



# Developing a new tool for scoliosis screening in a tertiary specialistic setting using artificial intelligence: a retrospective study on 10,813 patients: 2023 SOSORT award winner

Francesco Negrini<sup>1,2</sup> · Andrea Cina<sup>3,4</sup> · Irene Ferrario<sup>5</sup> · Fabio Zaina<sup>5</sup> · Sabrina Donzelli<sup>5</sup> · Fabio Galbusera<sup>3</sup> · Stefano Negrini<sup>6,7</sup>

Received: 1 August 2023 / Revised: 1 August 2023 / Accepted: 6 August 2023  
© The Author(s) 2023

## Abstract

**Purpose** The study aims to assess if the angle of trunk rotation (ATR) in combination with other readily measurable clinical parameters allows for effective non-invasive scoliosis screening.

**Methods** We analysed 10,813 patients (4–18 years old) who underwent clinical and radiological evaluation for scoliosis in a tertiary clinic specialised in spinal deformities. We considered as predictors ATR, Prominence (mm), visible asymmetry of the waist, scapulae and shoulders, familiarity, sex, BMI, age, menarche, and localisation of the curve. We implemented a Logistic Regression model to classify the Cobb angle of the major curve according to thresholds of 15, 20, 25, 30, and 40 degrees, by randomly splitting the dataset into 80–20% for training and testing, respectively.

**Results** The model showed accuracies of 74, 81, 79, 79, and 84% for 15-, 20-, 25-, 30- and 40-degrees thresholds, respectively. For all the thresholds ATR, Prominence, and visible asymmetry of the waist were the top five most important variables for the prediction. Samples that were wrongly classified as negatives had always statistically significant ( $p \ll 0.01$ ) lower values of ATR and Prominence. This confirmed that these two parameters were very important for the correct classification of the Cobb angle. The model showed better performances than using the 5 and 7 degrees ATR thresholds to prescribe a radiological examination.

**Conclusions** Machine-learning-based classification models have the potential to effectively improve the non-invasive screening for AIS. The results of the study constitute the basis for the development of easy-to-use tools enabling physicians to decide whether to prescribe radiographic imaging.

**Keywords** Adolescent idiopathic scoliosis · Machine learning · Prediction model

## Introduction

Scoliosis is the most common spinal disorder during growth. Adolescent idiopathic scoliosis (AIS) shows an overall prevalence ranging from 0.9 to 12%, with 2 to 3% as the most reported value in the literature [1–4]. AIS progresses more frequently in females than males; for Cobb angles between 10 and 20°, the percentage of affected girls is similar to boys' (1.3:1), but the ratio increases together with Cobb degrees: for Cobb angles ranging from 20 to 30° the girls-to-boys ratio is 5.4:1 and for angles values above 30° the ratio is 7:1 [5, 6]. Curves larger than 50° at the end of growth are associated with a higher risk of progressing through the lifespan [7], health problems in adult life, pain, disability, and progressive functional limitations [6, 8, 9].

✉ Francesco Negrini  
Francesco.negrini@uninsubria.it

<sup>1</sup> Department of Biotechnology and Life Sciences, University of Insubria, 21100 Varese, Italy

<sup>2</sup> Istituti Clinici Scientifici Maugeri IRCCS, 21049 Tradate, VA, Italy

<sup>3</sup> Spine Center, Schulthess Clinic, 8008 Zurich, Switzerland

<sup>4</sup> Biomedical Data Science Lab, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland

<sup>5</sup> ISICO (Italian Scientific Spine Institute), 20141 Milan, Italy

<sup>6</sup> Department of Biomedical, Surgical and Dental Sciences, University “La Statale”, 20122 Milan, Italy

<sup>7</sup> IRCCS Istituto Ortopedico Galeazzi, 20161 Milan, Italy

Early detection of scoliosis becomes fundamental for starting an early and less invasive treatment and improving final results. With screening, the average degree of the curve at diagnosis decreases, the number of prescribed braces increases because of early detection, and the number of performed spinal fusions reduces [10, 11]. Screening is based on a physical examination to identify the need for a radiograph to confirm the diagnosis. The primary evaluation is Adam's forward bending test, and a positive result is highly suggestive of scoliosis [12]. This test allows measuring the angle of trunk rotation (ATR); a 7° Bunnell ATR at the level of the prominence, as measured by the Scoliometer, is the usual cut-off point to indicate suspect scoliosis [3, 13]. Still, this diagnostic test has relatively low sensitivity and specificity [14–16]. A radiological examination is therefore indicated to confirm the positive results of Adam's test. However, increased neoplastic risk due to ionising radiation exposure is a relevant issue, especially in young subjects [17]. Other approaches aimed at avoiding using X-rays for scoliosis follow-up, replacing them, for example, with surface topography [18], but also proved insufficiently reliable in diagnosing spinal deformities. Indeed, radiographs remain needed in scoliosis follow-up and are considered the gold standard for diagnosing and monitoring the pathology [8].

In this study, we hypothesize the possibility of improving the decision to prescribe a radiological examination with a complete evaluation that does not rely only on ATR. An extensive database including other clinical information analysed through machine learning techniques could redefine the classical threshold, increasing its sensitivity and specificity. We aimed to identify a simple formula for radiographic referral of children with suspicion of scoliosis based on history and clinical examination in a specialistic setting.

## Materials and methods

### Study design

This is an observational, cross-sectional study. The study adheres to the STROBE checklist for cross-sectional studies [19].

### Setting

We recruited all patients in a tertiary referral outpatient clinic specialised in spine deformity conservative treatment. The local Ethics Committee approved the study, and all patients (or their parents, if minors) provided informed written consent.

