

ISSN 2281-4299



DIPARTIMENTO DI INGEGNERIA INFORMATICA
AUTOMATICA E GESTIONALE ANTONIO RUBERTI



SAPIENZA
UNIVERSITÀ DI ROMA

**Supervised and unsupervised learning to
classify scoliosis and healthy subjects
based on non-invasive rasterstereography
analysis**

Tommaso Colombo
Massimiliano Mangone
Andrea Bernetti
Marco Paoloni
Valter Santilli
Laura Palagi

Technical Report n. 8, 2019

Supervised and unsupervised learning to classify scoliosis and healthy subjects based on non-invasive rasterstereography analysis

Tommaso Colombo^{*1}, Massimiliano Mangone[#], Andrea Bernetti[#], Marco Paoloni[#], Valter Santilli[#], Laura Palagi^{*2}

^{*}Department of Computer, Control and Management Engineering A. Ruberti,
E-mail: tommaso.colombo/laura.palagi at uniroma1.it

[#]Department of Physical Medicine and Rehabilitation

E-mail:

massimiliano.mangone/andrea.bernetti/marco.paoloni/valter.santilli at
uniroma1.it

ABSTRACT

Objective. Classify scoliosis versus healthy patients using rasterstereography non invasive surface acquisition, without prior knowledge from X-ray data.

Methods. Data acquisition via Rasterstereography; unsupervised learning for clustering and supervised learning for predicting models. Comparison among Support Vector Machine and Deep Network architectures. K -fold cross validation procedure for assessing the results.

Results. The accuracy and the balanced accuracy of the best supervised model was close to 85%. Classification rates per class were measured using confusion matrix giving low percentage of misclassified patients.

Conclusion. Rasterstereography turns out to be a good tool to identify scoliosis vs healthy patients with the advantage of not exposing patient to unhealthy X-Ray. Furthermore, thanks to the portability and the low cost of the Rasterstereography, it is possible to use it to promote screening campaign.

Keywords. Data Mining; Rasterstereography; Non invasive support system; Scoliosis diagnosis; Support Vector Machine; Deep Learning.

¹The author acknowledges partial support within the project “Distributed optimization algorithms for Big Data” (No RM11715C7E49E89C - 2017) which has received funding from Sapienza, University of Rome.

²This author acknowledges partial support within the project “MIME-BCI: Mindfulness Meditation training supported by Brain-Computer Interfaces” (No PI1161550696379A - 2016) which has received funding from Sapienza, University of Rome.

1 Introduction

Adolescent idiopathic scoliosis (AIS) is a three-dimensional deformity of the spine, which is characterized by deformation of spinal curvatures on the sagittal, frontal and transverse plane. The diagnosis of AIS is made by X-rays, that allow to detect vertebral rotation and to compute Cobb angle, needed for AIS classification. X-rays, however, carry health risk from repetitive exposure to ionizing radiation [9] and cannot aid physician to detect postural changes associated to AIS. Postural assessment with the study of the “static” standing posture, represents a relevant issue in routinely practice of physicians involved in the management of back diseases, especially those involving children and adolescents, in whom a particular attention should be payed that the developing body is growing up correctly. Nowadays physicians usually perform postural evaluation on the basis of a clinical examination, mainly supported by their own experience, aimed to detect deformities in bending, by means of the Adams test, as well asymmetries between the two sides of the body. The main difficulty, here, is to define which postural parameters need to be considered for diagnosis, and the border line between normal and pathologic features.

Recently, rasterstereography has been proposed as an objective method for instrumented three-dimensional (3D) back shape analysis and reconstruction of spinal curvatures and deformities without radiation exposure [25, 10]. Rasterstereography is based on stereophotogrammetric surface measuring of the back, and it provides more than one-hundred different quantitative parameters concerning 3D subject’s posture with a single exam. The main problem with the application of rasterstereography to clinical practice, as well as, for example, to its use on AIS screening, is represented by the lack of a codified system to analyze and to interpret the whole amount of parameters derived from any single acquisition. No ranges of normality still exist for many of rasterstereography parameters, often resulting in a subjective interpretation of objective data.

Indeed, the analysis of data either from a single patient or from a population of subjects is one of the most critical issues in modern medicine. Despite technological advances, too much data could be difficult to understand and could slow down the diagnostic and therapeutic approach, potentially causing unpleasant consequences for patients and operators. Data Mining (DM) and more specifically Machine Learning (ML) techniques have obtained much interest in medicine field to obtain relevant information from different medical data sets. The use of these techniques in medical areas are changing the way to approach to the patients, because they could simplify

and get faster the clinical processes [2, 26].

Differently from classical statistical parametric inference, ML fits in the class of inductive statistical methods that infer from data both the model and its parameters. Indeed in non-parametric models no hypotheses on the distribution or relationships among data are set and both a non-linear model and the values of the parameters are constructed at the same time. These models can be particularly useful in definition of complex mappings where there is no clear known relationships linking the above-mentioned rasterstereography features to the detection of a disease like scoliosis. We hypothesize that DM techniques could be useful in the field of postural analysis, due to the above mentioned difficulties in finding patterns of normality, and that they could be applied to parameters objectively derived from rasterstereography. Particularly, we want to understand if we are able to determine specific datasets of a limited number of features able to assist physicians in distinguish between AIS subjects and healthy ones, only on the basis of rasterstereographic measurements.

For these reasons, the objective of the present study has been to apply unsupervised and supervised ML techniques to automatically distinguish AIS from healthy subjects, using a subset of rasterstereography parameters that can be identified as those that bring most of the information.

Unsupervised learning or clustering consists in detecting if samples can be split into groups, i.e. the clusters, which possess some similarities in a defined metric. In our case we know the number of clusters is inherently defined by the disease, indeed scoliotic and healthy subjects form the two groups. The most common algorithm to perform clustering is K-means [18], where the number of groups K one wants to identify must be given in input. We adopt a clustering strategy as a first step to check how good is the information tied to the only features without driving the classification by the known status of the subject. In the second phase we adopt a supervised learning and we use as label the status (healthy or scoliotic) of the subject. For supervised learning, we consider two of the most famous models namely Support Vector Machines (SVM) [5, 29] and Deep Networks (DN) [15]. Actually SVM has been already used for detecting postural diseases tied to scoliosis in [2] with a different settings of data. Recently DNs, which are FeedForward Neural networks with a deep layered structure, have received much attention in the ML community. In the postural field DNs have been mainly used in image recognition to identify damaged vertebrae in the spine (see e.g. [6, 14]). Besides obtaining a good classifiers to separate the subjects, the aim of our study stays also in the comparison between SVM and DNs. In both cases we consider features selection strategies and we propose

how to combine different strategy. The features selected have been validated by the physicians and are not trivially obtainable by medical observations. Machine Learning is part of a more general process, which is called Knowledge Discovery Databases (KDD) [12], which includes collecting and cleaning of data, extraction of significant patterns and the final evaluation of the results. This paper covers all these steps. Indeed the research has been conducted by the Department of Physical Medicine and Rehabilitation (PMR) for the first steps of the KDD process requiring acquisition of data (the patient selection, the postural evaluation, the scoliosis/healthy diagnosis and the acquisition by the rasterstereography) in league with the Department of Computer, Control, and Management Engineering “Antonio Ruberti” for the second part of the KDD process namely features extraction and construction of learners - both research groups are from Sapienza University of Rome. Of course the overall process required continuous interaction between the two groups of researchers.

The paper is organized as follows: in section 2 we describe the acquisition of data and the first cleaning based on physician observations that led to the definition of the target set for training, the features extraction procedure, and classification models and performance measures considered. In section 3 we apply the methods to the data and we report the performances. Finally in section 4 we report our conclusion on the use of Rasterstereography for detecting postural diseases.

