.0)	4.0 (2.0, 8.0)	5.0 (3.0, 10.0)	
Min - Max	0.0 - 13.0	0.0 - 13.0	0.0 - 13.0	
Missing	0	0	0	
Pre TL/L apical translation to CSVL (mm)				.05
Mean (SD)	22.5 (20.7)	28.3 (22.7)	26.3 (22.2)	
Median (Q1, Q3)	16.3 (5.1, 38.5)	24.9 (7.2, 43.1)	21.8 (6.7, 42.2)	
Min - Max	0.0 - 81.8	0.0 - 82.6	0.0 - 82.6	
Missing	9	13	22	
Pre T10 L2 kyphosis (°)				<.001
Mean (SD)	16.2 (21.6)	8.2 (19.1)	10.9 (20.3)	
Median (Q1, Q3)	14.5 (4.0, 25.0)	6.0 (-6.0, 20.5)	10.0 (-2.0, 23.0)	
Min - Max	−39.0 to 93.0	−34.0 to 70.0	−39.0 to 93.0	
Missing	0	0	0	
Pre T2 T5 kyphosis (°)				.22
Mean (SD)	9.5 (10.8)	10.9 (11.0)	10.4 (11.0)	
Median (Q1, Q3)	7.5 (3.8, 14.2)	10.0 (4.0, 16.0)	9.0 (4.0, 15.5)	
Min - Max	−29.0 to 42.0	−18.0 to 71.0	−29.0 to 71.0	
Missing	0	0	0	
Pre major Cobb (°)				.12
Mean (SD)	35.9 (24.4)	43.3 (23.7)	40.8 (24.1)	
Median (Q1, Q3)	36.5 (15.8, 53.2)	45.0 (24.0, 58.3)	42.0 (21.0, 57.0)	
Min - Max	0.0 - 105.0	0.0 - 96.0	0.0 - 105.0	
Missing	0	0	0	
Pre coronal balance (°)				.315
Mean (SD)	25.7 (25.4)	22.4 (22.2)	23.5 (23.3)	
Median (Q1, Q3)	18.9 (9.7, 31.6)	16.3 (8.0, 28.5)	17.7 (8.1, 29.4)	
Min - Max	0.0 - 136.5	0.0 - 150.4	0.0 - 150.4	
Missing	4	9	13	
Pre ODI score				<.001
Mean (SD)	46.4 (19.0)	33.9 (20.0)	38.2 (20.5)	
Median (Q1, Q3)	46.0 (32.0, 60.0)	34.0 (16.0, 48.0)	40.0 (23.5, 52.2)	
Min - Max	0.0 - 90.0	0.0 - 80.0	0.0 - 90.0	
Missing	5	14	19	
Pre SRS-22 subtotal				<.001
Mean (SD)	2.5 (0.6)	3.0 (0.7)	2.8 (0.7)	
Median (Q1, Q3)	2.5 (2.1, 2.9)	2.9 (2.5, 3.4)	2.8 (2.4, 3.3)	
Min - Max	1.4 - 4.2	1.5 - 4.5	1.4 - 4.5	
Missing	4	14	18	
Pre ODI personal care				.01
Mean (SD)	1.3 (1.2)	0.9 (1.1)	1.1 (1.2)	
Median (Q1, Q3)	1.0 (0.0, 2.0)	1.0 (0.0, 2.0)	1.0 (0.0, 2.0)	
Min - Max	0.0 - 4.0	0.0 - 5.0	0.0 - 5.0	
Missing	6	15	21	

BMI, body mass index; CSVL, central sacral vertical line; GAP, global alignment and proportion; max, maximum; min, minimum; n, number; ODI, Oswestry Disability Index; pre, preoperative; Q1, first quartile; Q3, third quartile; SRS-22, Scoliosis Research Society-22 Patient Questionnaire; TL/L, thoracolumbar/lumbar.

Figures 3-5, while results for the training set appear in Supplementary Table 5. Clusters were reported if the increase in error exceeded 0.5%. It is important to note that feature names may represent clusters rather than individual variables.

In the RF model, 24 clusters were found to significantly impact F-score performance (25 in training). Similarly, 5 clusters (9 in training) were identified as influential for the rule set. For accuracy, 24 clusters (25 in training) were significant in the RF, and 7 clusters (9

Table 3. Descriptive statistics of the operative and postoperative features

Features	Mechanical complications - present (n=100)	Mechanical complications - absent (n=195)	Total (n=295)	*P*
Op. LIV location*				<.001
Mean (SD)	1.0 (1.6)	2.4 (1.8)	1.9 (1.9)	
Median (Q1, Q3)	0.0 (0.0, 2.0)	3.0 (0.0, 4.0)	2.0 (0.0, 4.0)	
Min - Max	0.0 - 5.0	0.0 - 6.0	0.0 - 6.0	
Missing	0	0	0	
Op. sacro iliac fixation				<.001
Mean (SD)	0.7 (0.4)	0.4 (0.5)	0.5 (0.5)	
Median (Q1, Q3)	1.0 (0.0, 1.0)	0.0 (0.0, 1.0)	0.0 (0.0, 1.0)	
Min - Max	0.0 - 1.0	0.0 - 1.0	0.0 - 1.0	
Missing	0	0	0	
Op. iliac fixation or not				<.001
Mean (SD)	0.8 (0.6)	0.3 (0.5)	0.5 (0.6)	
Median (Q1, Q3)	1.0 (0.0, 1.0)	0.0 (0.0, 1.0)	0.0 (0.0, 1.0)	
Min - Max	0.0 - 2.0	0.0 - 2.0	0.0 - 2.0	
Missing	0	0	0	
Op. construct location**				<.001
Mean (SD)	4.5 (1.0)	3.5 (1.1)	3.8 (1.1)	
Median (Q1, Q3)	5.0 (3.0, 5.0)	3.0 (3.0, 5.0)	3.0 (3.0, 5.0)	
Min - Max	1.0 - 8.0	1.0 - 6.0	1.0 - 8.0	
Missing	0	0	0	
Op. delta lumbar Lordosis (°)				<.001
Mean (SD)	15.1 (20.9)	6.5 (17.0)	9.4 (18.8)	
Median (Q1, Q3)	16.0 (0.0, 32.2)	4.0 (-5.0, 16.5)	7.0 (-5.0, 22.0)	
Min - Max	−25.0 to 70.0	−44.0 to 68.0	−44.0 to 70.0	
Missing	0	0	0	
Op. delta thoracic kyphosis (°)				.177
Mean (SD)	5.1 (17.3)	2.6 (15.0)	3.4 (15.8)	
Median (Q1, Q3)	5.5 (-6.2, 19.2)	3.0 (-6.0, 13.0)	4.0 (-6.0, 15.0)	
Min - Max	−41.0 to 42.0	−75.0 to 39.0	−75.0 to 42.0	
Missing	0	0	0	
Op. number of levels fused				.003
Mean (SD)	11.8 (3.7)	10.4 (3.6)	10.9 (3.7)	
Median (Q1, Q3)	11.0 (9.0, 15.0)	10.0 (7.0, 13.0)	10.0 (8.0, 13.0)	
Min - Max	4.0 - 20.0	4.0 - 18.0	4.0 - 20.0	
Missing	0	0	0	
Op. estimated blood loss (mL)				<.001
Mean (SD)	1763.3 (1175.4)	1227.4 (1173.5)	1409.1 (1199.4)	
Median (Q1, Q3)	1475.0 (900.0, 2300.0)	900.0 (500.0, 1625.0)	1000.0 (600.0, 1862.5)	
Min - Max	229.0 - 5380.0	40.0 - 7900.0	40.0 - 7900.0	
Missing	0	0	0	
Op. osteotomy Detailed***				<.001
Mean (SD)	1.1 (1.1)	0.7 (1.0)	0.8 (1.1)	
Median (Q1, Q3)	1.0 (0.0, 2.0)	0.0 (0.0, 1.0)	1.0 (0.0, 1.0)	
Min - Max	0.0 - 5.0	0.0 - 5.0	0.0 - 5.0	
Missing	0	0	0	
Op. osteotomy posterior column number				.098
Mean (SD)	1.1 (1.7)	0.9 (1.9)	1.0 (1.8)	
Median (Q1, Q3)	0.0 (0.0, 2.0)	0.0 (0.0, 1.0)	0.0 (0.0, 1.0)	
Min - Max	0.0 - 7.0	0.0 - 9.0	0.0 - 9.0	
Missing	0	0	0	
Op. dual versus multiple rods				.27
Mean (SD)	0.3 (0.6)	0.2 (0.5)	0.2 (0.5)	
Median (Q1, Q3)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	
Min - Max	0.0 - 3.0	0.0 - 2.0	0.0 - 3.0	
Missing	0	0	0	
Op. hospitalization time (days)				<.001
Mean (SD)	14.9 (9.7)	11.1 (6.9)	12.4 (8.2)	
Median (Q1, Q3)	12.0 (9.8, 16.2)	10.0 (7.5, 12.5)	10.0 (8.0, 14.0)	
Min - Max	4.0 - 59.0	4.0 - 54.0	4.0 - 59.0	
Missing	0	0	0	