## Dataset

The inclusion criteria for the study were:

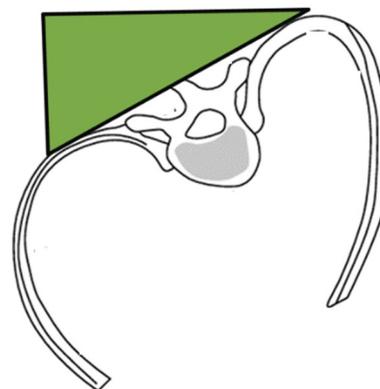
- juvenile or adolescent idiopathic scoliosis patients;
- between 4 and 18 years old;
- first consultation with a spine specialist at our institute;
- availability of a radiographic evaluation within three months of consultation;
- no history of previous bracing.

Our target variable was the Cobb angle of the major scoliotic curve in the coronal radiograph. We considered the following classical independent variables:

- sex,
- age,
- ATR measured with a Scoliometer (° Bunnell) [15],
- Prominence Height (mm) [13],
- Body Mass Index (BMI),
- Familiarity: at least a close relative who had treated scoliosis,
- Asymmetry defined as two or more in one of the TRACE parameters [20],
- localization of the major curve: lumbar, thoracolumbar, and thoracic.

Finally, we added some new independent variables. We considered the orthogonal triangle described by the Prominence Height (one cathetus) and the ATR (inclination of the hypotenuse). Using the trigonometric formulae we found:

- Prominence distance: the second cathetus of the triangle,
- Area of prominence: the area of the triangle (Fig. 1).



**Fig. 1** Visual representation of ATR, Prominence height, and the rectangular triangle that represent the area of the prominence

## Model derivation

Since directly regressing the Cobb angle from the measured parameters proved not feasible after preliminary tests, we decided to make the regression problem a binary classification problem, i.e., to detect if the angle is higher or lower than a predefined threshold. We, therefore, used different thresholds of the Cobb angle (15, 20, 25, 30, and 40 degrees) to split the dataset into two classes. For the sake of easy interpretability of the model, we used a logistic regression model for the classification task. We compared it to the currently used methodology to prescribe a radiological examination, namely an ATR angle above 5 and 7° Bunnell. Since we have five different thresholds, we developed five different logistic regression models, one for each Cobb angle threshold, to predict if the patient has a Cobb angle above or below the selected threshold using the following formula:

$$P(\text{above}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * x1 + \beta_2 * x2 + \beta_3 * x3 + \beta_4 * x4 + \beta_5 * x5 + \beta_6 * x6 + \beta_7 * x7 + \beta_8 * x8 + \beta_9 * x9 + \beta_{10} * x10 + \beta_{11} * x11)}} \quad (1)$$

where  $P(\text{above})$  is the probability that the patient has the Cobb angle above the angle threshold,  $\beta_i$  with  $i$  ranging from 0 to 11, are the coefficients that will be calculated from the model, and  $x_i$ , with  $i$  ranging from 1 to 11 are the independent variables of our model. The coefficients  $\beta$  for each model are reported in the Excel file in the Supplementary Material. Regarding  $x$ ,  $x1$  is the sex,  $x2$  the age,  $x3$  the ATR,  $x4$  the Prominence,  $x5$  the Prominence distance,  $x6$  the Area of the Prominence,  $x7$  the BMI,  $x8$  the Familiarity,  $x9$  the Asymmetry, and  $x10$  and  $x11$  the variables representing the location.

## Internal validation

We randomly split the dataset into 80% for training ( $N=5130$ ) and 20% for testing ( $N=1283$ ). We performed a 10-folds cross-validation (CV) only on the training set to analyse our model's performances and stability across different train-validation sets. To do so, we split the training set into ten groups, iteratively trained the model on nine of them and validated on the remaining one. We repeated this process ten times to cover many train validation sets. After cross-validation, we retrained the model using the full training set and evaluated the final performance on the test set. We performed the cross-validation and the final training for each threshold leading to the five different models. As preprocessing steps, we scaled the numerical variables (Age, ATR, Prominence, Prominence distance, Area, and BMI) to have zero mean and unitary variance.

In this way, all the variables are on the same scale, and no one dominates over the others.

## Discrimination and calibration

We evaluated our model on the repeated tenfold cross-validation and the test set. For the repeated tenfold cross-validation, we computed the Receiver Operating Characteristics (ROC) curves from which we obtained the Area Under the Curve (AUC); we also calculated the mean value and standard deviation for this metric for all the thresholds. The AUC allowed us also to calculate the Youden Index [21] to estimate the optimal classification threshold to maximise both sensitivity and specificity, namely the ability of the model to find positive cases (true positives) and negative cases (true negatives), respectively. Moreover, given the optimal classification threshold, we computed the accuracy, sensitivity, specificity, and F1 score for each run of the repeated cross-

validation. The  $F1$  score is a metric that summarizes the performances of the model by taking into account precision (positive predictive value) and recall (sensitivity), and it ranges from 0 to 1. It is the harmonic mean of precision and recall (Eq. 2).

$$F1 = \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (2)$$

Finally, we averaged the results of the runs to get, for each Cobb angle threshold, a mean value and a standard deviation for each metric.

We conducted the final evaluation of the model performance in the same way. First, we computed the ROC curves separately for each Cobb angle threshold. As we did for CV, we used them to calculate the optimal classification threshold (from 0 to 1) that maximises the sensitivity and specificity of the model. Then, by using this classification threshold, we computed the accuracy, sensitivity, specificity, and  $F1$  score and compared the sensitivity and specificity of our model to those we would have obtained using the ATR thresholds.