2 Materials and methods

2.1 Rasterstereography acquisition of data

The acquisition of data was performed through rasterstereography by the FormetricTM4D system (Diers International GmbH, Schlangenbad, Germany) reported in the figure 1. Briefly, parallel light lines are projected onto the back surface of undressed patients. The three-dimensional back shape leads to a deformation of the parallel light lines, which can be detected by a camera positioned at a different angle from the projector (triangulation system). Using a standardized mathematical analysis, the following specific landmarks are automatically determined by assigning concave and convex areas to the curved light pattern:

- (i) the spinous process of 7th cervical vertebra (Vertebra Prominens – VP);
- (ii) the spinous process of 12th thoracic vertebra (Th12);

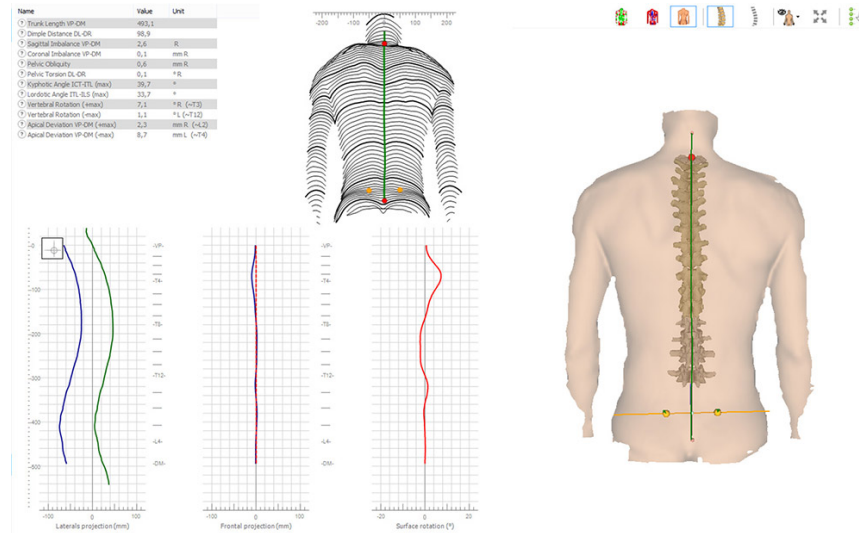


Figure 1: Formetric's output representation (from <https://diers.eu>)

- (iii) the midpoint between the lumbar dimples;
- (iv) the cervical-thoracic inflexion point (ICT);
- (v) the thoracic-lumbar inflexion point (ITL);
- (vi) and the lumbar-sacral inflexion point (ILS).

The patient is asked to stand still in an upright posture at a fixed distance from the camera for 6 seconds, during which a total number of 12 scans are performed. The mean value of the 12 measures is reported as output. Based on these landmarks, a three-dimensional model of the whole spine, the sagittal profile, and shape parameters describing this profile are generated. The accuracy of such measures and FormetricTM functioning can be found in [27, 21]. Derived parameters from automatic landmarks are, among the others, thoracic kyphosis angle, lumbar lordosis angle, flèche lombaire, flèche cervicale and kyphotic apex as described by Stagnara.

For each patient, the total number of rasterstereography features calculated by FormetricTM is 40. The features are all numerical and the full list is reported in Table 1 together with the units of measure.

Feature	Unit of Measure	Feature	Unit of Measure
Trunk length_VP-DM	mm	Flèche lombaire_(Stagnara)	mm
Trunk length_VP-SP	mm	Kyphosis angle_ICT-ITL	degree
Trunk length_VP-SP	%	Kyphosis angle_VP-ITL	degree
Dimple distance-DR	mm	Kyphosis angle_VP-T12	degree
Dimple distance-DL-DR	%	Lordotic angle_ITL-ILS_(max)	degree
Trunk inclination_VP-DM	degree	Lordotic angle_ITL-DM	degree
Trunk inclination_VP-DM	mm	Lordotic angle_T12-DM	degree
Lateral_flexion_VP-DM	degree	Pelvic inclination	degree
Lateral_flexion_VP-DM	mm	Surface rotation_(rms)	degree
Pelvic obliquity_DL-DR	degree	Surface rotation_(max)	degree
Pelvic obliquity_DL-DR	mm	Surface rotation_(+max)	degree
Pelvic torsion_DL-DR	degree	Surface rotation_(-max)	degree
Pelvic inclination_(dimple)	degree	Surface rotation_(width)	degree
Pelvis rotation	degree	Pelvic torsion	degree
Inflexion point_ICT	mm	Lateral deviation_VPDM_(rms)	mm
Kyphotic apex_KA_(VPDM)	mm	Lateral deviation_VPDM_(max)	mm
Inflexion point_ITL	mm	Lateral deviation_VPDM_(+max)	mm
Lordotic apex_LA_(VPDM)	mm	Lateral deviation_VPDM_(-max)	mm
Inflexion point_ILS	mm	Lateral deviation_(width)	mm
Flèche cervicale_(Stagnara)	mm	Pain_index_(Dr.Weiss)_rel	number

Table 1: The full list of FormetricTM features

2.2 Data description and preprocessing

The rasterstereographic collection of data was conducted for clinical purposes in the Department of PMR of Sapienza during the period January 1st, 2010 – December 31st, 2016. Each sample composing the initial database represents the rasterstereography record of one subject, selected according to the following inclusion criteria: (i) male or female and (ii) age between 14 and 30. We excluded subjects with: (i) clinical history of congenital/acquired pathologic condition of vertebrae (e.g. Scheuermann’s disease, spondylolysis, spondylolisthesis); (ii) history of vertebral fractures and/or vertebral surgery; (iii) diagnosis of disc protrusion/hernia at any spinal level; (iv) diagnosis of scoliosis secondary to neurologic, rheumatologic and/or congenital conditions; (v) diagnosis of AIS with Cobb angle measured on X-rays > 45 degrees; (vi) diagnosis of any neurologic and/or rheumatologic conditions.

Once analyzed inclusion and exclusion criteria of patients screened for eligibility, a total of 298 subjects has been enrolled. For each patient FormetricTM returns more than one measure each representing a sample in the dataset. In particular patients enrolled with diagnosis of scoliosis were 272 (~ 90% of total) for a total of 1111 FormetricTM samples. Healthy patients were 26 for a total of 194 FormetricTM samples. The number of samples of healthy/scoliotic is strongly imbalanced and this is a well-known cause of bias in the learning process [23]. To overcome drawbacks due to imbalance

we averaged the FormetricTM measures of the scoliotic patients (that were about 4 for each subject) obtaining 272 AIS averaged samples.

Acquisition date	2010 - 2016
Number of distinct patients	298
Healthy/scoliosis ratio of patients	0.1
Number of samples after balancing	466
Number of healthy samples after balancing	194
Number of AIS samples after balancing	272
Healthy/scoliosis ratio in the target set	0.7

Table 2: Summary of statistics on the dataset

Finally the target set is obtained by merging samples for the two population of AIS (averaged) and healthy (not averaged) patients, so that we obtain a dataset made up of $m = 466$ samples each represented by the 40 FormetricTM features and a label in $\{-1, 1\}$ that corresponds to healthy/scoliotic status. We summarize the main statistics of the target set in Table 2.

As mentioned in the introduction before undergoing the learning phase, data must undergo a cleaning and feature selection phase which usually reduces both the number of samples and the number of features which can be redundant with respect to the learning aim. Indeed learning machines performance are influenced both by the number of samples and the number of features of the target set used for training (see e.g. the surveys [4] [28]). Before undergoing a features selection phase done by a ML procedure, which is described in Section 2.6, data were briefly analyzed for cleaning and scaling purpose. Actually, the physicians recognized that some of the features obtained by the FormetricTM contain duplicate information, in the sense that they correspond to measures of the same quantity, expressed in different units (e.g. mm or degrees). Hence we decided to eliminate the duplicate measures and we finally got a number of distinct features equal to 33. The eliminated features are reported in table 3.