(*Continued*)

Table 3. Descriptive statistics of the operative and postoperative features (*Continued*)

Features	Mechanical complications - present (n=100)	Mechanical complications - absent (n=195)	Total (n=295)	*P*
Post GAP score				<.001
Mean (SD)	3.5 (1.9)	2.0 (1.9)	2.5 (2.0)	
Median (Q1, Q3)	4.0 (2.0, 5.0)	1.0 (0.0, 3.0)	2.0 (1.0, 4.0)	
Min - Max	0.0 - 6.0	0.0 - 6.0	0.0 - 6.0	
Missing	0	0	0	
Post RSA (°)				.002
Mean (SD)	0.6 (0.7)	0.4 (0.6)	0.4 (0.6)	
Median (Q1, Q3)	0.5 (0.0, 1.0)	0.0 (0.0, 1.0)	0.0 (0.0, 1.0)	
Min - Max	0.0 - 3.0	0.0 - 3.0	0.0 - 3.0	
Missing	0	0	0	
Post T2 T5 kyphosis (°)				.072
Mean (SD)	12.3 (9.0)	13.7 (8.8)	13.2 (8.8)	
Median (Q1, Q3)	12.0 (6.0, 17.0)	13.0 (8.0, 20.0)	13.0 (7.0, 19.0)	
Min - Max	−6.0 to 49.0	−11.0 to 43.0	−11.0 to 49.0	
Missing	0	0	0	
Post TL/L Cobb (°)				.78
Mean (SD)	15.7 (15.0)	15.9 (13.4)	15.8 (14.0)	
Median (Q1, Q3)	13.0 (2.0, 25.1)	14.3 (5.0, 26.0)	14.0 (3.0, 26.0)	
Min - Max	0.0 - 66.0	0.0 - 63.0	0.0 - 66.0	
Missing	5	21	26	
Follow-up duration (weeks)				<.001
Mean (SD)	52.5 (19.0)	41.7 (17.1)	45.3 (18.4)	
Median (Q1, Q3)	55.9 (35.3, 63.7)	35.9 (26.2, 52.9)	40.9 (27.9, 60.4)	
Min - Max	24.1 - 101.4	24.0 - 100.0	24.0 - 101.4	
Missing	0	0	0	

GAP, global alignment and proportion; LIV, lower instrumented vertebra; max, maximum; min, minimum; n, number; op, operative; post, postoperative; Q1, first quartile; Q3, third quartile; RSA, relative spinopelvic alignment; TL/L, thoracolumbar/lumbar.
*LIV location feature encoding is as follows: 0 ilium, 1 S1–S2, 2 L5, 3 L4, 4 L3-L2, 5 T12-L1, 6 T11, or higher.
**Construct location encoding is as follows: 0 thoracic only, 1 lumbar only, 2 thoracolumbar, 3 lumbosacral, 4 thoracolumbosacral, 5 cervicothoracic, 6 cervicothoracolumbar, 7 cervicothoracolumbosacral.
***Osteotomy detailed encoding is as follows: 0 none, 1 Ponte osteotomy, 2 Smith-Petersen osteotomy, 3 pedicle subtraction osteotomy, 4 bone-disc-bone osteotomy, 5 vertebral column resection.

in training) in the rule set. For sensitivity, 19 clusters (19 in training) were important in the RF model, and 4 clusters (7 in training) in the rule-based model.

The rankings of key features were largely consistent between training and test sets, except for the sensitivity metric, where minor reordering was observed between the 4th and 5th, and the 7th and 8th positions.

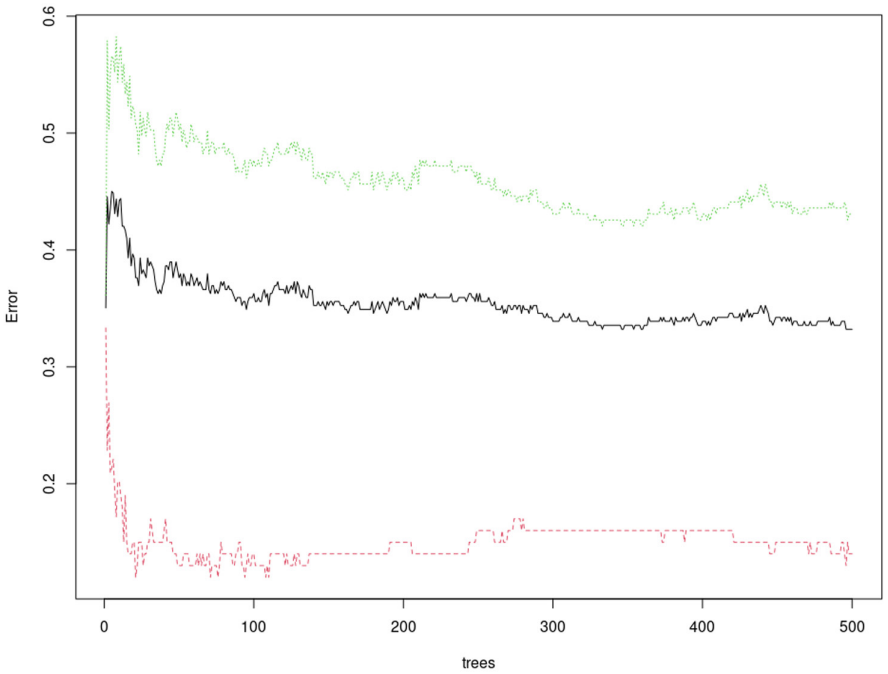

Figure 1. Cumulative error rate of final random forest model with respect to tree size. The green line indicates the error rate of predicting mechanical complications in patients without mechanical complications, and the red line indicates the error rate of predicting mechanical complications in patients with mechanical complications (Class errors) The class error of patients with mechanical complications was 0.14, whereas it was 0.43 for patients without mechanical complications for 500 trees.

Table 4. Performance Metrics of the Random Forest Model

Data subgroups / performance metrics	Training set (%)	Out-of-bag set (%)	Test set (%)
Accuracy	75	67	72
F-score	73	64	68
Sensitivity	99	86	91
Specificity	63	57	64
Precision	58	51	55
Negative predictive value	99	89	93
False positive rate	37	43	36
False discovery rate	42	49	45
False negative rate	1	14	9
Matthews correlation coefficient	59	41	51

Performance metrics are calculated with training, out-of-bag, and test sets. Each tree of the random forest model is built with approximately two-thirds of the training set, the other one-third of the data is the out-of-bag-set of specific to a particular tree. For each patient, a prediction is made without using the trees that include the patient in the training set.

Discussion

Using 295 individual health records as the training set, this work resulted in an RF model with 500 trees and a rule set of 8 rules. The RF model achieved an accuracy of 72% and a sensitivity of 91%, with a precision of 55%, while the rule set achieved an accuracy of 74%, a sensitivity of 81%, and a precision of 58%. Both models were evaluated on a holdout dataset consisting of 98 patient records. Although precision was low in both models, the primary objective was to achieve high sensitivity in identifying mechanically complicated patients, minimizing the risk of missed cases. Rather than targeting individual patients for intervention, the model helps recognize key risk factors, supports surgical planning, and optimizes postoperative care as part of a comprehensive prevention strategy. By using the class parameter as 2:1 and F-score as the performance metric, the study favors sensitivity in the sensitivity precision trade-off. Hence, using both models to administer preventive treatments may lead to unnecessary preventive treatments. That is why using expensive preventive treatments might not be possible.