We also analysed the most important variables that best predicted the outcome: whether the Cobb angle was below or above each threshold. Indeed, this can be easily done using a logistic regression model by looking at the significance of each coefficient assigned to each predictor. In particular, when the model is trained, we can look at the absolute values of the coefficients associated with each independent variable and rank them by importance from the highest to the lowest. Then, by looking at the  $p$ -values associated with each coefficient, we keep only those with a  $p$ -value lower than 0.05.

Finally, we compared the box plot of the numerical variables between the correctly classified samples (True Positives) and the samples that should have been classified as positives, namely Cobb angle above the threshold, but were wrongly classified as negatives (False Negatives). The purpose was to understand if there were significant differences between the distributions of the numerical variables for the group of true positives and that of the false negatives. First, we tested the normality of the two groups using the Shapiro–Wilk test, and then we applied the *t*-test (if both groups were normally distributed) or the Mann–Whitney test to find out if the true positives and the false negatives had significantly different distributions.

## Results

### Sample

We considered the entire database of 10,813 first clinical evaluations of children referred to our specialised clinic for a consult between 01/07/1996 and 04/05/2018. After

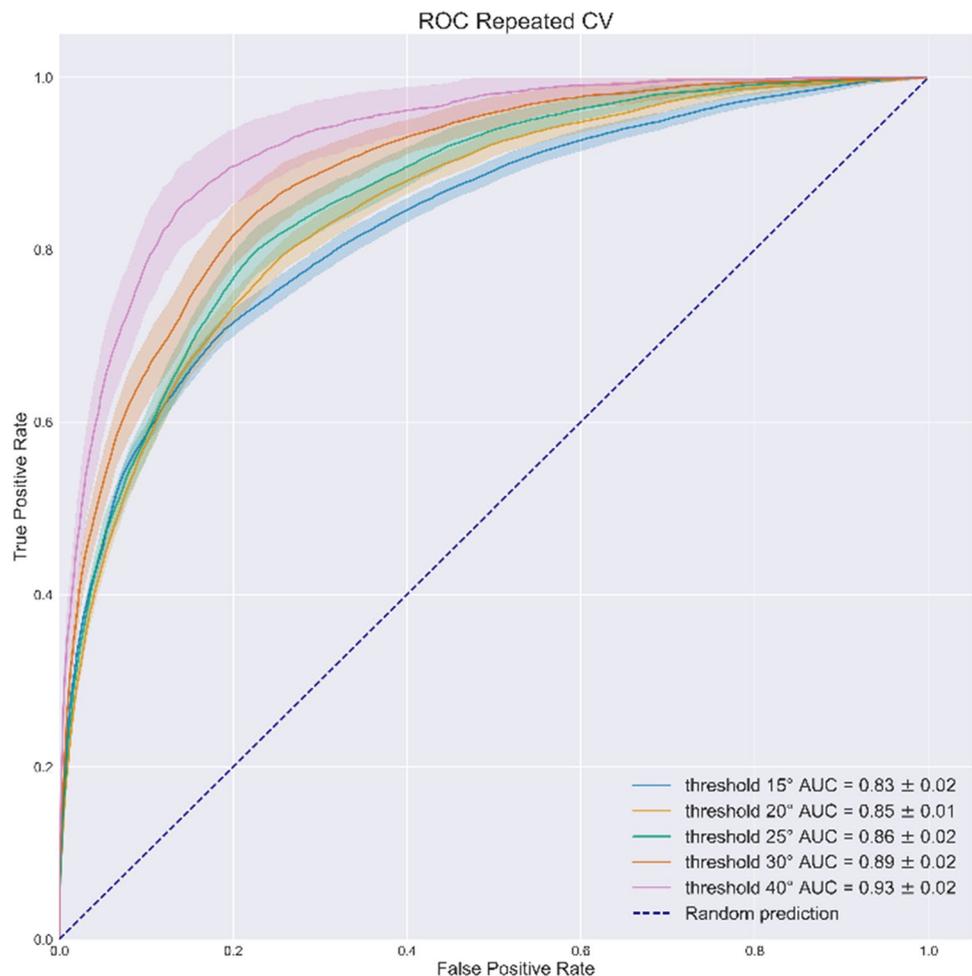
excluding all the patients who did not meet the criteria, we included 7378 children. After removing all children who did not have all the independent variables, we had a final sample of 6413 individuals.

The repeated tenfold cross-validation showed good performance metrics and stability results across the different folds. We obtained high AUC values with low standard deviation indicating the high robustness of our model (Fig. 2 and Table 1).

The *F1* score of our model outperforms the use of the 5 and 7° Bunnell thresholds of ATR with values of 0.77 ( $\pm 0.02$ ), 0.75 ( $\pm 0.01$ ), 0.70 ( $\pm 0.02$ ), 0.63 ( $\pm 0.03$ ), and 0.51 ( $\pm 0.06$ ) for 15, 20, 25, 30, and 40 degrees respectively. The *F1* scores for the 5 degrees threshold were 0.77 (15°), 0.69 (20°), 0.57 (25°), 0.44 (30°), and 0.22 (40°), while for the 7 degrees, threshold were 0.72 (15°), 0.70 (20°), 0.64 (25°), 0.53 (30°), and 0.30 (40°).