After this basic features reduction, a first run of classification using the tools described in section 2 was performed. The results obtained by unsupervised and supervised classification present inconsistencies. Indeed analyzing the obtained results, we understood that there are features related to trunk length, which are in turn tied to the age of the patient, that had a dominant role in classification thus resulting in a wrong classification . Hence

Feature	Unit of Measure	Eliminated
Trunk inclination_VP-DM	degree	Y
Trunk inclination_VP-DM	mm	N
Lateral flexion_VP-DM	degree	Y
Lateral flexion_VP-DM	mm	N
Pelvic obliquity_DL-DR	degree	Y
Pelvic obliquity_DL-DR	mm	N
Kyphosis angle ICT-ITL_(max)	degree	N
Kyphosis angle_VP-ITL	degree	Y
Kyphosis angle_VP-T12	degree	Y
Lordotic angle_ITL-ILS_(max)	degree	N
Lordotic angle_ITL-DM	degree	Y

Table 3: Duplicated features eliminated with physicians’ support: ‘Y’=eliminated, ‘N’=maintained

Feature	Unit of Measure
Trunk length_VP-DM	mm
Trunk length_VP-SP	mm
Trunk length_VP-SP	%
Dimple distance_DL-DR	mm
Dimple distance_DL-DR	%

Table 4: Eliminated features since highly dependent on trunk length

we decide to remove additional features that are strongly related to trunk length, reported in table 4. Thus the number of total features in the target set reduced further to 27.

Among the 27 remaining features, there are some that still depend on the trunk length that cannot be eliminated because they may bring useful information. For these features, reported in table 5, we divided each value by the trunk length thus obtaining an adimensional value. In this way the target set is not biased by the age of the patients.

At the end of this process we got a target set made up of $m = 467$ samples (referring to 299 patients) each characterized by $n = 27$ input features $x^i \in \mathbb{R}^{27}$ which are reported in Table 6 and one output label in $y^i \in \{-1, 1\}$ which identifies healthy/scoliotic samples. We denote the target set as

$$\mathcal{T} = \{(x^i, y^i) \in \mathbb{R}^n \times \{-1, 1\}, i = 1 \dots, m\}$$

Feature	Unit of Measure
Inflexion point ICT	mm
Kypothic apex KA_(VPDM)	mm
Inflexion point ITL	mm
Lordotic apex LA_(VPDM)	mm
Inflexion point ILS	mm

Table 5: Features dependent on trunk length normalized by trunk length_VP-DM in mm

2.3 Statistical analysis

Before entering a feature selection phase using standard tools in machine learning, we performed basic statistical analysis of the target data to check if it is possible to identify a high degree of correlation either among pairs of input features or among input features and the output class. In particular we perform a Pearson test on the target set \mathcal{T} . None of the features present a strong correlation with the output being the max score 0.59. However, some pairs of features are highly correlated with each other, i.e. they present a Pearson score greater than 0.8 (e.g x_5 and x_{16} which represent different procedure of measuring the pelvic inclination). The full Pearson matrix is reported in table 8 where variables with Pearson score greater than 0.8 are boldface in a green box. We use the identified relationships in connection with the features selection procedure in the next section 2.6. In Table 7 we report a summary of max/min indexes in the Pearson matrix.

2.4 Classification models

In this section we describe the three main classes of Machine Learning (ML) methods for classification that we used to analyze data in the target set

$$\mathcal{T} = \{(x^i, y^i) \in \mathbb{R}^n \times \{-1, 1\}, i = 1 \dots, m\}.$$

In particular as we mentioned in the introduction we are interested in using both unsupervised clustering and supervised classification by using Support Vector Machines (SVMs) and Deep Networks (DNs).

Unsupervised learning does not use any a priori information on the labels of the input data, with the aim of grouping ‘similar’ samples (clusters) on the basis of the features only. The known label y^i are used in the ‘a posteriori’ analysis to evaluate the performance as explained in Section 2.5.

Feature	Name	Unit of Measure
x_1^i	Trunk inclination_VP-DM	mm
x_2^i	Lateral flexion_VP-DM	mm
x_3^i	Pelvic obliquity_DL-DR	mm
x_4^i	Pelvic torsion_DL-DR	degree
x_5^i	Pelvic inclination_(dimple)	degree
x_6^i	Pelvis rotation	degree
x_7^i	Inflexion point_ICT/trunk length_VP-DM	adim
x_8^i	Kypothic apex_KA_(VPDM)/trunk length_VP-DM	adim
x_9^i	Inflexion point_ITL/trunk length_VP-DM	adim
x_{10}^i	Lordotic apex_LA_(VPDM)/trunk length_VP-DM	adim
x_{11}^i	Inflexion point_ILS/trunk length_VP-DM	adim
x_{12}^i	Flèche cervicale_(Stagnara)	mm
x_{13}^i	Flèche lombaire_(Stagnara)	mm
x_{14}^i	Kyphosis angle_ICT-ITL	degree
x_{15}^i	Lordotic angle_ITL-ILS_(max)	degree
x_{16}^i	Pelvic inclination	degree
x_{17}^i	Surface rotation_(rms)	degree
x_{18}^i	Surface rotation_(max)	degree
x_{19}^i	Surface rotation_(+max)	degree
x_{20}^i	Surface rotation_(-max)	degree
x_{21}^i	Surface rotation_(width)	degree
x_{22}^i	Pelvic torsion	degree
x_{23}^i	Lateral deviation_VPDM_(rms)	mm
x_{24}^i	Lateral deviation_VPDM_(max)	mm
x_{25}^i	Lateral deviation_VPDM_(+max)	mm
x_{26}^i	Lateral deviation_VPDM_(-max)	mm
x_{27}^i	Lateral deviation_(width)	mm

Table 6: List of features constituting the row $x^i \in \mathbb{R}^{27}$ of the clean data set

	features vs features	features vs output
Max abs correlation	0.96	0.59
Min abs correlation	0.00	0.04
Avg abs correlation	0.19	0.26

Table 7: Summary of Pearson coefficients (absolute values)

	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	y
x_1	-0.10	0.10	-0.02	0.18	-0.03	-0.26	-0.41	-0.06	0.06	-0.03	0.33	0.84	-0.36	-0.22	0.18	0.10	-0.11	-0.04	-0.12	0.05	-0.17	0.19	0.08	0.12	-0.06	0.17	0.26
x_2		-0.20	-0.33	-0.01	0.13	-0.04	0.10	0.09	0.06	0.09	-0.06	0.09	-0.05	-0.02	-0.02	-0.01	0.21	0.18	0.14	0.06	0.36	-0.02	-0.09	-0.08	-0.05	-0.03	0.05
x_3			-0.14	0.05	-0.04	-0.06	0.05	0.04	0.05	0.04	-0.04	-0.09	-0.07	0.02	0.05	-0.01	0.01	0.06	-0.03	0.07	0.13	0.09	-0.00	0.03	-0.11	0.10	-0.04
x_4				0.02	-0.16	0.02	0.00	0.01	-0.01	-0.04	-0.06	0.00	0.00	0.05	0.05	0.18	-0.14	-0.06	-0.16	0.07	-0.13	0.09	0.09	0.10	-0.02	0.10	0.11
x_5					-0.09	-0.09	-0.20	-0.31	-0.38	-0.44	-0.50	-0.32	-0.26	0.72	0.85	0.42	-0.07	0.14	-0.27	0.34	-0.04	0.24	0.14	0.18	-0.00	0.14	0.59
x_6						0.06	0.08	0.11	0.11	0.12	0.03	0.03	0.00	-0.11	-0.10	-0.07	0.27	0.26	0.18	0.10	0.24	-0.01	0.05	0.01	0.03	-0.01	-0.11
x_7							0.31	0.19	0.18	0.15	0.07	0.02	0.11	-0.02	-0.03	0.04	0.24	0.25	0.09	0.16	0.15	0.08	-0.02	0.08	-0.14	0.15	-0.21
x_8								0.74	0.73	0.70	-0.21	0.35	-0.09	0.03	-0.10	-0.12	0.13	0.06	0.10	-0.03	0.10	-0.04	0.09	0.03	0.11	-0.07	-0.38
x_9									0.96	0.86	0.07	0.09	-0.15	-0.17	-0.20	-0.23	0.11	-0.03	0.17	-0.15	-0.01	-0.11	0.04	-0.05	0.12	-0.14	-0.39
x_{10}										0.81	0.14	-0.01	-0.20	-0.27	-0.23	-0.25	0.11	-0.04	0.19	-0.18	-0.03	-0.09	0.07	-0.02	0.13	-0.11	-0.45
x_{11}											0.19	0.11	-0.14	-0.29	-0.30	-0.34	0.14	-0.05	0.26	-0.25	-0.03	-0.17	0.03	-0.10	0.16	-0.20	-0.47
x_{12}												-0.25	0.14	-0.65	-0.51	-0.35	-0.06	-0.23	0.16	-0.34	-0.10	-0.21	-0.08	-0.16	0.03	-0.13	-0.19
x_{13}													0.44	0.12	-0.31	-0.32	0.05	-0.13	0.23	-0.30	0.13	-0.34	-0.12	-0.24	0.15	-0.32	-0.39
x_{14}														0.09	-0.21	-0.30	0.00	-0.15	0.20	-0.29	-0.05	-0.40	-0.17	-0.34	0.13	-0.38	-0.14
x_{15}															0.79	0.23	-0.02	0.09	-0.11	0.16	0.01	-0.01	0.01	-0.04	0.07	-0.11	0.37
x_{16}																0.33	-0.04	0.13	-0.18	0.26	-0.00	0.16	0.12	0.11	0.05	0.05	0.48
x_{17}																	-0.13	0.32	-0.65	0.79	-0.03	0.59	0.27	0.52	-0.17	0.54	0.46
x_{18}																		0.83	0.70	0.23	0.48	-0.09	-0.25	-0.20	-0.20	-0.06	-0.08
x_{19}																			0.31	0.69	0.49	0.24	-0.07	0.12	-0.29	0.27	0.18
x_{20}																				-0.48	0.32	-0.49	-0.39	-0.55	-0.01	-0.47	-0.31
x_{21}																					0.21	0.60	0.23	0.53	-0.26	0.61	0.41
x_{22}																					0.07	-0.07	0.02	-0.16	0.15	-0.06	-0.06
x_{23}																						0.30	0.72	-0.34	0.84	0.26	0.26
x_{24}																							0.96	0.67	0.22	0.08	0.08
x_{25}																								0.21	0.69	0.16	0.16
x_{26}																										-0.52	-0.10
x_{27}																											0.20