Previously many other teams around the world tried to predict mechanical complications in ASD patients by applying machine learning algorithms. Yagi et al[20] and Scheer et al[21] achieved accuracy rates of 92% and 86% in predicting mechanical complications, respectively. By combining the 2 datasets, Pellise et al[22] developed risk stratification models for complications, hospital readmissions, and unplanned re-operations where the AUCs ranged from 67% to 92%.[22] Considering the works mentioned in this paragraph, this study's reported performance metrics were comparable with the literature.

Built-in feature importance graphs may favor categoric variables with high dimensions.[18] To be able to compare both RF and rule set feature importances and overcome the problem of the built-in function, permutation feature importances are calculated. Furthermore, many important features presented in Figure 4 are highly correlated, and this also affects the results.[23] Due to these reasons, before the feature importance analysis, the features were hierarchically clustered using the complete linkage method to form compact clusters where inside

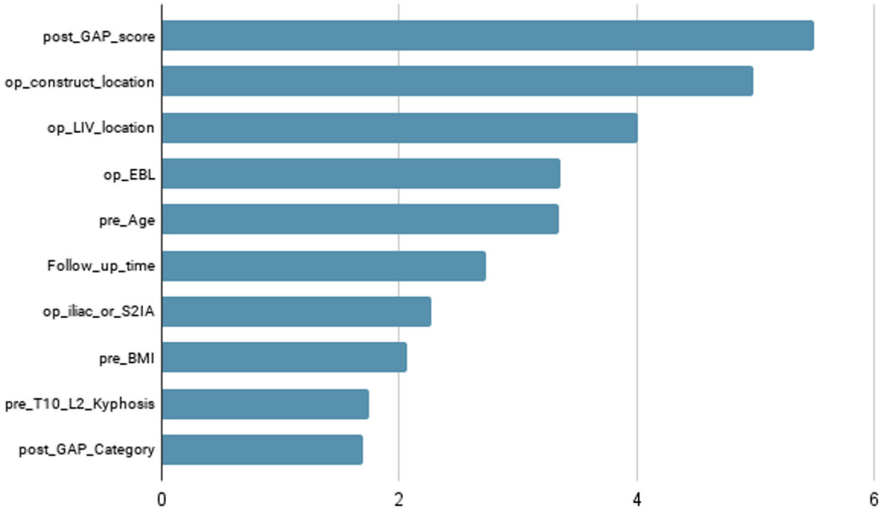

Figure 2. Feature importance graph using the built-in R function. Mean decrease Gini values averaged over all trees, and the first 10 features were presented. The encoding at the beginning of each feature name indicated that the feature is pre (preoperative), op (operative), or post (postoperative).

Table 5. The final rule set of simplified tree ensemble learner

Rule (Condition)*	Length of the rule	Prediction of mechanical complications	Frequency of the rule set in RF	Error
Estimated blood loss ≤ 550 mL and Osteotomy detailed ≤ 1.5 and pre ODI Personal care ≤ 3.5	3	Absent	0.11	0.00
Hospitalization time > 14.5 days and pre TL/L apical translation to CSVL ≤ 14.28°	2	Present	0.13	0.10
Pre T2 T5 kyphosis > 6.5° and post GAP score ≤ 4.5 and construct location ≤ 4.5	3	Absent	0.11	0.06
LIV location ≤ 0.5 and post TL/L Cobb > 19.735°	2	Present	0.21	0.10
LIV location ≤ 0.5 and dual versus multiple rods ≤ 0.5	2	Present	0.14	0.17
Follow-up time ≤ 57.2 months and delta Lumbar Lordosis > −16.5° and iliac fixation or not ≤ 0.5	3	Absent	0.11	0.23
Post GAP score > 1.5	1	Present	0.16	0.19
Else	1	Present	0.04	0.32

CSVL, central sacral vertical line; GAP, global alignment and proportion; LIV, lower instrumented vertebra; ODI, Oswestry Disability Index; RF, random forest; TL/L, thoracolumbar/lumbar.
*The final rule set derived with a simplified tree ensemble learner classifies patients iteratively. The first rule that is validated with the patient's data, determines the classification.

Table 6. Performance metrics of the rule set

Data subgroups	Training set (%)	Test set (%)
Accuracy	81	74
F-score	78	68
Sensitivity	95	81
Specificity	74	71
Precision	66	58
Negative predictive value	97	89
False positive rate	26	29
False discovery rate	34	42
False negative rate	5	19
Matthews correlation coefficient	66	49

Performance metrics are calculated with training and test sets consisting of 295, and 98 patients, respectively.

a cluster, absolute correlation between any 2 features was high. Data-driven feature grouping is preferred compared to leaving features as they are, as they might be misleading.[24]

The most effective features for predicting mechanical complications in the RF model were LIV location, postoperative GAP score, BMI, age, and preoperative apical translation to CSVL, in order of significance. LIV location was the most important predictor, in line with previous studies. The meta-analysis by Han and Ren (2022)[25] also emphasized LIV selection as a key factor in proximal junctional kyphosis, supporting the relevance of the findings.[26] Comparing important features of RF and the rule set with permutation feature importance, it was evident that by reducing the complexity, the rule set performance was dependent on only 4 to 6 features with different performance metrics. Besides LIV location and postoperative GAP score, the osteotomy, preoperative T2 T5 Kyphosis, and follow-up time features were found important for the rule set. On the other hand, all features that were used in the rule set were found important in the permutation importance analysis or a feature that is clustered was found important except for the dual versus multiple rods feature. Even though this feature was not found statistically significant to predict mechanical complications, it does help the predictive power of the rule set.

Lower instrumented vertebra location has been recurring among the most important 2 factors predicting mechanical complications in various other studies.[21,22,27,28] Postoperative GAP score has been evaluated by many other researchers after being proposed by Yilgor et al[1] and found important by the majority of researchers, given that they had an adequate dataset that had enough samples for each GAP categories.[22,27,29] Next, BMI was a prime candidate for modifying GAP score methodology to predict mechanical complications by many other studies and was found important.[29,30] Moreover, Yilgor et al[1] included age as a risk factor in the GAP score methodology, and it is found important in many other studies, including this study.[21,22,31]

The feature importance results of the rule set highlighted 3 different features than the ones explained in the previous paragraph: Osteotomy level and preoperative T2 T5 Kyphosis features, along with the follow-up time,[32] concluded that lower osteotomy level is associated with decreased reoperation rates, hence mechanical complications, after 3-column osteotomy. Last, the risk of a mechanical complication increases, as the follow-up time increases and it was shown by Gupta et al[33] by following patients up to 12 years. They reported that the GAP score's ability to predict mechanical complications increases from 68% to 91% as the duration of follow-up increases.

While the retrospective design of this study may have led to missing crucial information on mechanical complications, it is essential to acknowledge that such limitations are inherent in this type of research. The size of the cohort, although a constraint for experimenting with specific conditions, cannot diminish the value of this study, as the ESSG dataset is one of the largest and highest-quality ASD datasets. Despite its comprehensive nature, relying solely on data from 6 sites may introduce biases that could affect the generalization of the findings.

