Information Sciences 612 (2022) 514–535

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Automating the decision making process of Todd's age estimation method from the pubic symphysis with explainable machine learning

Juan Carlos Gámez-Granados [a], Javier Irurita [b], Raúl Pérez [c], Antonio González [c], Sergio Damas [c], Inmaculada Alemán [b], Oscar Cordón [c,*]

[a] Dept. Electrical & Computer Engineering, University of Córdoba, 14071 Córdoba, Spain
[b] Physical Anthropology Lab, Dept. of Legal Medicine, Toxicology and Physical Anthropology, University of Granada, 18071 Granada, Spain
[c] Dept. of Computer Science and Artificial Intelligence, and Dept. of Software Engineering, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, 18071 Granada, Spain

ARTICLE INFO

Article history:
Received 2 September 2021
Received in revised form 2 February 2022
Accepted 29 August 2022
Available online 5 September 2022

Keywords:
Forensic anthropology
Skeleton-based age assessment
Explainable artificial intelligence and machine learning
Ordinal classification
Oversampling methods

ABSTRACT

Age estimation is a fundamental task in forensic anthropology for both the living and the dead. The procedure consists of analyzing properties such as appearance, ossification patterns, and morphology in different skeletonized remains. The pubic symphysis is extensively used to assess adults' age-at-death due to its reliability. Nevertheless, most methods currently used for skeleton-based age estimation are carried out manually, even though their automation has the potential to lead to a considerable improvement in terms of economic resources, effectiveness, and execution time. In particular, explainable machine learning emerges as a promising means of addressing this challenge by engaging forensic experts to refine and audit the extracted knowledge and discover unknown patterns hidden in the complex and uncertain available data. In this contribution we address the automation of the decision making process of Todd's pioneering age assessment method to assist the forensic practitioner in its application. To do so, we make use of the pubic bone data base available at the Physical Anthropology lab of the University of Granada. The machine learning task is significantly complex as it becomes an imbalanced ordinal classification problem with a small sample size and a high dimension. We tackle it with the combination of an ordinal classification method and oversampling techniques through an extensive experimental setup. Two forensic anthropologists refine and validate the derived rule base according to their own expertise and the knowledge available in the area. The resulting automatic system, finally composed of 34 interpretable rules, outperforms the state-of-the-art accuracy. In addition, and more importantly, it allows the forensic experts to uncover novel and interesting insights about how Todd's method works, in particular, and the guidelines to estimate age-at-death from pubic symphysis characteristics, generally.

* Corresponding author.
E-mail addresses: jcgamez@uco.es (J.C. Gámez-Granados), javieri@ugr.es (J. Irurita), fgr@decsai.ugr.es (R. Pérez), A.Gonzalez@decsai.ugr.es (A. González), sdamas@go.ugr.es (S. Damas), ialeman@ugr.es (I. Alemán), ocordon@decsai.ugr.es (O. Cordón).

https://doi.org/10.1016/j.ins.2022.08.110
0020-0255/© 2022 The Author(s). Published by Elsevier Inc.

1. Introduction

Personal identity is a key element for the preservation and defense of Human Rights. Forensic anthropology (FA) applies skeletal analysis to medico-legal investigations [1]. The development of precise, fast, robust, and automatic methods for age estimation is currently an area with a strong research interest in FA [2]. Such methods support the forensic practitioner for the correct human identification (ID) of the dead (e.g., disaster victim identification scenarios) and the living (e.g., minors in border control).

There are several accepted methods for age-at-death estimation that are based on the standard bone development and degenerative processes in different skeletal remains [3]. There is a broad consensus in the fact that the analysis of pubic symphysis is especially reliable to assess adult age-at-death [4,5]. This analysis is performed using phase-based methods that estimate the subject's age-at-death from morphological characteristics related to the pubic changes experienced along life. In particular, Todd's method [6] is recognized as the first of these proposed methods. It has constituted the basis for later proposals (such as [7–9]) that have either modified the pubic bone characteristics or the age phases considered. Nevertheless, this family of methods mainly continues to be applied routinely in a similar, if not identical way to that followed when they were proposed. They mainly rely on the experience, expertise, and skills of the forensic practitioner. Hence, skeleton-based age estimation is a relevant but difficult task due to the limitations on the definition of the methods and the current technology [3].