The model's performance on the test set surpassed that of using the simple classical thresholds of 5 and 7° Bunnell to recommend a radiological examination. The optimal classification thresholds, as determined from the

**Fig. 2** ROC curves repeated (10 times) 10 folds cross-validation. Each curve is related to one threshold. The solid lines are the means and the shaded areas are the means  $\pm$  standard deviations of the 100 iterations. The dashed line is the random prediction that corresponds to an AUC of 0.5. On the *x*-axis the False Positive Rate (1—specificity) and on the *y*-axis the True Positive Rate (Sensitivity)

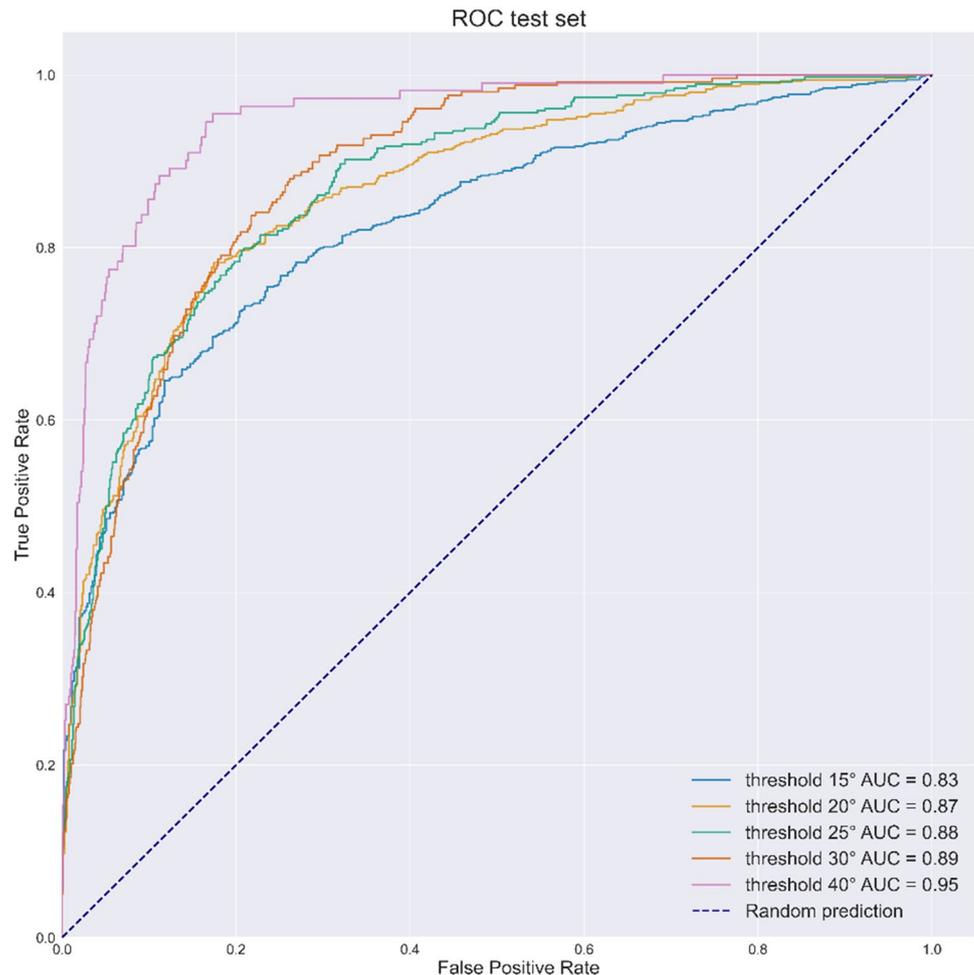


**Table 1** Results of CV

| Angle thresholds | AUC                | Classification thresholds | Sensitivity        | Specificity        |
|------------------|--------------------|---------------------------|--------------------|--------------------|
| 15°              | 0.83 ( $\pm$ 0.02) | 0.65 ( $\pm$ 0.04)        | 0.70 ( $\pm$ 0.04) | 0.83 ( $\pm$ 0.04) |
| 20°              | 0.85 ( $\pm$ 0.01) | 0.44 ( $\pm$ 0.05)        | 0.76 ( $\pm$ 0.05) | 0.79 ( $\pm$ 0.05) |
| 25°              | 0.86 ( $\pm$ 0.02) | 0.30 ( $\pm$ 0.04)        | 0.79 ( $\pm$ 0.04) | 0.79 ( $\pm$ 0.03) |
| 30°              | 0.89 ( $\pm$ 0.02) | 0.19 ( $\pm$ 0.03)        | 0.83 ( $\pm$ 0.05) | 0.80 ( $\pm$ 0.04) |
| 40°              | 0.93 ( $\pm$ 0.02) | 0.1 ( $\pm$ 0.03)         | 0.86 ( $\pm$ 0.05) | 0.85 ( $\pm$ 0.04) |

Mean values ( $\pm$  SD) of AUC (area under the curve) of the receiver operating characteristics (ROC) curves, classification thresholds, sensitivity, and specificity obtained at the repeated 10 folds cross-validation

**Fig. 3** ROC curves test set. Each curve is related to one threshold. The dashed line is the random prediction that corresponds to an AUC of 0.5. On the *x*-axis, the false-positive Rate (1—specificity) and on the *y*-axis the true positive rate (Sensitivity)



ROC curves (Fig. 3), were consistent with the model's performances on the cross-validation (Table 2).