Table 8: Pearson Correlation matrix among features: in a green box the most correlated ones

In this case our aim from a clinical point of view is to check whether the features describing each patient contain enough 'good' information to allow a natural division into two clusters. Similarity is usually measured by a metric distance between samples both intra-group and extra-group with the aim of maximizing the distance between samples in different clusters, while minimizing the distance between samples belonging to the same cluster. We select as clustering method an improved version of the basic K-Means [17] algorithm called K-Means++ [3], where we set the number of clusters K to $K = 2$ (healthy vs scoliosis). As metric distance we use the standard Euclidean norm that defines the distance between x^1 and $x^2 \in \mathbb{R}^n$ as

$$d(x^1, x^2) = \|x^1 - x^2\| = \sqrt{\sum_{i=1}^n (x_i^1 - x_i^2)^2}.$$

The comparison of the results obtained by unsupervised learning versus those obtained by a supervised procedure using the label to drive the learning procedure, can help in checking the existence of biasing features. Actually

this happens in the first stage of our study and led to normalization of data to avoid cluster implicitly based on the age of the patient. The results are reported in section 3 and they highlight that the data can be clustered sufficiently well.

On the other hand in supervised classification the task is to learn from 'labelled' examples (x^i, y^i) $i = 1, \dots, m$ as given in the target set \mathcal{T} . As supervised methods we used both the Support Vector Machines (SVMs) (see e.g. [35] and for algorithms the survey [29]) which are naturally defined for classification problems and Deep Networks (DNs) [15] which gained much attention in the recent years, with the additional aim, beyond the classification performance, of comparing the performance among these different tools.

Support Vector Machines (SVMs) are supervised binary classifiers in the class of kernel methods that learn the possibly nonlinear border between data belonging to different classes. Linear SVM classify point using hyperplanes defined by $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ and obtained by solving the following (Primal) optimization problem:

$$\min_{w \in \mathbb{R}^n} \|w\|_2^2 + C \sum_{i=1}^m \max\{0, 1 - y^i(w^T x^i + b)\} \quad (1)$$

where C plays the role of a penalty parameter on the misclassified points to be set in the tuning phase of the model. However usually Nonlinear SVMs are used which define a nonlinear decision function to predict the class of a new input x as

$$class(x) = sign \left(\sum_{i=1}^m \alpha_i k(x^i, x) + \beta \right),$$

where $k(\cdot, \cdot)$ is a kernel function which represents a measure of similarity, i.e. a scalar product among data points in a transformed nonlinear space. The use of kernels allows to define nonlinear separation surface avoiding the explicit nonlinear transformation in a higher dimensional space [31], thus avoiding the so called curse of dimensionality. Indeed the parameters $\alpha_i \in \mathbb{R}$, $i = 1, \dots, m$ and $\beta \in \mathbb{R}$ appearing in the nonlinear decision function are obtained as the solution of the dual formulation of the SVM training

problem

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y^i y^j k(x^i, x^j) \alpha_i \alpha_j - \sum_{j=1}^m \alpha_j \\ & \sum_{i=1}^m y^i \alpha_i = 0 \\ & 0 \leq \alpha_j \leq C \quad j = 1, \dots, m \end{aligned}$$

The kernel function may depend on hyper-parameters too and their values is set during the tuning process using a k -fold cross validation procedure. The DNs are multilayer feedforward neural networks organized into layers numbered from $\ell = 0$ (input layer) to $\ell = L$ corresponding to the output layer as reported in picture 2.

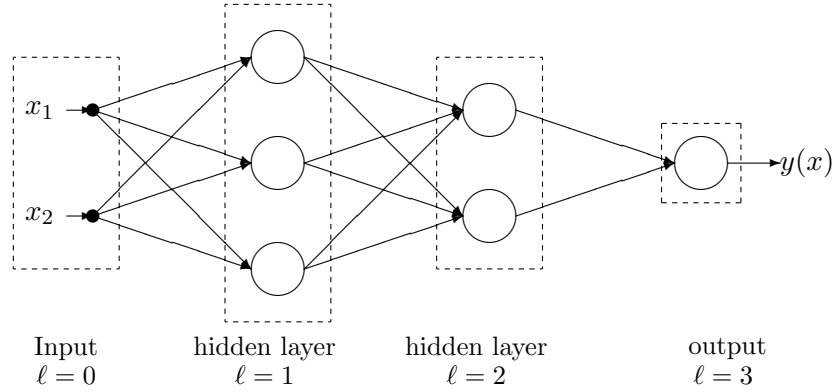


Figure 2: DN with two inputs ($n = 2$), two hidden layers ($L = 3$) and a single output

Assuming a linear output unit and hidden units with activation function given by $g(\cdot)$ (that we assume be the same for all the layers) the expression of the output

$$\tilde{y}(w; x) = W^L g(W^{L-1}; g(\dots; g(W^1; x))) \dots \quad (2)$$

where W^ℓ with $\ell = 1, \dots, L$ is the vector of weights from layer $\ell - 1$ to layer ℓ and its size depends on the number of unit in each layer, In particular, being N^ℓ the number of units in layer ℓ we have $W^\ell \in \mathbb{R}^{N^{\ell-1} \cdot N^\ell}$.

The parameters W^i , $i = 1, \dots, L$ are obtained by minimizing the regularized empirical error

$$R = \sum_{i=1}^m (y_i - \tilde{y}(w; x^i))^2 + C \|w\|^2$$

where C is an hyper-parameter. We set $C = 0$ in our numerical testing. Since we are tackling a classification problem, we apply a sigmoid to the output $\tilde{y}(w; x)$ of the network, so that the final output is $\tanh(\tilde{y}(w; x))$ that returns a value in $\{-1, 1\}$.