Artificial intelligence (AI) tools can be applied to exploit all the potentials of current ID approaches. Specifically, eXplainable Artificial Intelligence (XAI) [10] can fill the technological gap required to move from current subjective observational approaches to a new paradigm based on accurate, reliable, and understandable ID methods in FA.

Our goal in this contribution is thus twofold. First, we aim to design an explainable rule-based system which automates the decision making process of Todd's pubic symphysis age estimation method. The method's operation will be systematized based on the expert knowledge expressed in the general procedure enunciated by Todd in its seminal paper, thus assisting the practitioner. We also aim to uncover new knowledge from Todd's method itself by combining the expertise of the forensic practitioner and the insights uncovered by the explainable machine learning (ML) approach, explicitly represented in the rule-base description. This will allow us to revisit the method to objectively analyze its operation.

We face the associated ML task as an ordinal classification problem [11]. This is a more natural approach as Todd's age-at-death phases are ranked and an error in an estimated phase is definitively more relevant when the distance with the true phase is greater. We will consider an ordinal classification ML algorithm which has recently reported good performance, New Structural Learning algorithm in a Vague environment for Ordinal classification (NSLVOrd) [12]. The learning task includes additional challenges. First, the ordered set of classes is rather big (10 phases), thus making difficult to achieve high accuracy degrees. Second, it is significantly difficult to obtain an appropriate dataset. We will use an extensive and well distributed pubic bone collection, by the Physical Anthropology Lab at the University of Granada. However, the number of instances is small for a ML method considering the large number of pubic bone traits and classes used. The dataset is also strongly imbalanced as: i) the number of deceased people naturally increases with age; ii) the last degenerative stage is overpopulated since it corresponds to an age interval starting at 50 (as Todd's method was proposed when life expectancy was definitively shorter).

Thus, we must deal with an especially complex ML problem: an imbalanced ordinal classification problem with a small sample size and a high dimension [13,14]. It is well known that minimum-error oriented classifiers tend to ignore the less populated classes, leading to wrong conclusions and deceptive intelligent systems [15]. We will thus apply oversampling methods [14,16] to balance the number of instances in each of the 10 age-at-death phases before applying NSLVOrd.

Both the accuracy and the interpretability of every rule base obtained by NSLVOrd through an extensive experimentation will be evaluated and compared with two widely used ML methods, decision trees (in particular, C4.5 (J48) [17]) and random forests [18]. A deep neural network [19] will also be considered to determine an accuracy threshold on the problem complexity even if this classifier would not be useful for the forensic expert due to its low explainability. Finally, an expert analysis of the obtained explainable rule-based system will be developed by two forensic anthropologists to validate it and evaluate its reliability when compared with previous approaches and existing findings in the area.

2. Background

2.1. Skeleton-based age estimation

Age estimation from the study of bone remains is a fundamental tool for ID in numerous contexts as missing persons, multiple accidents, fires, natural disasters, acts of terrorism, mass graves, crimes of *lessa* humanity, etc. [3,2]. Assessing age of an adult person from skeletal remains is a complex task and there are different methods in FA to be applied depending on the evolution of the normal bone degenerative processes in the individual [3]. The most useful anatomic regions are thus those less affected by external factors as physical activity or job. Age assessment methods based on the pubic symphysis [6,8,9], the auricular surface of the ilium [20], and the sternal end of fourth rib [21] have shown a good accuracy. The analysis of pubic bone development and degenerative processes is the most extended when applicable, being recognized as the most reliable alternative [4,5].

2.2. Todd's family of methods for age estimation

The first studies using the pubic symphysis to estimate age-at-death date from 1920 and correspond to Todd, who analyzed 306 skeletons of Caucasian males in the range of 18–60 years (obtained from autopsies) to establish 10 development and degenerative stages [6]. To do so, Todd defined a morphological description of the pubic symphysis in each of the age-at-death phases and accompanied them with photographs to facilitate comparisons when applying the method.