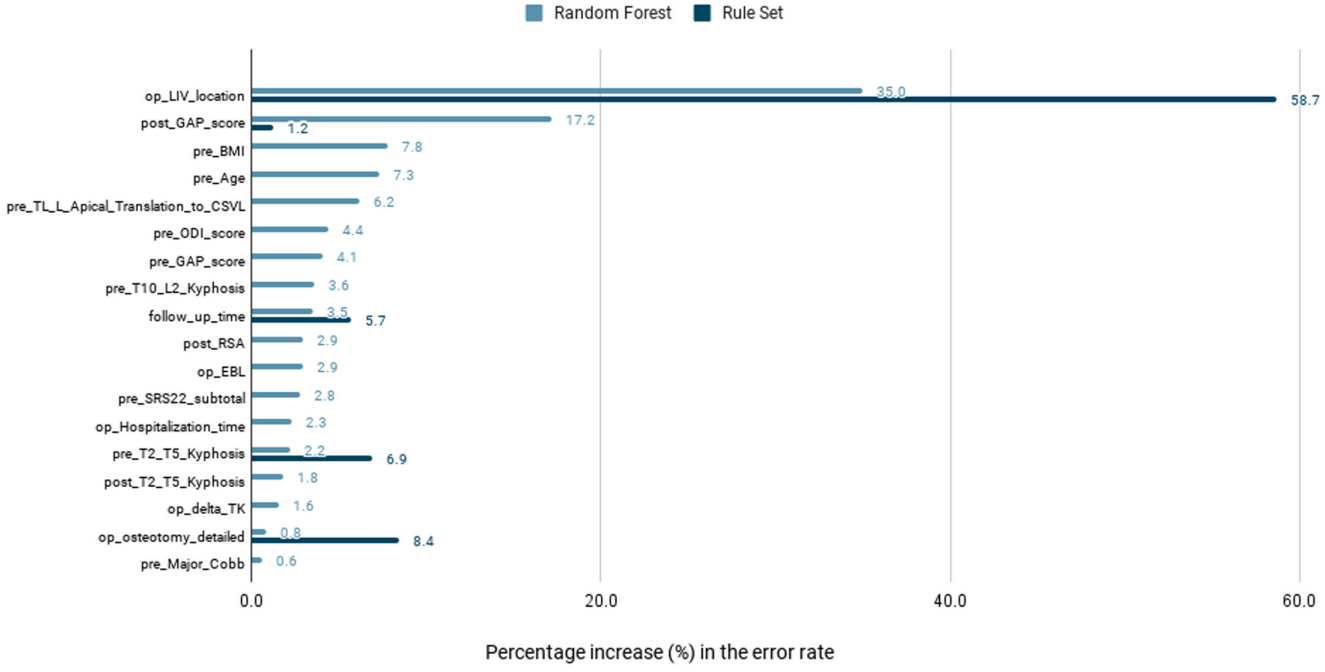

Figure 3. Permutation importance graph with respect to F-score. The features that increased the error rate after the permutation step are in the figure if the increased error rate is greater than 0.5%. The mean increases in the error rates of the F-score are presented. Recall that the feature names may not solely present a single feature.

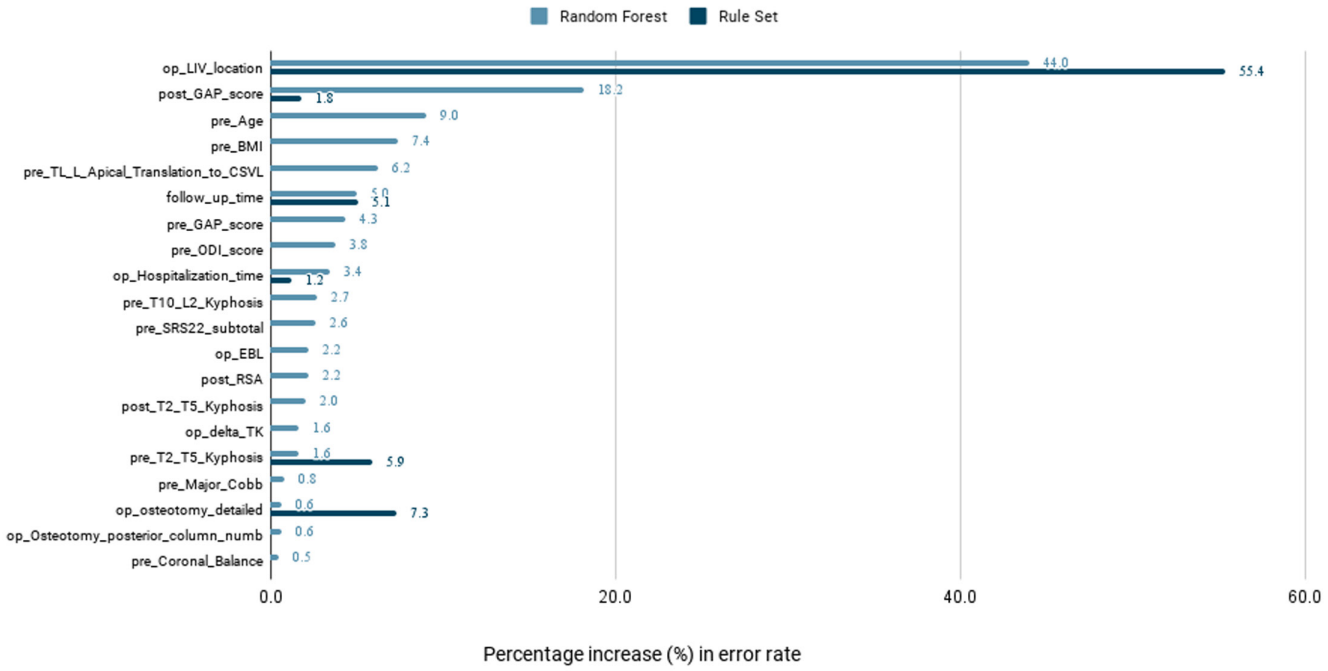

Figure 4. Permutation importance graph with respect to accuracy. The features that increased the error rate after the permutation step are in the figure if the increased error rate is greater than 0.5%. The mean increases in the error rates of accuracy are presented. Recall that the feature names may not solely present a single feature.

This study excludes severely disproportioned ASD patients according to the GAP score because of their remarkably high-risk factor of mechanical complications. This targeted approach not only provides valuable insights into operation planning strategies for proportioned and moderately disproportioned ASD patients but also opens up new avenues for specific preventive post-surgery care after operations. Having an adequate dataset even though excluding 1 subgroup is one of the key strengths of the study.

Preventive measures can be taken at different stages to minimize mechanical complications. Preoperatively, modifiable factors like BMI can be managed, while intraoperative planning, considering key factors like LIV selection, helps reduce risks. Postoperative follow-ups provide insight into patient progress, allowing timely interventions such as targeted rehabilitation. The prediction model aids in this process by identifying risk factors, guiding decision-making, and enhancing prevention strategies for better surgical outcomes.

During the cross-validation process, the focus on sensitivity measures emerged as paramount in ensuring high-risk patients would be detected. By closely examining these measures, the proposed RF model and rule set show their potential to be used in clinics. The

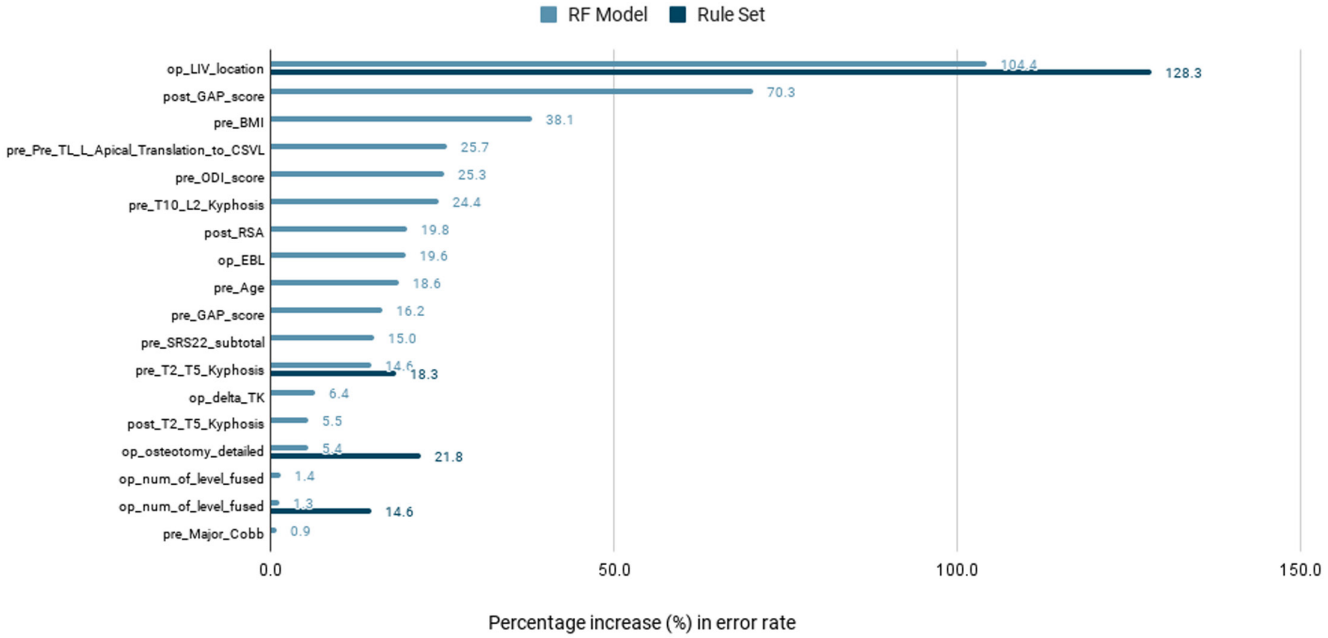

Figure 5. Permutation importance graph of the random forest model with respect to sensitivity. The features that increased the error rate after the permutation step are in the figure if the increased error rate is greater than 0.5%. The mean increases in the error rates of sensitivity are presented. Recall that the feature names may not solely present a single feature.

complexity of the RF model enabled the predictions to have higher sensitivity (91%) in the test set. The rule set achieved 2% higher accuracy than the model but had 10% lower sensitivity. However, it was designed for clinical usability, emphasizing simplicity over performance. While the model demonstrated higher sensitivity, its complexity may hinder practical implementation. Future work for mechanical complication prediction in ASD patients would benefit from leveraging larger datasets from a wider range of sites, as it would open up new possibilities for uncovering hidden patterns and trends that might otherwise go unnoticed.