2.5 Performance measures

The ultimate task of a machine learning is to give good performance on new unseen samples with unknown label. This task is called generalization ability of a learner and it is in contrast with the perfect learning of the training data that leads to the so-called over-fitting phenomenon (see e.g. [28] and references therein). Hence both in supervised and in unsupervised learning, two main phases must be distinguished: the training phase, where the machine is trained using the samples of the target set, and a prediction phase, where the trained machine is used to predict the label (namely the belonging class or cluster) of future unseen data samples. In order to check the performance of a learner without being biased by the learning process itself, usually the training phase is repeated by inserting some randomness in the process.

In particular in the case of unsupervised learning, we use all the target set as training set and we repeat M times the K-means++ procedure [3], which has a random seed to start with, and we averaged the results. In particular, to check the correctness of the clusters \mathcal{C}_i , $i = 1, 2$ obtained at the end of each of the M runs of the training phase we use the known labels y^i . In particular we measure the correctness of the clusters with the *purity* which is a simple and transparent accuracy measure. Each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy is measured by counting the average number of correctly assigned samples. Formally we can write it as:

$$ACC = \frac{1}{2m} \max \left\{ \sum_{i=1}^m |y^i + \text{class}_i|, \sum_{i=1}^m |y^i - \text{class}_i| \right\},$$

with

$$\text{class}_i = \begin{cases} +1, & x^i \in \mathcal{C}_1, \\ -1, & x^i \in \mathcal{C}_2. \end{cases}$$

In supervised learning, the target set is usually split into two parts: a training set, used only in the learning phase to train the machine, and a test set, used only in a post-learning analysis to quantify the generalization performance. In this way the performance indicators of the learning machine are

computed on patients never shown to the learning process. However due to the limited dimension of the target set, we prefer to use the full target set and use instead a cross validation procedure. IN this case the available data are split randomly into M subsets and the machine is trained M times using $M - 1$ subsets and leaving out one of the subsets that is instead use as validation set to compute the Key Performance Indicators (KPIs). The average of these KPIs over the M runs represents an estimation of the generalization performance.

The Key Performance Indicators (KPIs) to measure the quality of a classification machine used in this paper are:

- a 2×2 confusion matrix, where each elements represents the (averaged) percentage of classification of instances True Negative (TN), True Positive (TP), False positives (FP) and False Negatives (FN) as shown below

Real / Predicted	Scoliotic	Healthy
Scoliotic	TP	FN
Healthy	FP	TN

- classification accuracy (i.e. percentage of correct classified patients)

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}$$

- Balanced Accuracy (BACC)

$$BACC = \frac{1}{2} \left[\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right]$$

The Key Performance Indicators (KPIs) are reported as an average on the test set when supervised learning is used. The perfect learning corresponds to a 100% accuracy and it consists in having the sum over the diagonal of the confusion matrix being 100. This situation, whenever it appears, is in general untrustworthy being a signal of overfitting.

The training procedure both in the case of unsupervised and supervised learning can be summarized in the following scheme.

Training procedure scheme
<ul style="list-style-type: none"> • Given the target set \mathcal{T} given by m-by-$(n + 1) = 466 \times 28$ samples; • Repeat M times <ol style="list-style-type: none"> 1. Extract randomly the training and the validation sets as a percentage of the m samples (patients): <ul style="list-style-type: none"> - for supervised learning respectively 70% and 30% - for unsupervised learning respectively 100% and 0% ; 2. Train a classifier using the training set 3. Compute the KPIs of the obtained classifier • Average the KPIs over the M runs

Table 9: Scheme of the Training procedure and avergare evaluation of KPI

2.6 Features selection procedure

A critical aspect for the success of any learning procedure stays in the reduction of the number of input features. Indeed it may happen that some features are redundant and/or add noisy information so that eliminating them will help the learning task. Further the features selection can give insight on which are the features that hold the most significant information and hence can give doctors indications about the key measures of Formetric 4D. To this aim we perform a feature selection phase before entering the true learning phase. In the literature different methods for features selection have been proposed. We choose four different algorithms, described below, and we define a ranking of the features by assigning as vote how many algorithm have selected it.

Further we use this ranking to reduce the dimension of the training set in the experiments to test whether the most selected features actually include the most significant patterns. To this aim, we define the *minimal features set* as the set of those features chosen by at least 3 of the 4 algorithms.

As tools for features selection we used four different algorithms, both supervised and unsupervised, listed below:

1. L2-regularized SVM [5]

$$\min_{w \in \mathbb{R}^n} \|w\|_2^2 + C \sum_{i=1}^m \max\{0, 1 - y_i(w^T x^i + b)\}$$

2. L1-regularized SVM [39]

$$\min_{w \in \mathbb{R}^n} \|w\|_1 + C \sum_{i=1}^m \max\{0, 1 - y_i(w^T x^i + b)\}$$

3. Mutual information (MI) which is a non-negative value that measures the dependency between two random variables. The function relies on nonparametric methods based on entropy estimation from k-nearest neighbors distances as described e.g. in [20, 30].
4. Analysis Of Variance (ANOVA) [16]

We emphasize that the selection of the features depends on the available data. In order to take care of this aspect, we perform M replicates selecting each time randomly 70% of the available data. For each algorithm, we choose the features selected more than 80% of the times over the M replicates. The feature selection procedure can then be described as follows:

At the end of the features selection we construct a *Reduced Target Set* which is obtained from the full target set by considering only the minimal set of features. We use the Reduced Target Set as a term of comparison w.r.t. the full data set.

2.7 Classifiers tuning and training

In both the two models used for supervised classification, DNN and SVM, some hyper-parameters must be fixed in order to get the best performance of the machines.

The SVM classifiers has two type of parameters: the regularization parameter C that controls misclassified points and the kernel parameters. Both linear and Gaussian kernels were tested. Linear kernel $k(x^i, x) = x^{iT} x$ does not depend on hyper-parameters, while the Gaussian kernel

$$k(x^i, x) = e^{-\gamma \|x^i - x\|^2},$$

Features Selection procedure

- Given the target set in a table m -by- $(n + 1) = 466 \times 28$
- Repeat M times
 - Randomly extract the 70% of the target set
 - Apply the four different feature selection algorithms
- For each algorithm, rank the features according to the number of times selected over the M runs
- For each algorithm, choose the features selected more than 80% of the times
- Construct the *minimal features set* by choosing the features selected by more than 3 algorithms out of the 4

Table 10: Features Selection procedure

	hyper parameters	
SVM	C	γ
DNN	N^ℓ	L

Table 11: Hyper parameters to be set in SVM/DNN classifiers

which corresponds to a transformation of the data points x^i into an infinite dimensional space, presents the hyper-parameter γ called the width of the kernel.

As regard DNN, we need to choose the number of layers L and neurons per layer N^ℓ , $\ell = 1, \dots, L$ and the activation function $g(z)$. We tested both the ReLU (Rectified Linear Unit) $g(z) = \max\{0, z\}$ and the sigmoid function $g(z) = \frac{e^z}{e^z + 1}$ which do not include any hyper parameter in the library that we use in the implementation (ScikitLearn [11]). We perform the tuning procedure for different values of L and N^ℓ , $\ell = 1, \dots, L$.

We summarize the different hyper-parameters to be fixed for each configuration in Table 11.

The selection of the correct hyper-parameters C, γ has been done by using a k -fold cross validation procedure with a grid search on the hyper-parameters. The average of these errors over the k runs represents an estimation of the

generalization performance and we select the setting of hyper-parameters that give the best average validation error. The process can be summarized as follows:

<i>k</i> -fold cross validation
<ul style="list-style-type: none"> • Given the target set \mathcal{T}; • Define a grid \mathcal{H} of hyper-parameters values • For each $h \in \mathcal{H}$, select the h-th hyperparameter pair <ul style="list-style-type: none"> – Repeat M times <ol style="list-style-type: none"> 1. Extract randomly the training and the validation sets; 2. Train a classifier using the training set 3. Compute the KPIs on the validation set – Average the KPIs over the M runs • Select the $h \in \mathcal{H}$ that gives the best average KPIs

Table 12: *k*-fold cross validation

3 Results and discussion

3.1 Recap of the procedure and toolboxes

We sum up the overall procedure that led to the final learning system.