In general, there is a clear consensus regarding the methodological limitations of Todd's proposal, starting with the reliability of the sample used, composed mostly of unidentified corpses whose age was estimated during the autopsy. In addition, those pubic symphyses that did not meet the expected standards were eliminated from the sample, which distorts the results, showing a lower variability than the actual one. Finally, the obtained results were analyzed only based on his experience, without any statistical analysis, and very narrow age ranges were used for the phases, so even if the method may be precise it is not accurate. Nevertheless, despite these criticisms regarding the design of the method, the criteria that Todd used to describe the degenerative process of the pubic bone more than 100 years ago are still used by the methodology recommended in the area nowadays.

Due to these limitations, numerous studies have tested Todd's method performance and proposed adaptations. Gilbert and McKern [7] adapted the method to work with a continuous age-at-death variable derived from the aggregation of numerical values attributed to each bone trait but at the cost of a lower accuracy. At present, the most accepted adaptation by the scientific community is the Suchey-Brooks' method [8], based on reducing the age-at-death phases to 6, slightly modifying the criteria to define them, and introducing error margins in the phases based on the 95% confidence intervals obtained from the sample of subjects analyzed. Some authors have proposed new population-specific adaptations [9,22,23]. They mainly involve slight modifications of the age estimation intervals or of the pubic bone traits that define each phase. Nevertheless, more recent studies point to the possibility that these population differences are not such and are due to methodological errors in the validation studies [24].

Despite all the validation studies and adaptations made in the last 100 years, Todd's method and its variants continue to be applied routinely in a similar way to that of the original proposal. This means it involves the subjective assessment of the morphological traits of the pubic symphyses and its correspondence with the most similar of the pre-established age-at-death phases, according to the experience of the forensic anthropologist. Hence, the main limitation posed by Todd's method and its subsequent adaptations is the high subjectivity inherent both in the morphological description of the pubic symphyses and in the choice of the corresponding development and degenerative phase [5]. This subjectivity derives mainly from the lack of systematization of the method, which offers overly generic descriptions that make discrimination between phases difficult.

Skeleton-based age assessment methods should offer the same result regardless the observer applying them. However, numerous authors have shown that the methods' effectiveness depends almost exclusively on the expertise of the forensic anthropologist [22,25]. Without despising the necessary experience of the observer, there is a need to use methods which allow us to reduce the subjectivity of these processes, offering measurable tools assisting the decision making. The use of advanced techniques such as AI in FA can benefit the area in terms of the systematization of the methods, thus extending their application by expert forensic anthropologists but also for practitioners. Some methods have shown that the error is significantly reduced when avoiding the subjective phase selection process and analyzing each trait independently [26,27]. In our work we also avoid the subjective process of phase selection by the specialist, with the difference that we do not use transition analysis to evaluate each trait independently, as the aforementioned works do, but instead we opt for the use of explainable ML techniques. It would ease their validation and permit the development of new methods with a higher degree of certainty and discriminant capability. Future works will be necessary to compare the performance of the two different approaches.

Our main goal in this contribution is thus to reduce the existing subjectivity in the decision making of the age-at-death phase thanks to the use of a ordinal classifier learned from examples and refined by forensic anthropology experts. After analyzing the different alternatives, we finally decided to consider the original Todd's method [6] to carry out our study. The reasons underlying this decision is that it is the method that analyzes the highest number of variables (pubic bone traits) and considers the greater number of age-at-death phases. These characteristics make it the most suitable choice for extracting the underlying knowledge on the existing age estimation procedures from the pubic symphysis. Furthermore, this involves starting from scratch from the origin of the problem statement, thus eliminating possible uncontrolled inferences.