Compared to using values of 5 and 7° Bunnell as thresholds to recommend a radiological examination, our model achieved a superior balance between sensitivity and specificity. The *F1* scores on the test set for 15, 20, 25, 30, and 40 degrees were 0.75, 0.78, 0.70, 0.62, and 0.50, respectively, higher than those obtained using the 5 and 7° Bunnell thresholds. The best trade-off between sensitivity and

specificity was achieved with the 40 degrees threshold, with values of 0.95 and 0.83, respectively. These values consistently outperformed the values of 0.97/0.36 (sensitivity/specificity) and 0.93/0.59 (sensitivity/specificity) obtained using the 5 and 7° Bunnell thresholds, respectively, to recommend a radiograph.

The most important variables included in the model for all the thresholds were sex, ATR, and localisation of the curve. Prominence and BMI were among the most important

**Table 2** Results on the test set

| Thresholds |                | Model       |             | ATR 5       |             | ATR 7       |             |
|------------|----------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Angle      | Classification | Sensitivity | Specificity | Sensitivity | Specificity | Sensitivity | Specificity |
| 15         | 0.69           | 0.64        | 0.88        | 0.81        | 0.56        | 0.63        | 0.82        |
| 20         | 0.47           | 0.78        | 0.83        | 0.87        | 0.50        | 0.72        | 0.75        |
| 25         | 0.28           | 0.80        | 0.79        | 0.91        | 0.44        | 0.79        | 0.69        |
| 30         | 0.19           | 0.83        | 0.78        | 0.94        | 0.41        | 0.86        | 0.65        |
| 40         | 0.07           | 0.95        | 0.83        | 0.97        | 0.36        | 0.93        | 0.59        |

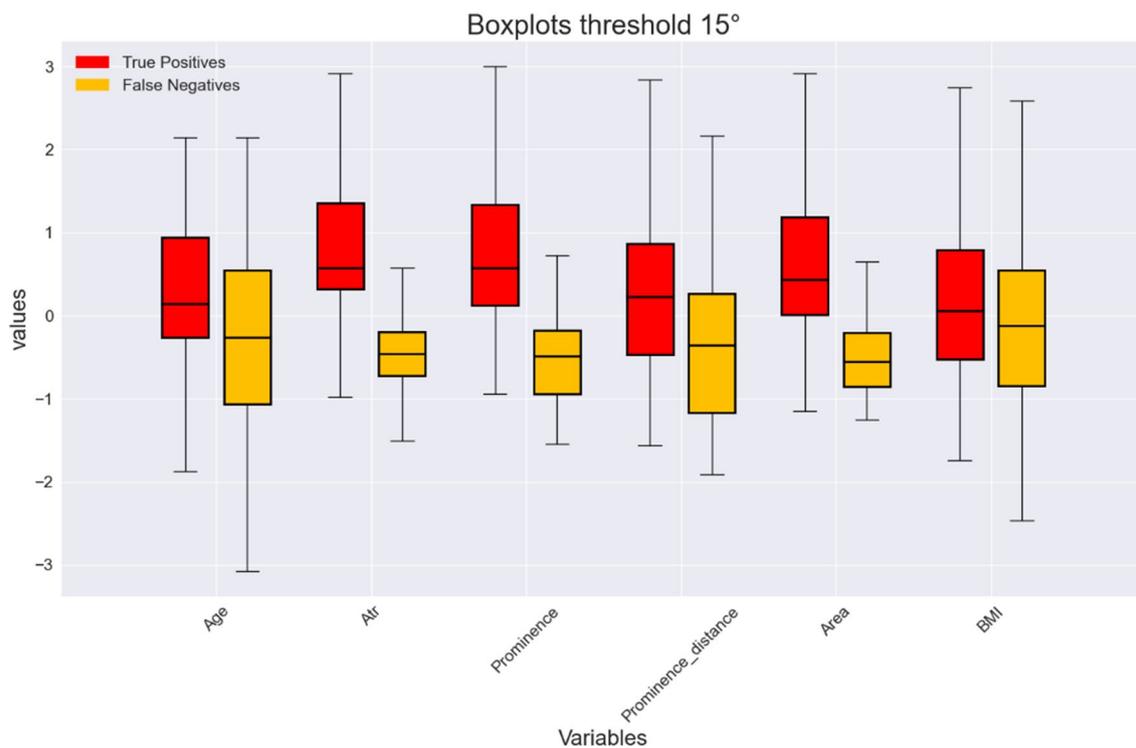
The first two columns report the angles thresholds with the corresponding classification thresholds to maximise both sensitivity and specificity. The other columns report the sensitivity and specificity of our model, of the 5 degrees ATR thresholds, and of the 7 degrees ATR threshold

variables for 20, 25, and 30° Cobb thresholds. The models developed using these three thresholds where the two classes were more balanced had more significant variables (8, 7, and 6, respectively), indicating that they needed more information from different parameters to perform well. Interestingly, familiarity did not have any impact on the prediction.