1. **Data acquisition.**
2. **Data Balancing.** Since healthy/scoliotic classes are not balanced we randomly oversampled the less numerous class
3. **Data cleaning.**
4. **Features reduction.** We define a ranking of features using different procedures
5. **Training of unsupervised classifiers.**

6. **Hyper parameters tuning of supervised classifiers.**
7. **Training of supervised classifiers.**
8. **KPIs analysis and medical comments.**

We use standard tools for unsupervised and supervised training. For the numerical testing we use the Python’s ScikitLearn [11] library which is an open source ML library, for clustering and SVM. In particular ScikitLearn’s ‘cluster.KMeans’ package with $K = 2$, with the ‘k-means++’ initialization [3] for clustering, and LIBSVM [5] for SVM. For training the DNN we use ADAM [19] as implemented on Keras [7] over TensorFlow paradigm [1]. In all our results we repeat the procedure $M = 100$ times. In addition to the standard assessment of the system’s performance we design experiments to check the role of each step of the procedure. In particular:

1. to quantify the effect of features selection we perform the experiments using the target data (after cleaning) with the full features and the reduced one;
2. to check the intrinsic information in the data we perform unsupervised clustering both on the full target set and on the reduced one;
3. to check the effect of tuning we test linear and Gaussian kernel in SVM classification and deepness and wideness in DNN, by choosing different values of L, N^ℓ , and the activation function selecting either the Relu or the sigmoid function.

3.2 Features reduction

We ran the feature selection procedure as detailed in Section 2.6. In Table 13 we report a detailed output: for each of the four algorithms we indicate with a "x" the features selected at least in the 80% of random runs. The last columns counts how many algorithms selected the feature.

We propose to select as the most significant features those ones that are selected by at least 3 out of the 4 algorithms. These are the first 14 values in Table 13 which constitutes the minimal feature set:

$$\{x_4, x_5, x_7, x_9, x_{10}, x_{11}, x_{13}, x_{15}, x_{16}, x_{17}, x_{19}, x_{20}, x_{25}, x_{26}\}.$$

In order to check the effectiveness of the selection procedure with respect to standard statistic tools, we perform a selection by solely analyzing Pearson’s

Variables	name	L2 SVM	L1 SVM	MI	ANOVA	Final score
x_4	Pelvic torsion_DL-DR	x	x	x	x	4
x_5	Pelvic inclination_(dimple)	x	x	x	x	4
x_7	Inflexion point ICT /trunk length_VP-DM	x	x	x	x	4
x_9	Inflexion point ITL /trunk length_VP-DM	x	x	x	x	4
x_{10}	Lordotic apex_LA_(VPDM) /trunk length_VP-DM	x	x	x	x	4
x_{13}	Flèche lombaire_(Stagnara)	x	x	x	x	4
x_{17}	Surface rotation_(rms)	x	x	x	x	4
x_{25}	Lateral deviation_VPDM_(+max)	x	x	x	x	4
x_{11}	Inflexion point ILS /trunk length_VP-DM		x	x	x	3
x_{15}	Lordotic angle_ITL-ILS_(max)		x	x	x	3
x_{16}	Pelvic inclination	x		x	x	3
x_{19}	Surface rotation_(+max)	x	x		x	3
x_{20}	Surface rotation_(-max)	x		x	x	3
x_{26}	Lateral deviation_VPDM_(-max)	x	x	x		3
x_1	Trunk inclination_VP-DM		x		x	2
x_2	Lateral flexion_VP-DM	x		x		2
x_6	Pelvis rotation	x			x	2
x_8	Kyphotic apex_KA_(VPDM) /trunk length_VP-DM	x		x		2
x_{12}	Flèche cervicale_(Stagnara)			x	x	2
x_{14}	Kyphosis angle ICT-ITL			x	x	2
x_{18}	Surface rotation_(max)	x	x			2
x_{21}	Surface rotation_(width)	x			x	2
x_{24}	Lateral deviation_VPDM_(max)			x	x	2
x_{27}	Lateral deviation_(width)	x	x			2
x_3	Pelvic obliquity_DL-DR			x		1
x_{23}	Lateral deviation_VPDM_(rms)				x	1
x_{22}	Pelvic torsion					0

Table 13: Features ranking

correlation matrix in Table 8 and selecting those features which have a Pearson correlation with the output ≥ 0.2 . In this case the Pearson’s selected features are

$$\{x_1, x_5, x_7, x_8, x_9, x_{10}, x_{11}, x_{13}, x_{15}, x_{16}, x_{17}, x_{20}, x_{21}, x_{23}, x_{27}\}.$$

If we add also the Pearson’s selection in the voting procedure for choosing the most prominent features and we consider those features that obtained at least four votes we obtain the set

$$\{x_4, x_5, x_7, x_9, x_{10}, x_{11}, x_{13}, x_{15}, x_{16}, x_{17}, x_{20}\}.$$

We name this set Pearson minimal set. We note that the Pearson minimal set is not the intersection of the Pearson’s feature set and the minimal features set. Indeed the feature x_4 (Pelvic torsion_DL-DR) has a low value of the Pearson coefficient (0.11) and it would not be selected by solely the Pearson selection rule but it is instead selected by all the other four algorithms.

After the features selection has been performed, we have available four training sets with the same number of samples but one with the full 27 features and the others with a reduced number of features.

Full / Selected features	Healthy	Scoliotic
Healthy	25.7 / 32.4	16.1 / 9.2
Scoliotic	22.2 / 18.7	36.0 / 39.8

Table 14: Confusion matrix of the unsupervised classifier: the first number refers to the full target set whilst the second number refers to the minimal set.

We apply the learning process, both unsupervised and supervised, using these four datasets, namely the full dataset, the minimal set, the Pearson set and the Pearson minimal set, to verify how these selections impact over the performance. Nonetheless, it turns out that among the three, the features selection in the minimal set led to the highest KPIs. Hence in the results we report only the comparison between the full set and the minimal set.

The results reported in sections 3.4 and 3.3 show that such a cut in dimensionality does not strongly affect performance of the ML which is measured by the KPI indicators suggesting that features in Table 6 have a key role in classifying scoliosis.

3.3 Performance of unsupervised classifiers

We first apply an unsupervised learning process to the two target sets defined in section 2.6 where the labels are taken out. The unsupervised learner is applied with the task of grouping the patients with the aim of “reconstructing” the two original classes.

We report the results in terms of accuracy and confusion matrix. The accuracy reached using the full set of 27 features and only the 14 selected features of the minimal set 61.7% and 72.2%, respectively. It is interesting to see how the accuracy increased with the features reduction. Indeed this was foreseeable since lower dimension makes the task easier for a learning machine that does not exploit the labels to cluster the samples in groups. The confusion matrices of the unsupervised learning machine obtained on the two target sets are shown in Table 14. In both the experiments, false positives and false negatives are well balanced, showing once more the robustness of the target data, namely that the rastereographic information allow to clearly separate the two clusters, i.e. subjects with scoliosis and healthy ones.

It is worth to mention that we performed the clustering procedure as a first step at the very beginning of the project and the results pointed out some

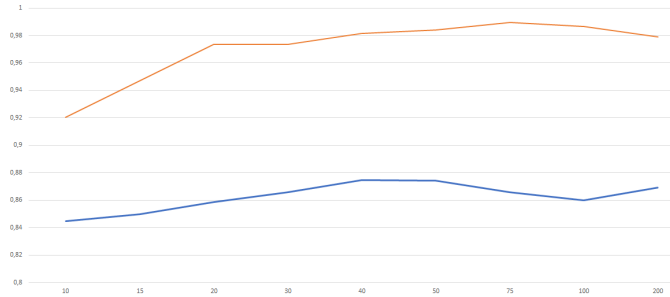


Figure 3: Test (blue) and Training (red) accuracy for increasing number of neurons in the unique hidden layer

	ACC	BACC
Full set	87.5%	87.4%
Minimal set	83.7 %	83.4%

Table 15: Accuracy (ACC) and Balanced Accuracy (BACC) of a shallow Network with $N^1 = 40$.

bias in the data that induced wrong clusters. We used the information derived by the first clustering results to clean up the data.