2.3. State of the art in computer and artificial intelligence approaches for pubic symphysis-based age assessment

Different computer-based methods have been considered with the goal of objectifying and in some cases automating age estimation methods from the pubic symphysis. The main advances have been aimed at replacing the direct study of dry bone with the use of digitized 3D models, either considering the same criteria used by the original methods [28] or the use of new traits.

A significantly relevant work is [29] where Slice and Algee-Hewitt introduce the SAH score, an index to quantify the variation of the pubic symphysis surface morphology associated with aging in Todd's method. It is computed from a principal component analysis of the laser scans and is used to design a linear regression model evaluated in a study considering 41 modern American male skeletons. It obtains a similar performance to Suchey-Brooks' method, reporting a root mean square error (RMSE) of 17.15 years. Later, in [30], Stoyanova et al. design another linear model that estimates age from a single variable automatically extracted from the pubic symphysis 3D models, the bending energy. This approach is tested with scans from 44 documented white modern American male skeletons and 12 casts representing the Suchey-Brooks phases in the range of 16–100 years, reporting an RMSE of around 19 years. In [31], they extend their previous work by extracting new curvature-based features from the 3D model and integrating all the latter into two multi-variate regression models. The validation over a dataset of 93 white male individuals in 16–90 years (68 individuals and 25 casts) obtains an RMSE between 13.7 and 16.5 years.

Koterova et al. propose 9 automatic regression methods for age estimation jointly considering 3 traits from the pubic symphysis and 4 from the auricular surface of the ilium [32]. The methods are validated on a multi-ethnic dataset of 941 adult individuals in the range of 19–100 years. Classical regression approaches, K-nearest neighbors, Bayesian models, regression trees, and neural networks were considered. Multi-linear regression was the best predictor with an RMSE of 12.1 years and a mean absolute error (MAE) value of 9.7 years.

Dudzik and Langley focus on the uncertainty arising the bone trait definitions and the difficulty of forensic anthropologists to understand them [5]. The bone characteristics in the stages prior to advanced degenerative changes are analyzed in 237 American individuals. Multinomial logistic regression and decision trees are applied over a small sample of 47 young individuals in the range of 18–40 years using 5 pubic symphysis traits. They report a 94% of accuracy but only consider 3 different phases.

In [33], we modeled the problem as a classical classification ML task with 10 independent age-at-death phases. The main morphological traits of the bone described by Todd were analyzed and modeled as 9 linguistic variables based on the forensic expert knowledge. A small set of pubic symphyses from 74 individuals were manually labeled by two different forensic anthropologists. Compact rule bases composed of 17 to 20 rules were obtained using fuzzy decision trees, reporting an ordinal MAE of 1.07 in the whole dataset and of 1.68 in the test set in a 10-fold cross validation. Nevertheless, the derived rules did not cover all the phases. Hence, the validation process was clearly not trustful due to the limitations in the dataset (small and not balanced) and the ML approach followed.

Meanwhile, there are few related studies that apply ML to automate other kinds of skeleton-based age assessment methods. Stepanovsky et al. propose 22 automatic methods for dental age estimation in children and adolescents [34]. The methods are based on the analysis of categorical variables by means of ML or basic mathematical operations. An experimental comparison is made in terms of both accuracy and "usability" to support forensic anthropologists' work. The results obtained on a sample of 976 orthopantomographs from Czech kids show that 3 methods based on multiple linear regressions, regression trees, and support vector machines provided the best accuracy while remaining user-friendly.

Alternatively, in [35] Aja-Fernández et al. deal with the RUS Tanner-Whitehouse 3 method [36] automation by means of fuzzy decision trees. The method explores hand radiographs of pediatric patients and classifies them in 8 or 9 maturity stages, depending on the specific bones considered. Each bone undergoes a shape analysis process by the expert, being individually classified into one or several maturity phases, and the final decision is taken from an overall aggregation of those partial decisions. This is the most similar proposal to the one in the current contribution as the authors take the descriptive nature of the automatic classifier as an advantage to extract the underlying expert knowledge and to refine the predictions.