Euorpean Spine Study Group: A. Gómez, A. Yucekul, AF. Mannion, A. Bourghli, A. Matamalas, Ç. Yilgor, D. Haschtmann, D. Jeszensky, JM. Vital, J. Bagó, JM. Sánchez, L. Boissiere, M. Loibl, N. Fernández-Baíllo, S. Haddad, S. Núñez, T. Fekete, Y. Yavuz.

Data Availability Statement: Clinical data was collected by the authors of Global Alignment and Proportion (GAP) Score in 6 different sites across Europe and the data is unavailable to access.

Ethics Committee Approval: Institutional review board approval was obtained for patient enrollment and data collection protocols at each site. This study was approved by the Acıbadem University Medical Research Evaluation Board (Acıbadem Üniversitesi Tıbbi Araştırmalar Değerlendirme Kurulu – ATADEK) with the approval reference number 2019-11/34. This study took place at Acıbadem University and ATADEK is the medical board of Acıbadem University.

Informed Consent: Written informed consent was obtained from the patients who agreed to take part in the study.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – B.B., N.D., C.Y., A.Y., T.Z., Sl.Ha., Sh.Ha., F.K., I.O., J.P., F.K., F.J.S.P., F.P., A.A., C.B., O.U.S. Design – B.B., N.D., C.Y., A.Y., T.Z., Sl.Ha., Sh.Ha., F.K., I.O., J.P., F.K., F.J.S.P., F.P., A.A., C.B., O.U.S. Supervision – C.Y., A.A., C.B., O.U.S. Resources – B.B., N.D., C.Y., A.Y., T.Z., Sl.Ha., Sh.Ha., F.K., I.O., J.P., F.K., F.J.S.P., F.P., A.A., C.B., O.U.S. Materials – B.B., N.D., C.Y., A.Y., T.Z., Sl.Ha., Sh.Ha., F.K., I.O., J.P., F.K., F.J.S.P., F.P., A.A., C.B., O.U.S. Data Collection and/or Processing – B.B., N.D., C.Y., A.Y., T.Z., Sl.Ha., Sh.Ha., F.K., I.O., J.P., F.K., F.J.S.P., F.P., A.A. Analysis and/or Interpretation – B.B., N.D., C.Y. Literature Search – B.B., N.D., C.Y. Writing – B.B., N.D., C.Y. Critical Review – C.Y., A.A., C.B., O.U.S.

Acknowledgments: Clinical data was collected by the authors of Global Alignment and Proportion (GAP) Score in 6 different sites across Europe and the data is unavailable to access. Barış Balaban is a bursar of the Scientific and Technological Research Council of Türkiye (TUBITAK) 2244 – Industry Doctorate Program in conjunction with Bilmed Computer and Software Company. Prof. Dr. Ugur Sezerman supervised this work as the department head of the Bioinformatics and Biostatistics Department at Acibadem University, Türkiye.

Declaration of Interests: Baris Balaban declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Nuri Demirci declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Caglar Yilgor reports a relationship with Medtronic that includes consulting. Altug Yucekul declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Tais Zulemyan declares that she has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Sleiman Haddad reports a relationship with Medtronic that includes consulting and speaking and lecture fees. Shahnawaz Haleem declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Feyzi Kilic declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Ibrahim Obeid reports a relationship with Alpatec Spine that includes royalties. Ibrahim Obeid reports a relationship with Medicrea that includes royalties. Ibrahim Obeid reports a relationship with Spineart that includes royalties. Ibrahim Obeid reports a relationship with DePuy Synthes that includes consulting and funding grants. Ibrahim Obeid reports a relationship with Medtronic that includes consulting. Javier Pizones reports a relationship with Stryker that includes consulting. Javier Pizones reports a relationship with Medtronic that includes consulting and funding grants. Javier Pizones reports a relationship with DePuy Synthes that includes funding grants. Frank Kleinstueck reports a relationship with DePuy Synthes that includes speaking and lecture fees. Francisco Javier Sanchez Perez-Grueso declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Ferran Pellise reports a relationship with Medtronic that includes consulting and funding grants. Ferran Pellise reports a relationship with Stryker Spine that includes consulting. Ferran Pellise reports a relationship with DePuy Synthes that includes funding grants. Ahmet Alanay reports a relationship with Medtronic that includes funding grants. Ahmet Alanay reports a relationship with DePuy Synthes that includes funding grants. Ahmet Alanay reports a relationship with Globus Medical Inc that includes consulting or advisory. Ahmet Alanay reports a relationship with Highridge Medical that includes royalties and consulting. Cetin Bagci declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Osman Ugur Sezerman declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The ESSG reports a relationship with Medtronic that includes funding grants. The ESSG reports a relationship with DePuy Synthes that includes funding grants.

Funding: The authors declared that this study has received no financial support.

References

1. Yilgor C, Sogunmez N, Yavuz Y, et al. Global Alignment and Proportion (GAP) Score: development and validation of a new method of analyzing spinopelvic alignment to predict mechanical complications after adult spinal deformity surgery. *Spine J*. 2017;17(10):S155-S156. [CrossRef]
2. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? *arXiv (Cornell University)*. Published online January 1, 2022. [CrossRef]
3. Rokach L, Maimon O. Decision trees. In: Maimon O, Rokach L, eds. *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag; 2005:165-192. [CrossRef]
4. Quinlan J. Bagging, boosting, and C4.5. In: *Proceedings of the Thirteenth American Association for Artificial Intelligence National Conference on Artificial Intelligence*. Menlo Park: AAAI Press; 1996:725-730.
5. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. [CrossRef]
6. Hu L, Li L. Using Tree-Based Machine Learning for Health Studies: literature review and Case series. *Int J Environ Res Public Health*. 2022;19(23):16080. [CrossRef]
7. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123-140. [CrossRef]
8. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 2024. Available at: https://www.R-project.org/.
9. Wickham H, François R, Henry L, et al. dplyr: a Grammar of Data Manipulation_. R package version 1.1.2, 2023. Available at: https://CRAN.R-project.org/package=dplyr.
10. Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Softw*. 2008;28(5):1-26. [CrossRef]
11. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. Published online January 1, 2007. Available at: http://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf.
12. Friedman JH, Popescu BE. Predictive learning via rule ensembles. *Ann Appl Stat*. 2008;2(3). [CrossRef]
13. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. *Knowledge Discovery and Data Mining*. Published online August 27, 1998:80-86. Available at: https://sci2s.ugr.es/keel/pdf/algorithm/congreso/1998-Liu-CBA.pdf.
14. Deng H, Runger G, Tuv E, Bannister W. CBC: an associative classifier with a small number of rules. *Decis Support Syst*. 2014;59:163-170. [CrossRef]
15. Deng H. Interpreting tree ensembles with inTrees. *Int J Data Sci Anal*. 2019;7(4):277-287. [CrossRef]
16. Deng H, Runger G. Gene selection with guided regularized random forest. *Pattern Recognit*. 2013;46(12):3483-3489. [CrossRef]
17. Probst P, Boulesteix AL. To tune or not to tune the number of trees in random forest. *J Mach Learn Res*. 2017;18(181):1-18. Available at: https://jmlr.org/papers/volume18/17-269/17-269.pdf.
18. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics*. 2010;26(10):1340-1347. [CrossRef]
19. Müllner D. fastcluster: fast Hierarchical, Agglomerative Clustering Routines for R and Python. *J Stat Softw*. 2013;53(9):1-18. Available at: http://www.jstatsoft.org/v53/i09/.
20. Yagi M, Hosogane N, Fujita N, et al. Predictive model for major complications 2 years after corrective spine surgery for adult spinal deformity. *Eur Spine J*. 2019;28(1):180-187. [CrossRef]
21. Scheer JK, Osorio JA, Smith JS, et al. Development of validated computer-based preoperative Predictive Model for Proximal junction Failure (PJF) or Clinically Significant PJK with 86% accuracy based on 510 ASD patients with 2-year follow-up. *Spine*. 2016;41(22):E1328-E1335. [CrossRef]
22. Pellisé F, Serra-Burriel M, Smith JS, et al. Development and validation of risk stratification models for adult spinal deformity surgery. *J Neurosurg Spine*. 2019;31(4):587-599. [CrossRef]
23. Tolosi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*. 2011;27(14):1986-1994. [CrossRef]
24. Gregorutti B, Michel B, Saint-Pierre P. Grouped variable importance with random forests and application to multiple functional data analysis. *Comp Stat Data Anal*. 2015;90:15-35. [CrossRef]
25. Han X, Ren J. Risk factors for proximal junctional kyphosis in adult spinal deformity after correction surgery: A systematic review and meta-analysis. *Acta Orthopaedica et Traumatologica Turcica*. 2022;56(3):158–165. [CrossRef]
26. Kwan KYH, Lenke LG, Shaffrey CI, et al. Risk factors for proximal junctional kyphosis in adult spinal deformity: a systematic review and metaanalysis. *Eur Spine J*. 2022;31(12):2971-2984. [CrossRef]
27. Pellisé F, Serra-Burriel M, Vila-Casademunt A, et al. Quality metrics in adult spinal deformity surgery over the last decade: a combined analysis of the