3.4 Performance of supervised classifiers

We set $M = 100$ in the training procedure scheme in Table 9. We first trained the DNN on the full data set using different architectures with the aim of exploring the role of wideness and deepness. We tried both sigmoid and ReLu as activation functions, but the ReLu performs significantly worst. Hence we report the results only for the sigmoid. In particular, we trained a shallow network (i.e. one hidden layer) with increasing number of neurons N in order to understand the role of wideness. We increased the number of neurons N from 5 to 200 and we report the results in terms of average test (blue) and training accuracy (red) in Figure 3. The averaged classification accuracy is almost everywhere higher than 80% being $40 \leq N \leq 50$ the best range of neurons. We observe that the training accuracy reaches almost value 1 for $N \geq 50$. In Table (15) we report the accuracy and balanced accuracy of the best configuration which corresponds to $N^1 = 40$ in Table 15.

We also performed a test increasing the deepness from $L = 2$ to $L = 10$ with

	ACC	BACC
Full set	86.3%	86.6%
Minimal set	85.5 %	85.5%

Table 16: Accuracy (ACC) and Balanced Accuracy (BACC) of Deep Network with $L = 3$.

a fixed the number of neurons per layer $N^\ell = 20$ for all $\ell = 1, \dots, L$. The results are in the Figure 4.

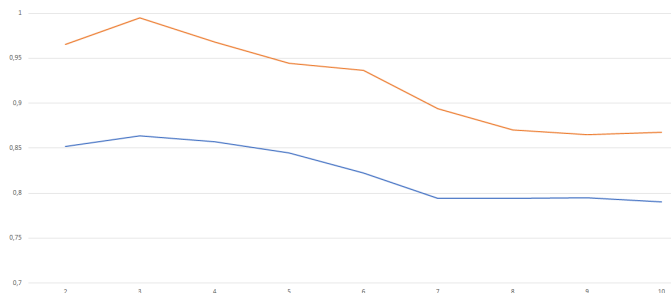


Figure 4: Test (blue) and Training (red) accuracy for increasing number of layers with $N^\ell = 20$ for all $\ell = 1, \dots, L$

The best results correspond to $L = 3$ (two hidden layers). The picture shows that increasing the deepness of the network do not produce better results. This may be due also to the well known difficulties in training such deep networks using gradient based method like Adam. The accuracy and balance accuracy are reported in Table (16)

We also analyze behavior on the target set with only the 14 features selected as described in section 2.6.

For a comparison we report the average confusion matrix obtained in the two case, full versus reduced target set. The results are in Tables 17 and 18 for the two configurations.

From the results we get that the performance decreases with the reduction of the features only by 2 – 3%. This seems to suggest that the 14 features selected bring the most significant information. Indeed the physicians analyzed them and the comments are reported at the end of the section.

We also used the nonlinear SVM classifier as implemented in LIBSVM with the kernel chosen as RBF function with spread γ . The best values of the parameters C and γ appearing in (1) have been set with the tuning procedure

Full / minimal set	Healthy	Scoliotic
Healthy	16.82/16.25 %	2.59/3.55 %
Scoliotic	3.81/4.11 %	23.78/23.09 %

Table 17: Confusion matrix of shallow network with $N = 40$ with all the features and the 14 in the minimal set respectively

Full / minimal set	Healthy	Scoliotic
Healthy	35.77/ 33.62 %	6.15/ 7.43 %
Scoliotic	8.32/ 8.64 %	49.77/ 50.32 %

Table 18: Confusion matrix of deep network with $L = 3$ with all the features and the 14 in the minimal set respectively

as described in section 2.7. The procedure has been replicated both for the full dataset and for the minimal features set. The resulting values of C and γ are reported in Table 19.

The classification accuracy and balanced accuracy reached by the SVM classifier both for the full and the minimal dataset are reported in Table 20.

We also report in Table 21 the confusion matrix of the SVM classifier for the two sets, i.e. full dataset and minimal features set. In both cases, the false positives and false negatives are well balanced, namely the percentage of healthy patients classified as scoliotic is almost equal to the percentage of scoliotic patients classified as healthy.

We observed that a reduction of the input dimensionality by almost 50% brought only a slight deterioration of the accuracy thus confirming the potential role played by the selected features is in line with the results of the DNN classifier. This result could help physicians in the diagnostic process, when it is usually needed to look at multiple different variables and potentially invasive tests (i.e. X-ray) to detect a scoliosis.

	Full	Minimal
C	10	10
γ	10^{-3}	10^{-2}

Table 19: Parameters of SVM defined by the tuning procedure

	ACC	BACC
Full set	84.9%	84.7%
Minimal set	82.2 %	81.5%

Table 20: Accuracy (ACC) and Balanced Accuracy (BACC) of SVM.

Full / Minimal set	Healthy	Scoliotic
Healthy	34.0 / 31.9	7.0 / 9.6
Scoliotic	8.1 / 8.3	50.9 / 50.3

Table 21: Results of the supervised SVM classifier on the full and minimal features set

3.5 Clinical comments

We analyze in this section the 14 features (i.e. rasterstereography parameters) identified by means of the the feature selection process with the aim of verifying their clinical role. Among the identified parameters reported as the first 14 in Table 13 there are measures on the three planes lateral, sagittal and frontal. Five of the selected parameters, identified by th $x_{17}, x_{19}, x_{20}, x_{25}, x_{26}$, are commonly related to the evaluation and diagnosis of scoliosis, in fact lateral deviation and vertebral rotation are well known clinical signs of the disease. Indeed the scoliosis is defined as a lateral curvature of more than 10 degrees as measured by the Cobb Technique on standing anterior posterior radiograph of the spine. Scoliosis is a complex three-dimensional spinal deformity, that forms a complex curve that leads to deformities not only in the coronal plane but in all three planes, which is caused by the self-rotating movement of the spine. An important feature of idiopathic scoliosis deformity is the vertebral axial rotation which accompanies the vertebral lateral deviation. Mechanical interactions within the spine have been implicated in causing vertebral rotation with lateral deviation. This rotation is thought to be significant for initiation and progression of scoliosis. The magnitude of vertebral axial rotation correlates with the lateral deviation of vertebrae from the spinal axis, and the rotation is maximal near the curve apex ([36], [8], [37], [33]).

Nine parameters, identified by th $x_4, x_5, x_7, x_9, x_{10}, x_{11}, x_{13}, x_{15}, x_{16}$, are instead related to the sagittal plane and these results may seem unexpected because the scoliosis is predominantly characterized by alterations on the

frontal and the transversal plane. However recent clinical papers seem to suggest a role of these parameters. Sullivan et al [34] underlined the importance of sagittal plane and the need of a global assessment in the evaluation of scoliosis. They find a strong correlation between scoliosis severity and loss of 3D kyphosis. Increasing severity of coronal plane curvature is associated with a progressive loss of thoracic kyphosis. Moreover previous study have suggested that thoracic hypokyphosis is a primary event in the development of Idiopathic Scoliosis (IS) [32]. However, because of historic restrictions of planar imaging of this multidimensional deformity, there is little information regarding the correlation of thoracic kyphosis with increasing severity of idiopathic scoliosis. Modern, low-radiation-exposure 3D imaging systems have now made routine clinical 3D imaging feasible. These imaging modalities offer the possibility to study the components of the scoliotic deformity in the planes of origin for each vertebra, free of the distortions on 2D images [25]. Sullivan et al [34] found a strong linear correlation between the magnitude of the main thoracic coronal curve and loss of 3D thoracic kyphosis. Three of these sagittal parameters, x_4, x_5, x_{16} , are related to sagittal alignment of the pelvis, but this was expected since an influence of the sagittal parameters of the column in the identification of patients is known. Legaye et al [22] demonstrated the key importance of the anatomical parameter of pelvic incidence in the regulation of the sagittal curves and this is maintained when the scoliosis disease occurs. Moreover Fei Han et al [13] underlined that patients with degenerative scoliosis (DS) may have a higher pelvic incidence, which may impact the pathogenesis of DS. In fact an unbalancing of incidence pelvic should cause scoliosis if the degeneration speed of the two sides differs [24]. The limits of traditional 2D imaging have restricted the evaluation of the adolescent idiopathic scoliosis effects on the sagittal plane, in fact it is impossible to perform simultaneous evaluation of the frontal and sagittal profiles of the spine. Moreover when sagittal evaluation is performed, the patient is asked to put arms forward which is a non-natural position. Since RX evaluation exposes the patients to ionizing radiations, it can not be made frequently in clinical follow up frequency. The rasterstereography provides a three-dimensional reconstruction of the spine curvature and the patient can assume a natural position. Moreover it provides a dynamic evaluation: indeed, in six seconds, twelve frames are recorded and the average results are obtained. So it is particularly suited to the sagittal plane parameters of the spine without any ionizing radiations exposure and any risk for the patient [38]. Summarizing, the features selected by the procedure seems to have some clinical usefulness thus validating the overall learning procedure based on DNN or SVM.