3. Proposal

3.1. Definition of the machine learning task to be solved to automate Todd's age estimation method

The existing age estimation methods from skeletal remains can be distinguished by either providing a specific age-at-death value or interval as output. To automate any of those methods there is a need to define the appropriate ML task. Component-scoring methods providing an estimated age-at-death value as Gilbert-McKern's [7] must be modeled as a regression problem. Meanwhile, phase-based methods reporting an estimated age-at-death interval as Todd's [6] and Suchey-Brooks' [8] involve a classification problem.

We will consider an advanced approach to automate Todd's method, ordinal classification [11]. The assumption is that the natural ordinal structure of the 10 age-at-death phases can be specifically exploited for the problem solving. We will apply NSLVOrd, an ordinal classification method that has shown good performance (see Section 3.3).

There are some additional issues to be considered. The most important is related to the quality of the dataset. We have available an extensive and well distributed collection of pubic bones (see Section 4.1). Nevertheless, the number of instances is reduced for a ML method considering the 9 traits characterizing the development and degenerative processes in the pubic bone (see Section 3.2) and the 10 output classes. Besides, the dataset is strongly imbalanced. The number of samples in the first phases is significantly smaller than in the last phases, and the last phase naturally includes more than twice as many samples as the previous one. This makes the ML even more complex. The amount of false negatives must be reduced while not increasing excessively the number of false positives. We thus use oversampling [16] to rebalance the number of instances in the 10 phases (see Section 3.4).

Problems like ours that combine imbalanced data, small sample sizes and high dimensionality have been recognized as especially difficult in the specialized ML literature [13]. Learning algorithms often fail to derive general rules due to the difficultly to form conjunctions over the large number of input features with limited samples, leading to the extraction of too specific rules and thus of overfitted rule-based systems [14]. Our age assessment problem is even more complex because of the phase order in the output. A careful ML design like the one proposed is crucial to handle it correctly.

3.2. Identification and modeling of the pubic symphysis' features and categorical values

In [6], Todd provided a really thorough description of the morphological aspects of the pubic symphysis. However, such descriptions are sometimes imprecise and they can lead to confusion between age groups, especially for new practitioners. There are some main traits focused on the development and degenerative changes in the bone but they are not clearly established. For the sake of both objectivity and systematization, we consider 9 variables whose categorical values describe a morphological characteristic of the pubic symphysis analyzed by Todd (see Table 1).

Fig. 1 shows a graphical representation of some of these morphological aspects. The descriptions of the different variables and categorical values are as follows:

- **V1: Articular face** describes the obliteration process of the epiphysis on the symphyseal surface and has been divided into 6 levels according to the presence of ridges and grooves.
- **V2: Irregular porosity** describes the extensive erosion of the bone surface, characterized by the progressive appearance of irregular porosity. Divided into 3 levels.
- **V3: Upper symphysial extremity; V5: Lower symphysial extremity; V6: Dorsal margin;** and **V9: Ventral margin** describe the formation of the upper, lower, dorsal and ventral margins, respectively. Each variable has 2 values for the presence or absence of the respective margin, but variable 9 which includes 3 more levels to describe the degenerative process of the ventral margin.
- **V4: Bony nodule** describes the formation of the ossification nodule in the upper margin. It considers 2 values for presence or absence.
- **V7: Dorsal plateau** indicates the presence or absence of texture difference between the dorsal half and the ventral half. It thus considers 2 values.
- **V8: Ventral bevel** describes the progressive and bevelled elevation of the ventral area. Divided into 3 levels.

3.3. Rule-based learning method considered: NSLVOrd

NSLVOrd [12] is a fuzzy/categorical ordinal rule learning algorithm based on the Iterative Rule Learning approach. It generates rules one by considering the instance covering the rule base already generated produce on the dataset. In each iteration, a genetic algorithm searches in the combinatorial space of tentative rules to derive the best one under the current conditions. An example of a possible rule for our age estimation system would be:

Rule 1: Example of a tentative rule

```
IF "Articular Face" IS 'Grooves Remains' AND "Bony Nodule" IS 'Absent' AND "Dorsal Plateau
    " IS 'Present' AND "Ventral Bevel" IS 'Absent' AND "Ventral Margin" IS ('Absent' OR '
    Partially Formed')
THEN Phase IS Ph05-27-30 with weight = 0.75
```

Table 1
Pubic symphysis' features and categorical values.