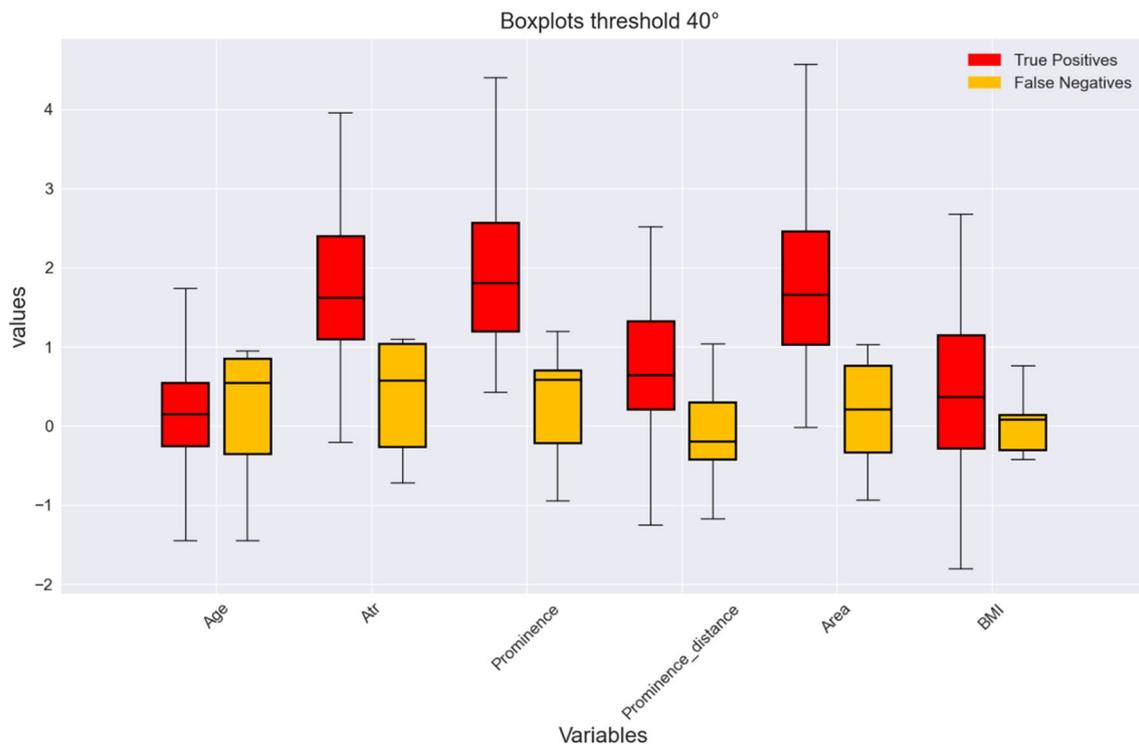
Finally, for the numerical variables (Age, ATR, Prominence, Prominence distance, Area, and BMI), we compared the distributions between the true positives and the false negatives. For the lower thresholds (15, 20, and 25° Cobb), all

the numerical values were significantly different ( $p < 0.01$ ) between the groups. In particular, the true positives had significantly higher values with respect to false negatives, especially for ATR and Prominence (Fig. 4).

This is consistent with the fact that ATR and Prominence are important variables for the classification task and that higher values of these parameters are associated with a higher Cobb angle. Regarding the 30 and 40 degrees threshold, Age and BMI were not significantly different between the two groups (Fig. 5).



**Fig. 4** Box plot numerical variables threshold 15 degrees. The red boxes represent the true positives, while the yellow boxes the false negatives. On the x-axis, the numerical variables and on the y-axis the standardised values



**Fig. 5** Box plot numerical variables threshold 40 degrees. The red boxes represent the true positives while the yellow boxes the false negatives. On the x-axis the numerical variables and on the y-axis, the standardised values

## Discussion

Based on the positive results of this study, machine-learning-based classification models have the potential to effectively improve the non-invasive screening for AIS and reduce the need for radiographic investigation. We developed five different logistic regression models, one for each Cobb angle threshold, to predict if the patient has a Cobb angle above or below the selected threshold. The results of the test set showed that the model outperformed the use of the 5 and 7 degrees thresholds for radiograph prescription for all the thresholds.

Traditionally, only ATR values were used for scoliosis screening. Thus, most of the previous studies use only, or mainly, ATR to propose radiographic examination in screening population. Ashworth et al. in 1988 claimed the Scoliometer has a sensitivity of about 100% and a specificity of about 47% when an ATR of 5° Bunnell is chosen; the specificity increases to 86% at ATR of 7° Bunnell, but the sensitivity drops to 83% [22]. A bigger screening study involving 33,596 children performed in Taiwan found a positive predictive value of 9.5 for 7° Bunnell for curve > 20° [3]. A 1999 US study based on scoliosis screening using Adams test plus Scoliometer (cut-off 6° Bunnell in two repeated measures) reported a 71.1% sensitivity and 97.1% specificity [23, 24].

During Adam's forward bending test, using the combination of Scoliometer and a simple ruler, it is also possible to collect Prominence Height, a measure that has been proposed as a good complimentary tool [13]. ATR and Prominence Height are complementary measures of the same phenomenon, that is prominence, and together describe a rectangular triangle on the back of the patient (Fig. 1).

Other two easily assessed parameters that could contribute to obtaining a reliable tool for radiographic prescription are familiarity and aesthetic impairment. It is known that scoliosis runs in the family, and the role of genetics in its etiology has been proposed, given the increased prevalence in the progeny of scoliotic patients [25]. Furthermore, aesthetic impairment due to scoliosis is often the only or most relevant symptom of early-stage scoliosis [9].

While ATR has widely been used, other parameters such as Prominence Height, aesthetic impairment or familiarity, have never been comprehended before in a model to improve scoliosis screening.

The comparison between our comprehensive model and ATR thresholds commonly used to prescribe a radiological examination showed that our model has better performance considering sensitivity and specificity (Table 2). If we look at the sensitivities for the 5° Bunnell threshold, we can see that they are higher compared to our model but at the cost of

very low values for the specificities, leading to a high risk of prescribing a radiograph when it is not necessary. In particular, for the Cobb threshold of 40° Cobb, we can see that the sensitivity is almost the same (0.97 vs. 0.95 of our model) but the specificity is considerably higher (0.36 vs 0.83 of our model). Regarding the 7° Bunnell threshold, our model is superior considering both sensitivities and specificities (Table 3). Another major point in favour of our model is that the classification thresholds can be modified to maximise the sensitivity or the specificity. The *F1* score takes into account the precision and recall (sensitivity) together. Since the precision is the number of true positives (patients that actually are above the Cobb angle threshold correctly predicted by the model) divided by all the positive model's prediction it is affected by the classification threshold. Indeed, for higher Cobb angle thresholds where the classification threshold is very low, the number of false positives can increase affecting the *F1* score values. Despite this, the performance of our model is still superior to simply using the 5 and 7° Bunnell thresholds to discriminate between patients that require a radiological examination and those who do not.