4 Conclusion

In this study it has been showed that both supervised and unsupervised learning using rasterstereography data give high accuracy results in classifying AIS patient versus healthy one. As expected the accuracy is higher in the supervised case. However the use of clustering procedure allowed to group patients in well separated clusters which showed strong intra-group similarity. Th supervised algorithms used. both Deep Networks and SVMs, performed quite well with accuracy over 80%.

Those evidences confirm as data mining can represent a new approach to identify patients with AIS from healthy ones Moreover clustering procedure can represent a useful method to classify AIS patient after rasterstereography collecting the maximum available data for the patient relating them each other in a set of category. Finally our results confirm that a subset of rasterstereography parameters can be used in the screening of AIS patients, although X-ray imaging cannot be replaced by this method.

Rasterstereography tool can be used to perform a scoliosis screening in order to improve the selection of patient that need to underwent X-ray examination. Furthermore thanks to the fact that Rastereography machine is easily carried, it can be used to propose once again scholar screening for pre-adolescent pupils.

Acknowledgment

We thank Dr. Benedetta Zucchi, Dr. Alessandro Cerino and Dr. Sabina Pellanera for their support on medical aspects.

References

References

- [1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.
- [2] Mathias M. Adankon, Jean Dansereau, Hubert Labelle, and Farida Cheriet. Non invasive classification system of scoliosis curve types using least-squares support vector machines. *Artificial Intelligence in Medicine*, 56(2):99–107, 2012.

- [3] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.
- [4] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.
- [5] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.
- [6] Hao Chen, Chiyao Shen, Jing Qin, Dong Ni, Lin Shi, Jack C. Y. Cheng, and Pheng-Ann Heng. Automatic localization and identification of vertebrae in spine ct via a joint learning model with deep neural networks. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 515–522, Cham, 2015. Springer International Publishing.
- [7] François Chollet et al. Keras. <https://keras.io>, 2015.
- [8] J.R. Cobb. Outline for the study of scoliosis. *Instr Course Lect AAOS*, 5:261–275, 1948.
- [9] Michele Morin Doody, John E Lonstein, Marilyn Stovall, David G Hacker, Nickolas Luckyanov, Charles E Land, et al. Breast cancer mortality after diagnostic radiography: findings from the us scoliosis cohort study. *Spine*, 25(16):2052–2063, 2000.
- [10] B. Drerup and E. Hierholzer. Back shape measurement using video rasterstereography and three-dimensional reconstruction of spinal shape. *Clinical Biomechanics*, 9(1):28–36, 1994.
- [11] Pedregosa et al. Scikit-learn: Machine learning in python. *JMLR*, 12:2825–2830, 2011.
- [12] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, 1996.
- [13] Han Fei, Wei-shi Li, Zhuo-ran Sun, Shuai Jiang, and Zhong-qiang Chen. Effect of patient position on the lordosis and scoliosis of patients with degenerative lumbar scoliosis. *Medicine*, 96(32), 2017.

- [14] Bilwaj Gaonkar, David Hovda, Neil Martin, and Luke Macyszyn. Deep learning in the small sample size setting: cascaded feed forward neural networks for medical image segmentation, 2016.
- [15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [16] Gudmund R Iversen, Albert R Wildt, Helmut Norpoth, and Helmut P Norpoth. *Analysis of variance*. Number 1. Sage, 1987.
- [17] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651 – 666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).
- [18] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.
- [19] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *Proc. 3rd Int. Conf. Learn. Representations*, 2014.
- [20] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.
- [21] Vishwajeet Kumar, Ashley Cole, Lee Breakwell, and Antony Louis Rex Michael. Comparison of the diers formetric 4d scanner and plain radiographs in terms of accuracy in idiopathic scoliosis patients. *Global Spine Journal*, 6(1_suppl):s-0036, 2016.
- [22] Jean Legaye, G. Duval-Beaupere, J. Hecquet, and C. Marty. Pelvic incidence: a fundamental pelvic parameter for three-dimensional regulation of spinal sagittal curves. *European Spine Journal*, 7(2):99–103, 1998.
- [23] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.
- [24] Wei-shi LI, Zhuo-ran SUN, and Zhong-qiang CHEN. Radiographic analysis of sagittal spino-pelvic alignment in asymptomatic chinese adults. *Chinese Journal of Orthopaedics*, (5):447–453, 2013.
- [25] Peter O. Newton, Takahito Fujimori, Joshua Doan, Fredrick G. Reighard, Tracey P. Bastrom, and Amirhossein Misaghi. Defining the

- “three-dimensional sagittal plane” in thoracic adolescent idiopathic scoliosis. *JBJS*, 97(20):1694–1701, 2015.
- [26] Po Shun Ngan, Man Leung Wong, Wai Lam, Kwong Sak Leung, and Jack CY Cheng. Medical data mining using evolutionary computation. *Artificial Intelligence in Medicine*, 16(1):73–96, 1999.
- [27] Johnny Padulo and Luca Paolo Ardigò. Formetric 4d rasterstereography. *BioMed research international*, 2014, 2014.
- [28] Laura Palagi. Global optimization issues in deep network regression: an overview. *Journal of Global Optimization*, pages 1–39, 2018.
- [29] Veronica Piccialli and Marco Sciandrone. Nonlinear optimization and support vector machines. *4OR*, 2018.
- [30] Brian C. Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357, 2014.
- [31] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.
- [32] Alastair J. Stirling, Denise Howel, Peter A. Millner, Safa’a Sadiq, David Sharples, and Robert A. Dickson. Late-onset idiopathic scoliosis in children six to fourteen years old.: A cross-sectional prevalence study. *JBJS*, 78(9):1330–1336, 1996.
- [33] Ian AF. Stokes and Mack Gardner-Morse. Analysis of the interaction between vertebral lateral deviation and axial rotation in scoliosis. *Journal of biomechanics*, 24(8):753–759, 1991.
- [34] T. Barrett Sullivan, Fredrick G Reighard, Emily J. Osborn, Kevin C. Parvaresh, and Peter O. Newton. Thoracic idiopathic scoliosis severity is highly correlated with 3d measures of thoracic kyphosis. *JBJS*, 99(11):e54, 2017.
- [35] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.
- [36] O. Yaman and S. Dalbayrak. Idiopathic scoliosis. *Turkish neurosurgery*, 24(5):646–657, 2014.
- [37] Mingyuan Yang, Yuechao Zhao, Xin Yin, Ziqiang Chen, Changwei Yang, Li Li, and Ming Li. Prevalence, risk factors, and characteristics

of the “adding-on” phenomenon in idiopathic scoliosis after correction surgery: A systematic review and meta-analysis. *Spine*, 43(11):780–790.

- [38] F. Zaina, S. Donzelli, M. Lusini, and S. Negrini. How to measure kyphosis in everyday clinical practice: a reliability study on different methods. *Studies in health technology and informatics*, 176:264, 2012.
- [39] Ji Zhu, Saharon Rosset, Robert Tibshirani, and Trevor J Hastie. 1-norm support vector machines. In *Advances in neural information processing systems*, pages 49–56, 2004.