	Variable Name	Categorical Values	
V1	**Articular Face**	Regular Porosity	Ridges Formation
		Ridges And Grooves	Grooves Shallow
		Grooves Remains	No Grooves
V2	**Irregular Porosity**	Absence	Medium
		Much	
V3	**Upper Symphysial Ext.**	NotSpecified	Defined
V4	**Bony Nodule**	Absent	Present
V5	**Lower Symphysial Ext.**	NotSpecified	Defined
V6	**Dorsal Margin**	Absent	Present
V7	**Dorsal Plateau**	Absent	Present
V8	**Ventral Bevel**	Absent	In Process
		Present	
V9	**Ventral Margin**	Absent	Partially Formed
		Formed Without Bony Outgrowths	Formed With Few Bony Outgrowths
		Formed With Recesses And Protrusions	

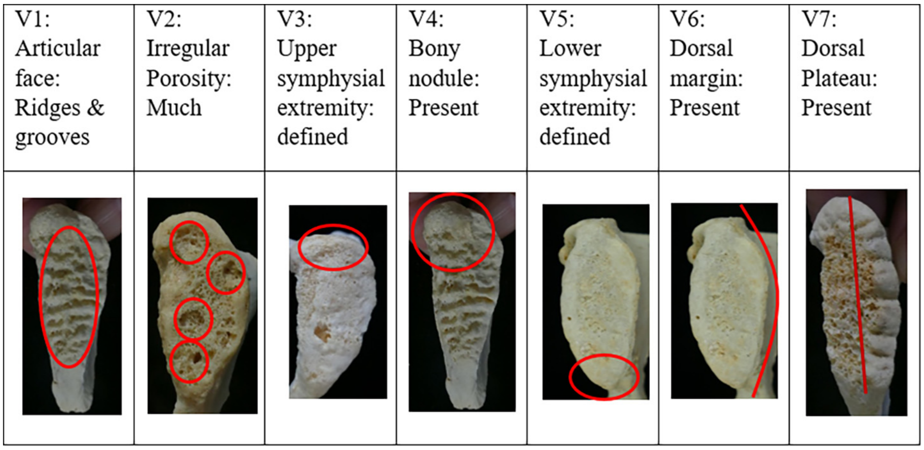
V1: Articular face: Ridges & grooves
V2: Irregular Porosity: Much
V3: Upper symphysial extremity: defined
V4: Bony nodule: Present
V5: Lower symphysial extremity: defined
V6: Dorsal margin: Present
V7: Dorsal Plateau: Present

Fig. 1. Some example of traits defining the typical development and degenerative morphological changes in the pubic symphysis during growth. Regions of interest are highlighted in red.

Each candidate rule is represented as an individual in the genetic population. The coding scheme has 3 different parts. The real-coded "variable part" is used to define the variables in the rule antecedent (e.g. **Articular Face**). The binary-coded "value part" represents the specific values for the selected variables (e.g. *Grooves Remains*). The integer-coded "consequent part" encodes the rule consequent (e.g. *Ph05-27–30*). NSLVOrd uses an embedded feature selection mechanism at rule level by means of the "variable part". An information measure is used to determine an initial relevance of each variable in relation to the consequent. A numerical threshold (that also undergoes evolution) determines the variables that will eventually appear in the rule, that is, all the variables whose associated value in the "variable part" is greater o equal to the threshold value will be considered in the description of the rule. The remainder will be considered irrelevant.