The analysis of the most important variables showed that sex, ATR, and localisation of the curve are the most important independent variables to take into consideration during the evaluation of scoliosis. For 20, 25, and 30 degrees thresholds also Prominence was among the most important variables confirming results reported in the literature [14, 22]. The box plots of the final evaluation allowed us to understand why the model wrongly classified patients who were actually above the threshold (False Negatives). As expected, the most important numerical values for the classification (mainly ATR and Prominence) were significantly differently distributed between the two groups leading the model to classify patients below the threshold (class 0) with low values of ATR and Prominence.

It should be noted that we used a model-specific classification threshold for all the models to maximise at the same time the sensitivity and specificity. The optimal classification threshold varied a lot because increasing the Cobb angle threshold to binarise the outcome led to different ratios between the number of samples of each one of the two classes. In particular, for the 15° Cobb threshold, we had a ratio  $\frac{class0}{class1}$  of 0.65, while for the highest threshold (40° Cobb), the ratio became 10.5, meaning that by increasing the Cobb angle, we had fewer patients belonging to class 1. The threshold where the two classes were almost balanced was 20° Cobb with a ratio of 1.3. Indeed, for the 40° Cobb threshold model, the classification threshold was very low to “correct” the fact that we have very few patients belonging to class 1 (above threshold). If we had used the standard classification threshold (0.5), we would have obtained a very high

value for accuracy (0.93) and specificity (0.98) but a very low ability to detect the patients above the threshold (sensitivity = 0.41). This shows that, in some scenarios, accuracy is not a reliable metric to evaluate model performances since it can lead to wrong conclusions. So, depending on the task, one can choose to maximise the sensitivity, the specificity, or both by varying the classification threshold.

In the Supplementary Material, we provide the calculator in Excel format to easily implements the formula reported in the paper (Eq. 1) and make it simply usable by clinicians during everyday practice to test the model in different populations. The calculator uses the coefficient (see Eq. 1) to discriminate patients exposed to higher risk of having a curve reaching the pre-defined threshold. The document has 5 sheets one for each threshold that contain models' coefficients, a table with a green header where a user can input the data, a table with a red header where the values are normalised before applying the model, and a table of the results. The latter shows the probability of the classification, the 95% confidence interval of the probability as well as the classification according to the optimal classification threshold (BELOW means that according to the model is unlikely that the possible underlying scoliotic curve reach the radiographic threshold, ABOVE means that is likely that the possible underlying scoliotic curve reaches the radiographic threshold). The user should only input the values into the green table. The tool can be easily used to improve decision-making in a clinical setting.

The present study has a few limitations. The dataset has been collected from a single clinic so it was not possible to perform an external validation that could be useful to evaluate the model on a different population and understand the scalability and generalisability of the model. Moreover, the measurements of the Cobb angle were performed by a single annotator making it impossible to investigate the agreement among different annotators and as a consequence a variability of the Cobb angle.

Despite the limitations, the use of machine learning classification models is a novelty for the topic. In the spine domain, the main previous clinical applications of machine learning techniques include image processing, diagnosis, decision support, operative assistance, rehabilitation, surgery outcomes, complications, hospitalisation and cost [26]. Regarding AIS, a group of researchers used machine learning applied to x-ray images to predict AIS progression [27], while another group developed a machine learning model for three-dimensional (3D) radiographic outcomes prediction as a function of preoperative spinal parameters [28]. However, to our knowledge, it is the first time that machine learning techniques have been used to improve scoliosis screening.

## Conclusion

The machine-learning-based classification model included in the present paper can potentially improve clinical decision-making in everyday clinical settings. After decades of utilising only ATR to choose whether to perform radiographs in young children, this new tool lets us include in the decision process other readily available clinical characteristics of the patients, with the ultimate goal of reducing false positives and false negatives. On the other hand, further studies on different and less selected populations will verify the model's generalisability.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00586-023-07892-1>.