largest prospective multicenter data sets. *J Neurosurg Spine*. 2021;1:1-19. [CrossRef]
28. Balaban B, Yilgor C, Yucekul A, et al. Building clinically actionable models for predicting mechanical complications in postoperatively well-aligned adult spinal deformity patients using XGBoost algorithm. *Inform Med Unlocked*. 2023;37:101191. [CrossRef]
29. Hiltunen S, Repo JP, Pekkanen L, Kautiainen H, Kyrölä K. Mechanical complications and reoperations after adult spinal deformity surgery: a clinical analysis with the gap score. *Eur Spine J*. 2023;32(4):1421-1428. [CrossRef]
30. Noh SH, Ha Y, Obeid I, et al. Modified global alignment and proportion scoring with body mass index and Bone Mineral Density (GAPB) for improving predictions of mechanical complications after adult spinal deformity surgery. *Spine J*. 2020;20(5):776-784. [CrossRef]
31. Barone G, Giudici F, Martinelli N, et al. Mechanical complications in adult spine deformity surgery: retrospective evaluation of incidence, clinical impact and risk factors in a single-center large series. *J Clin Med*. 2021;10(9):1811. [CrossRef]
32. Kawabata A, Sakai K, Yamada K, et al. The lower osteotomy level is associated with decreased revision surgery due to mechanical complications after three-column osteotomy in patients with adult spinal deformity: a multi-institutional retrospective study. *Glob Spine J*. Published online August 18. 2025;15(2):498-505. [CrossRef]
33. Gupta MC, Yilgor C, Moon HJ, et al. Evaluation of global alignment and proportion score in an independent database. *Spine J*. 2021;21(9):1549-1558. [CrossRef]

Supplementary Table 1. Descriptive Statistics of Other 30 Features

Features	Mechanical Complications - Present (n=100)	Mechanical Complications - Absent (n=195)	Total (n=295)	p-value
Pre Operation Construct Rigidity Soft Landing				0.72
Mean (SD)	2.8 (0.8)	2.9 (0.7)	2.9 (0.7)	
Median (Q1, Q3)	3.0 (2.0, 3.0)	3.0 (3.0, 3.0)	3.0 (3.0, 3.0)	
Min - Max	1.0 - 4.0	1.0 - 4.0	1.0 - 4.0	
Missing	0	0	0	
Pre Operation Gender (1 F, 2 M)				0.47
Mean (SD)	1.2 (0.4)	45.2 (18.9)	49.6 (18.8)	
Median (Q1, Q3)	1 (1, 1)	1 (1, 1)	54.1 (1, 1)	
Min - Max	1 - 2	1 - 2	1 - 2	
Missing	0	0	0	
Pre Operation Height (m^2)				0.33
Mean (SD)	1.6 (0.1)	1.6 (0.1)	1.6 (0.1)	
Median (Q1, Q3)	1.6 (1.6, 1.7)	1.6 (1.6, 1.7)	1.6 (1.6, 1.7)	
Min - Max	1.4 - 1.9	1.4 - 1.9	1.4 - 1.9	
Missing	0	0	0	
Pre Operation Level of Physical Labor				0.002
Mean (SD)	3.2 (1.7)	3.7 (1.5)	3.5 (1.6)	
Median (Q1, Q3)	4.0 (2.0, 4.0)	4.0 (4.0, 5.0)	4.0 (3.0, 5.0)	
Min - Max	0.0 - 5.0	0.0 - 5.0	0.0 - 5.0	
Missing	0	0	0	
Pre Operation Menapause (-1:Male)				0.038
Mean (SD)	0.4 (0.8)	0.2 (0.7)	0.3 (0.8)	
Median (Q1, Q3)	1.0 (0.0, 1.0)	0.0 (0.0, 1.0)	0.0 (0.0, 1.0)	
Min - Max	-1.0 - 1.0	-1.0 - 2.0	-1.0 - 2.0	
Missing	0	0	0	
Pre Operation Number of Pregnancies				0.07
Mean (SD)	2.3 (1.8)	2.0 (1.8)	2.1 (1.8)	
Median (Q1, Q3)	2.0 (1.0, 4.0)	2.0 (0.0, 3.0)	2.0 (0.0, 3.0)	
Min - Max	0.0 - 5.0	0.0 - 5.0	0.0 - 5.0	
Missing	0	0	0	
Pre Operation Pelvic Incidence (°)				0.065
Mean (SD)	56.0 (13.4)	52.9 (11.6)	54.0 (12.3)	
Median (Q1, Q3)	54.5 (46.0, 64.0)	52.0 (44.5, 60.0)	53.0 (45.0, 62.0)	
Min - Max	24.0 - 86.0	28.0 - 87.0	24.0 - 87.0	
Missing	0	0	0	
Pre Operation Pelvic Obliquity (°)				0.29
Mean (SD)	2.9 (0.7)	3.6 (3.9)	3.3 (3.5)	
Median (Q1, Q3)	2.0 (2.0, 3.0)	3.0 (1.0, 4.0)	2.0 (1.0, 4.0)	
Min - Max	0.0 - 18.0	0.0 - 29.0	0.0 - 29.0	
Missing	2	10	12	
Pre Operation Prior Decompression				0.14
Mean (SD)	0.1 (0.3)	0.1 (0.2)	0.1 (0.3)	
Median (Q1, Q3)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	
Min - Max	0.0 - 1.0	0.0 - 1.0	0.0 - 1.0	
Missing	0	0	0	
Pre Operation Prior Fusion				0.001
Mean (SD)	0.3 (0.5)	0.2 (0.4)	0.2 (0.4)	
Median (Q1, Q3)	0.0 (0.0, 1.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	
Min - Max	0.0 - 1.0	0.0 - 1.0	0.0 - 1.0	
Missing	0	0	0	
Pre Operation Smoking				0.28
Mean (SD)	0.6 (0.8)	0.5 (0.8)	0.5 (0.8)	
Median (Q1, Q3)	0.0 (0.0, 1.0)	0.0 (0.0, 1.0)	0.0 (0.0, 1.0)	
Min - Max	0.0 - 2.0	0.0 - 2.0	0.0 - 2.0	
Missing	0	0	0	

(*Continued*)

Supplementary Table 1. Descriptive Statistics of Other 30 Features (*Continued*)