Fig. 2 illustrates a possible encoding for Rule 1. The "variable part" is composed of 9 values, one for each variable involved in the problem. The order of the variable coding is the same showed in Table 1 (1-Articular Face, 2-Irregular Porosity, ...). Only 5 variables present a relevance greater than the threshold value (0.21). These variables are thus selected: the first (**Articular Face**), fourth (**Bony Nodule**), and seventh to ninth (**Dorsal Plateau, Ventral Bevel, Ventral Margin**). The "value part" encodes the categorical values for these variables. The first of them is **Articular Face** and its associated value in the "value part" is 000010, where each bit represents each value it can take and an 1 indicates it is actually taken. Thus, the domain of **Articular Face** has six categorical values and only the fifth (*Grooves Remains*) is taken. When more than one value is taken, the rule considers the OR operation over the values (as for **Ventral Margin**). Finally, the "consequent part" (with value 4) is representing the fifth phase (*Ph05-27–30*) since the 10 age-at-death phases are encoded in $\{0, \ldots, 9\}$.

The weight represents the confidence of the rule and it is calculated as $\frac{n^{+}(R)}{n(R)}$, with $n^{+}(R)$ and $n(R)$ being the number of instances from the training set that are correctly classified and the total number of instances fired by the rule, respectively.

The classifiers derived from NSLVOrd always return an output value by considering a default rule. It is used when no other rule in the rule base is fired. We have chosen to set the default rule to the majority class and to add this rule at the end of the learning process. When the dataset is balanced, the intermediate class in the order is considered instead.

Several metrics from the specialized literature in ordinal classification have been considered [11,14]. For a better understanding, we start from a standard confusion matrix:

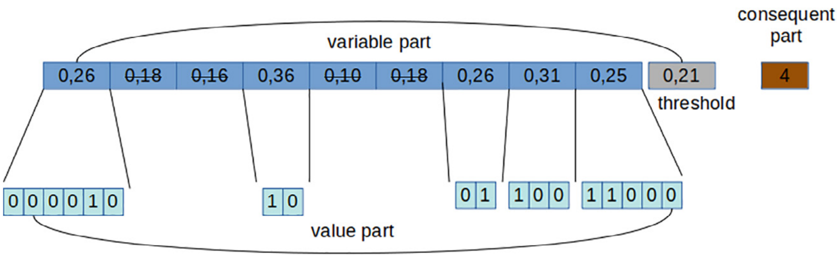

Fig. 2. Coding scheme for the example rule.

$$\begin{array}{cc|ccccc|c}
 & & \multicolumn{5}{c}{\text{predicted class}} & \\
 & & 1 & \cdots & j & \cdots & Q & \\ \hline
 & 1 & n_{11} & \cdots & n_{1j} & \cdots & n_{1Q} & n_{1\bullet} \\
 & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\
\text{actual class} & i & n_{i1} & \cdots & n_{ij} & \cdots & n_{iQ} & n_{i\bullet} \\
 & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\
 & Q & n_{Q1} & \cdots & n_{Qj} & \cdots & n_{QQ} & n_{Q\bullet} \\ \hline
 & & n_{\bullet 1} & \cdots & n_{\bullet j} & \cdots & n_{\bullet Q} & N
\end{array}$$

of a classification problem with Q classes and N instances; n_{ij} is the number of instances of the $i-th$ class which have been predicted as belonging to the $j-th$ class by the classifier; n_{i*} is the number of instances of the $i-th$ class; and n_{*j} is the number of instances classified as the $j-th$ class.

- *Accuracy* or *Correct Classification Rate (CCR)*: rate of correctly classified instances in the dataset: $CCR = \frac{1}{N}\sum_{i=1}^{Q} n_{ii}$.
- *Precision*: Rate of i-th class instances correctly classified considering the predicted instances as the i-th class: $Precision_i = \frac{n_{ii}}{n_{\bullet i}}$.
- *Recall*: rate of i-th class instances correctly classified considering the actual instances in i-th class: $Recall_i = \frac{n_{ii}}{n_{i\bullet}}$.
- *F1-Score*: balance between Precision and Recall, thus also defined in [0,1]: $F1 - Score_i = \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i}$.
- *Ordinal Mean Absolute Error (OMAE)*: average deviation in absolute value between predicted and actual class, thus defined in [0,Q-1]: $OMAE_i = \frac{1}{n_{i\bullet}}\sum_{k=1}^{Q}|i-k|n_{ik}$.