**Acknowledgements** This study was supported and funded by the Italian Ministry of Health—Ricerca Corrente [2022].

**Funding** Open access funding provided by Università degli Studi dell'Insubria within the CRUI-CARE Agreement.

## Declarations

**Conflict of interest** SN owns stock of ISICO. FN and IF are related to SN. All other authors have no conflicts of interest to declare.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Konieczny MR, Senyurt H, Krauspe R (2013) Epidemiology of adolescent idiopathic scoliosis. *J Child Orthop* 7:3–9. <https://doi.org/10.1007/s11832-012-0457-4>
- Brooks HL, Azen SP, Gerberg E et al (1975) Scoliosis: a prospective epidemiological study. *J Bone Joint Surg Am* 57:968–972
- Huang SC (1997) Cut-off point of the Scoliometer in school scoliosis screening. *Spine (Phila Pa 1976)* 22:1985–1989. <https://doi.org/10.1097/00007632-199709010-00007>
- Wong H-K, Hui JHP, Rajan U, Chia H-P (2005) Idiopathic scoliosis in Singapore schoolchildren: a prevalence study 15 years into the screening program. *Spine (Phila Pa 1976)* 30:1188–1196. <https://doi.org/10.1097/01.brs.0000162280.95076.bb>
- Parent S, Newton PO, Wenger DR (2005) Adolescent idiopathic scoliosis: etiology, anatomy, natural history, and bracing. *Instr Course Lect* 54:529–536
- Lonstein JE (2006) Scoliosis: surgical versus nonsurgical treatment. *Clin Orthop Relat Res* 443:248–259. <https://doi.org/10.1097/01.blo.0000198725.54891.73>
- Weinstein SL, Dolan LA, Wright JG, Dobbs MB (2013) Effects of bracing in adolescents with idiopathic scoliosis. *N Engl J Med* 369:1512–1521. <https://doi.org/10.1056/NEJMoa1307337>
- Negrini S, Donzelli S, Aulisa AG et al (2018) 2016 SOSORT guidelines: orthopaedic and rehabilitation treatment of idiopathic scoliosis during growth. *Scoliosis Spinal Disord* 13. <https://doi.org/10.1186/s13013-017-0145-8>
- Negrini S, Grivas TB (2005) Why do we treat adolescent idiopathic scoliosis? What we want to obtain and to avoid for our patients. SOSORT 2005 Consensus paper. *Scoliosis* 1:4. <https://doi.org/10.1186/1748-7161-1-4>
- Altaf F, Drinkwater J, Phan K, Cree AK (2017) Systematic review of school scoliosis screening. *Spine Deform* 5:303–309. <https://doi.org/10.1016/j.jspd.2017.03.009>
- Montgomery F, Willner S (1993) Screening for idiopathic scoliosis. Comparison of 90 cases shows less surgery by early diagnosis. *Acta Orthop Scand* 64:456–458. <https://doi.org/10.3109/17453679308993666>
- Berg AO (1993) Screening for adolescent idiopathic scoliosis: a report from the United States preventive services task force. *J Am Board Fam Pract* 6:497–501
- Ferraro C, Venturin A, Ferraro M et al (2017) Hump height in idiopathic scoliosis measured using a humpmeter in growing subjects: relationship between the hump height and the Cobb angle and the effect of age on the hump height. *Eur J Phys Rehabil Med* 53:377–389. <https://doi.org/10.23736/S1973-9087.16.04227-1>
- Bunnell WP (1993) Outcome of spinal screening. *Spine (Phila Pa 1976)* 18:1572–1580
- Bunnell WP (1984) An objective criterion for scoliosis screening. *J Bone Joint Surg Am* 66:1381–1387
- Grosso C, Negrini S, Boniolo A, Negrini A (2002) The validity of clinical examination in adolescent spinal deformities. *Stud Health Technol Inform* 91:123–125
- Ron E (2003) Cancer risks from medical radiation. *Health Phys* 85:47–59. <https://doi.org/10.1097/00004032-200307000-00011>
- Applebaum A, Ference R, Cho W (2020) Evaluating the role of surface topography in the surveillance of scoliosis. *Spine Deform* 8:397–404. <https://doi.org/10.1007/s43390-019-00001-7>
- von Elm E, Altman DG, Egger M et al (2007) The Strengthening of Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 370:1453–1457. [https://doi.org/10.1016/S0140-6736\(07\)61602-X](https://doi.org/10.1016/S0140-6736(07)61602-X)
- Negrini S, Donzelli S, Di Felice F et al (2020) Construct validity of the Trunk Aesthetic Clinical Evaluation (TRACE) in young people with idiopathic scoliosis. *Ann Phys Rehabil Med* 63:216–221. <https://doi.org/10.1016/j.rehab.2019.10.008>
- Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF (2008) Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J* 50:419–430. <https://doi.org/10.1002/bimj.200710415>
- Ashworth MA, Hancock JA, Ashworth L, Tessier KA (1988) Scoliosis screening. An approach to cost/benefit analysis. *Spine (Phila Pa 1976)* 13:1187–1188. <https://doi.org/10.1097/00007632-198810000-00024>
- Yawn BP, Yawn RA, Hodge D et al (1999) A population-based study of school scoliosis screening. *JAMA* 282:1427–1432. <https://doi.org/10.1001/jama.282.15.1427>
- Dunn J, Henrikson NB, Morrison CC et al (2018) Screening for adolescent idiopathic scoliosis: evidence report and systematic review for the US preventive services task force. *JAMA* 319:173–187. <https://doi.org/10.1001/jama.2017.11669>

25. Aulisa L, Papaleo P, Pola E et al (2007) Association between IL-6 and MMP-3 gene polymorphisms and adolescent idiopathic scoliosis: a case-control study. *Spine* 32:2700–2702. <https://doi.org/10.1097/BRS.0b013e31815a5943>
26. Ren G, Yu K, Xie Z et al (2022) Current applications of machine learning in spine: from clinical view. *Global Spine J* 12:1827–1840. <https://doi.org/10.1177/21925682211035363>
27. Tajdari M, Pawar A, Li H et al (2021) Image-based modelling for Adolescent Idiopathic Scoliosis: Mechanistic machine learning analysis and prediction. *Comp Methods Appl Mech Eng* 374:113590. <https://doi.org/10.1016/j.cma.2020.113590>
28. Pasha S, Shah S, Newton P, Harms Study Group (2021) Machine learning predicts the 3D outcomes of adolescent idiopathic scoliosis surgery using patient-surgeon specific parameters. *Spine (Phila Pa 1976)* 46:579–587. <https://doi.org/10.1097/BRS.0000000003795>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.