Features	Mechanical Complications - Present (n=100)	Mechanical Complications - Absent (n=195)	Total (n=295)	p-value
Pre Operation Spine Problems (years)				0.11
Mean (SD)	4.5 (1.0)	4.3 (1.0)	4.4 (1.0)	
Median (Q1, Q3)	5.0 (4.0, 5.0)	5.0 (4.0, 5.0)	5.0 (4.0, 5.0)	
Min - Max	1.0 - 5.0	0.0 - 5.0	0.0 - 5.0	
Missing	0	0	0	
Pre Operation T5 T12 Kyphosis (°)				0.70
Mean (SD)	35.1 (20.8)	33.8 (20.0)	34.3 (20.2)	
Median (Q1, Q3)	33.0 (19.0, 48.5)	30.0 (20.0, 45.0)	31.0 (19.5, 45.0)	
Min - Max	1.0 - 88.0	-2.0 - 94.0	-2.0 - 94.0	
Missing	0	0	0	
Pre Operation Weight (kg)				< 0.001
Mean (SD)	68.3 (11.5)	63.4 (12.7)	65.0 (12.5)	
Median (Q1, Q3)	67.0 (58.8, 77.2)	62.0 (54.0, 71.5)	64.0 (56.0, 72.0)	
Min - Max	45.0 - 97.0	39.0 - 110.0	39.0 - 110.0	
Missing	0	0	0	
Operation Biggest Osteotomy Type				< 0.001
Mean (SD)	1.0 (0.9)	0.6 (0.8)	0.8 (0.9)	
Median (Q1, Q3)	1.0 (0.0, 2.0)	0.0 (0.0, 1.0)	1.0 (0.0, 1.0)	
Min - Max	0.0 - 3.0	0.0 - 3.0	0.0 - 3.0	
Missing	0	0	0	
Operation Decompression				0.048
Mean (SD)	0.2 (0.4)	0.1 (0.3)	0.2 (0.4)	
Median (Q1, Q3)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	
Min - Max	0.0 - 1.0	0.0 - 1.0	0.0 - 1.0	
Missing	0	0	0	
Operation Decompression Level				0.032
Mean (SD)	0.5 (0.1)	0.2 (0.6)	0.3 (0.8)	
Median (Q1, Q3)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	
Min - Max	0.0 - 4.0	0.0 - 3.0	0.0 - 4.0	
Missing	0	0	0	
Operation Domino Technique				0.002
Mean (SD)	0.4 (0.5)	0.2 (0.4)	0.3 (0.4)	
Median (Q1, Q3)	0.0 (0.0, 1.0)	0.0 (0.0, 1.0)	0.0 (0.0, 0.5)	
Min - Max	0.0 - 1.0	0.0 - 1.0	0.0 - 1.0	
Missing	0	0	0	
Operation Interbody Fusion				0.003
Mean (SD)	0.5 (0.5)	0.3 (0.5)	0.4 (0.5)	
Median (Q1, Q3)	1.0 (0.0, 1.0)	0.0 (0.0, 1.0)	0.0 (0.0, 1.0)	
Min - Max	0.0 - 1.0	0.0 - 1.0	0.0 - 1.0	
Missing	0	0	0	
Operation Number of Total Rods				0.30
Mean (SD)	2.3 (0.6)	2.2 (0.6)	2.3 (0.6)	
Median (Q1, Q3)	2.0 (2.0, 2.0)	2.0 (2.0, 2.0)	2.0 (2.0, 2.0)	
Min - Max	2.0 - 4.0	2.0 - 4.0	2.0 - 4.0	
Missing	0	0	0	
Operation Number of Transverse Connectors				0.003
Mean (SD)	0.6 (0.9)	0.3 (0.7)	0.4 (0.8)	
Median (Q1, Q3)	0.0 (0.0, 1.0)	0.0 (0.0, 0.0)	0.0 (0.0, 1.0)	
Min - Max	0.0 - 4.0	0.0 - 4.0	0.0 - 4.0	
Missing	0	0	0	
Operation Surgery Stages				< 0.001
Mean (SD)	1.1 (0.4)	1.0 (0.2)	1.1 (0.3)	
Median (Q1, Q3)	1.0 (1.0, 1.0)	1.0 (1.0, 1.0)	1.0 (1.0, 1.0)	
Min - Max	1.0 - 2.0	1.0 - 3.0	1.0 - 3.0	
Missing	0	0	0	

(*Continued*)

Supplementary Table 1. Descriptive Statistics of Other 30 Features (*Continued*)

Features	Mechanical Complications - Present (n=100)	Mechanical Complications - Absent (n=195)	Total (n=295)	p-value
Operation Surgery Time (minutes)				0.30
Mean (SD)	332 (182.1)	310.2 (179.9)	317.6 (180.6)	
Median (Q1, Q3)	300 (207.5, 413.2)	285 (180, 386.5)	285 (187.5, 390)	
Min - Max	120.0 – 1115.0	60.0 - 1265.0	60.0 - 1265.0	
Missing	0	0	0	
Operation Upper Instrumented Vertebra implant				0.83
Mean (SD)	1.6 (0.9)	1.5 (0.9)	1.6 (0.9)	
Median (Q1, Q3)	1.0 (1.0, 3.0)	1.0 (1.0, 2.0)	1.0 (1.0, 3.0)	
Min - Max	1.0 - 4.0	0.0 - 4.0	0.0 - 4.0	
Missing	0	0	0	
Post Operation Coronal Balance (°)				0.064
Mean (SD)	16.2 (13.2)	18.3 (12.9)	17.6 (13.0)	
Median (Q1, Q3)	11.9 (5.4, 25.5)	16.1 (9.5, 26.2)	14.5 (7.3, 25.9)	
Min - Max	1.0 - 55.9	0.0 - 69.0	0.0 - 69.0	
Missing	3	13	16	
Post Operation Major Cobb (°)				0.22
Mean (SD)	19.8 (16.5)	21.6 (14.5)	21.0 (15.2)	
Median (Q1, Q3)	18.5 (6.0, 27.2)	18.0 (12.0, 31.0)	18.0 (10.0, 30.3)	
Min - Max	0.0 - 85.0	0.0 - 65.0	0.0 - 85.0	
Missing	4	18	22	
Post Operation Pelvic Obliquity (°)				0.41
Mean (SD)	2.2 (2.1)	2.2 (2.4)	2.2 (2.3)	
Median (Q1, Q3)	1.9 (1.0, 2.7)	2.0 (1.0, 3.0)	2.0 (1.0, 3.0)	
Min – Max	-1.2 - 14.4	-8.9 - 13.8	-8.9 - 14.4	
Missing	7	22	29	
Post Operation T10 L2 Kyphosis (°)				0.011
Mean (SD)	6.7 (12.6)	3.4 (11.2)	4.5 (11.8)	
Median (Q1, Q3)	6.0 (0.8, 12.2)	3.0 (-3.0, 10.0)	4.0 (-2.0, 11.0)	
Min - Max	-47.0 - 43.0	-35.0 - 39.0	-47.0 - 43.0	
Missing	0	0	0	
Post Operation T5 T12 Kyphosis (°)				0.033
Mean (SD)	38.2 (13.8)	34.5 (14.1)	35.8 (14.1)	
Median (Q1, Q3)	38.0 (28.8, 46.0)	34.0 (25.0, 44.0)	35.0 (26.0, 44.5)	
Min - Max	2.0 - 80.0	-23.0 - 76.0	-23.0 - 80.0	
Missing	0	0	0	
Post Operation TL L Apical Translation to CSVL (°)				0.24
Mean (SD)	16.8 (13.9)	18.4 (13.2)	17.8 (13.4)	
Median (Q1, Q3)	15.0 (5.2, 23.4)	15.0 (8.9, 25.8)	15.0 (7.2, 25.6)	
Min - Max	0.4 - 63.2	0.3 - 62.6	0.3 - 63.2	
Missing	16	35	51	
Min - Max	2.0 - 80.0	-23.0 - 76.0	-23.0 - 80.0	

Supplementary Table 2. Mean (Standard Deviations) of Cross-validation Results

Grid Search	Tuned Parameters Set*	F-score	AUC	Sensitivity	Specificity	AUPRC	Precision
1	10, 25, 1, null	0.51 (0.08)	0.72 (0.05)	0.45 (0.09)	0.84 (0.05)	0.53 (0.06)	0.60 (0.09)
2	250, 50, 25, 8	0.62 (0.03)	0.78 (0.04)	0.80 (0.06)	0.61 (0.06)	0.61(0.06)	0.52 (0.03)
3	500, 55, 19, 6	0.63 (0.03)	0.79 (0.04)	0.84 (0.07)	0.59 (0.05)	0.63 (0.06)	0.51 (0.02)

AUC: area under the receiving-operating characteristic curve, AUPRC: area under the precision-recall curve
*RF parameters are presented in this column which are the number of trees in the forest, mtry, nodesize and, maxnodes. In the first grid search, default values of nodesize (1) and maxnodes (Null- trees are fully grown) are used.

Supplementary Table 3. Frequent Patterns of Final Random Forest Model

Condition	Support	Confidence	Prediction of Mechanical Complications
Construct location ≤ 4.5	0.082	0.626	Absent
Construct location > 4.5	0.055	0.735	Present
LIV location ≤ 0.5	0.047	0.622	Present
LIV location > 0.5	0.041	0.693	Absent
Iliac fixation or not ≤ 0.5	0.039	0.628	Absent
Post GAP score ≤ 2.5	0.031	0.614	Absent
Post GAP score category ≤ 0.5	0.031	0.662	Absent
Iliac fixation or not > 0.5	0.024	0.643	Present
Post GAP score ≤ 1.5	0.023	0.742	Absent
Post GAP score category > 0.5	0.021	0.614	Present
Post GAP score > 1.5	0.020	0.741	Present
Post GAP score ≤ 3.5	0.020	0.766	Absent
Sacro iliac fixation ≤ 0.5	0.019	0.637	Absent
Post GAP score > 2.5	0.019	0.582	Present
Post GAP score ≤ 4.5	0.017	0.825	Absent
Surgery year ≤ 2015	0.017	0.754	Present
Post GAP score > 4.5	0.014	0.811	Present
Post GAP score > 3.5	0.014	0.677	Present
Surgery year > 2015	0.013	0.784	Absent
Pre T10 L2 Kyphosis > 1.5°	0.013	0.696	Present
Sacro iliac fixation > 0.5	0.013	0.609	Present
Pre T10 L2 Kyphosis ≤ 1.5°	0.013	0.679	Absent
Pre SRS-22 Pain ≤ 3.55	0.012	0.607	Present
Estimated Blood Loss > 550mL	0.011	0.532	Present
Number of levels fused > 7.5	0.010	0.861	Present
Pre SRS-22 Pain > 3.55	0.010	0.775	Absent
Estimated Blood Loss ≤ 550mL	0.010	0.592	Absent

LIV: Lower Instrumented Vertebra, GAP: Global Alignment and Proportion, SRS-22: Scoliosis Research Society-22 Patient Questionnaire
Frequent patterns were derived using the default values of the parameters of "getFreqPattern" of inTrees library. The support values sorted the table and there weren't any frequent patterns with more than 0.01 support and with 2 or more features.

Supplementary Table 4. Significant Clustered Features

Feature	Clustered Features
Op LIV location	Op sacroiliac fixation, op construct location, op iliac fixation or not
Post GAP score	Post GAP category, post RPV
BMI	Weight
Age	Body Mass Density features (Pre BMD 3 cat, pre BMD young normal)
Pre TL/L Apical Translation to CSVL	Pre TL/L Cobb, post TL/L Cobb, post TL/L Apical Translation to CSVL
Pre ODI score	Pre ODI walking and standing scores
Pre GAP score	Pre GAP category, op delta LL, pre RSA, pre RLL, pre RPV
Follow-up time	Year of the surgery
Op EBL	Op Intraoperative transfusions
Pre SRS-22 Subtotal	Pre NRS back, pre ODI Pain Intensity, pre SF-36 PCS, pre SRS-22 Function, pre SRS-22 Pain
Op osteotomy detailed	Op osteotomy none or posterior or anterior, Op biggest Osteotomy type, op 3-column osteotomy location simple, op 3-column osteotomy location
Pre Major Cobb	Pre MT Cobb, post Major Cobb, post MT Cobb
Op Number of levels fused	Op UIV location anatomical
Op Osteotomy posterior column number	Op Osteotomy posterior column number cat1, op Osteotomy posterior column number cat2

op: Operative, LIV: Lower Instrumented Vertebra, post: postoperative, GAP: Global Alignment and Proportion, RPV: Relative Pelvic Version, BMI: Body Mass Index, pre: preoperative, BMD: Bone Mineral Density, TL/L: Thoracolumbar/Lumbar , CSVL: Central Sacral Vertical Line, RSA: Relative Spinopelvic Alignment, RLL: Relative Lumbar Lordosis, EBL: Estimated Blood Loss, SRS-22: Scoliosis Research Society-22 Patient Questionnaire, SF-36: Short Form-36, PCS: Physical Component Summary, MT: Major Thoracic, UIV: Upper Instrumented Vertebra, cat: category.

Supplementary Table 5. Permutation Feature Importance Results for the Training Set - Increase in Error Rates

	F-score		Accuracy		Sensitivity	
Feature	**RF**	**Ruleset**	**RF**	**Ruleset**	**RF**	**Ruleset**
op_sacro_iliac_fixation	57.7%	129.5%	58.2%	112.6%	1814.9%	756.9%
post_GAP_score	36.9%	46.3%	29.9%	39.3%	1497.1%	0.0%
pre_BMI	25.8%	0.0%	17.9%	0.0%	1194.6%	0.0%
pre_Age	25.3%	0.0%	19.7%	0.0%	1009.9%	0.0%
pre_TL_L_Cobb	24.1%	42.9%	17.1%	34.4%	1080.3%	0.0%
pre_ODI_score	22.0%	0.0%	14.1%	0.0%	1080.3%	0.0%
pre_GAP_score	21.6%	43.7%	14.6%	36.3%	992.0%	265.4%
pre_T10_L2_Kyphosis	20.9%	0.0%	12.8%	0.0%	1063.6%	0.0%
follow_up_time	20.8%	53.1%	15.3%	44.0%	850.5%	331.1%
op_EBL	20.2%	0.0%	12.3%	0.0%	1021.5%	0.0%
post_RSA	20.1%	0.0%	12.2%	0.0%	1020.7%	0.0%
pre_SRS22_subtotal	20.1%	0.0%	12.7%	0.0%	978.0%	0.0%
op_Hospitalization_time	19.5%	44.4%	13.6%	38.4%	0.0%	252.3%
pre_T2_T5_Kyphosis	19.2%	54.5%	11.6%	44.9%	971.7%	343.7%
post_T2_T5_Kyphosis	18.8%	0.0%	12.0%	0.0%	890.9%	0.0%
op_delta_TK	18.6%	0.0%	11.6%	0.0%	899.4%	0.0%
op_biggest_Osteotomy_type	17.7%	56.8%	10.5%	46.8%	889.6%	356.7%
pre_Major_Cobb	17.5%	0.0%	10.8%	0.0%	846.8%	0.0%
op_Osteotomy_posterior_column_number	17.2%	0.0%	10.4%	0.0%	839.7%	0.0%
pre_Coronal_Balance	17.2%	0.0%	10.4%	0.0%	0.0%	0.0%
op_num_of_level_fused	17.0%	0.0%	10.0%	0.0%	850.8%	0.0%
pre_T5_T12_Kyphosis	16.9%	0.0%	10.0%	0.0%	0.0%	0.0%
pre_Diagnosis	16.9%	0.0%	10.0%	0.0%	0.0%	0.0%
pre_Leg_Length_Discrepancy	16.8%	0.0%	9.9%	0.0%	0.0%	0.0%
pre_Number_prior_Spine_Surgery	16.8%	0.0%	9.9%	0.0%	0.0%	0.0%
pre_Pre_ODI_Personal_Care	0.0%	44.5%	0.0%	37.2%	0.0%	269.1%

RF: Random Forest, op: operative, post: postoperative, pre: preoperative, GAP: Global Alignment and Proportion, BMI: Body Mass Index, TL_L: Thoracolumbar/Lumbar, ODI: Oswestry Disability Index, EBL: Estimated Blood Loss, RSA: Relative Spinopelvic Alignment, SRS-22: Scoliosis Research Society-22 Patient Questionnaire, TK: Thoracic Kyphosis.