Precision, *Recall*, *F1-Score* and *OMAE* can be computed over the whole dataset as an average of the values for each class, either weighted by the number of instances in each class or not. They can also be computed over the whole dataset but then they take the same value and are equivalent to the *CCR*.

Ordinal classifiers can also be evaluated using regression metrics, such as the following ones, where $\hat{Y}_i$ is the predicted value (usually taken as a representative value from the elements in the corresponding class) and Y_i is the actual value:

- *Mean Absolute Error (MAE)*: average deviation in absolute value between predicted values and actual values: $MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{Y}_i - Y_i|$.
- *Root Mean Square Error (RMSE)*: square root of the average deviation in squared differences: $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2}$.

3.4. Imbalanced classification problem issues: oversampling

There is an extensive list of strategies to deal with imbalanced classification [14]. The complexity of the current ML task advises against the use of complex and specific imbalanced learning approaches. In our case, the use of oversampling based on generating a balanced dataset by adding instances from the minority class seems to be the best alternative. There are many different oversampling variants [14,16]. Synthetic Minority Oversampling Technique (SMOTE) [37] is the *de facto* standard. It generates new instances of the minority classes by interpolating the feature values from selected neighboring instances of the respective class. The resulting balanced dataset, both composed of the original real instances and some new artificial but real-like instances, can improve the generalization capability of the classifier.

We will use 7 different oversampling methods [16]. First, two simple alternatives based on the uniform (basic oversampling) and random replication (random oversampling) of the existing instances in each of the first 9 age-at-death phases until obtaining a balanced distribution with respect to the number of instances of the majority class (phase Ph10-50+). The other methods include the original SMOTE and 4 variants: Support Vectors SMOTE (SVMSMOTE), Borderline SMOTE (Borderline-SMOTE), ADAptive SYNthetic sampling approach for imbalanced learning (ADASYN), and Kmeans-SMOTE. Their composition is out of the scope of this contribution and the interested reader is referred to [16] for specific details. As they deal with numerical variables, the original variable values are first transformed into an integer representation, the oversampling technique is run, and the inverse transformation is applied to achieve the artificially generated nominal value. The implementation used is available at https://github.com/scikit-learn-contrib/imbalanced-learn (scikit-learn Python ML package developed at INRIA).

4. Experiments and analysis of results

4.1. Dataset: The Pubic Symphysis Collection at the University of Granada Physical Anthropology Lab

The sample considered is part of the collection of skeletonized pubic symphyses available in the Physical Anthropology lab at the University of Granada, Spain. It results from autopsy studies developed since 1991 under a collaboration with the Institute of Legal Medicine and Forensic Sciences: 837 individuals in the range of 17–82 years, 197 women and 637 men (the other 3 have not been identified). Detailed information is available regarding sex, age, cause of death and, in many cases, additional information such as approximate weight, alcohol or drug consumption, and population origin, which may be important for future studies. The collection was acquired following the criteria of the Ethics Committee of the University of Granada.

To avoid sex bias in our study, we restrict the sample to the 637 males, since this is the majority sex group and Todd's method was originally proposed for men [6]. We consider incomplete or unreliable *ante mortem* information and a deficient state of conservation as additional exclusion criteria, reducing the sample to 566 individuals (1,127 adding left and right sides, with a few individuals lacking one piece).

The pubic symphysis traits and categorical labels in Section 3.2 have been considered by two forensic anthropologists to label each pubis, thus building the dataset used by the ML methods. They first practiced with a total of 200 pubic symphyses, both to avoid the bias associated with the acquisition of habits during data collection and to identify possible refinements in the definitions. After that, they labeled the whole sample, including both pubis sides, always in random order and without knowing the real age-at-death of the individual.

4.2. Experimental setup

We restrict the age range of the sample to 18–60 years, a similar interval to the one used by Todd. The final dataset is thus composed of 439 samples of left laterality and 453 samples of right laterality with the distribution shown in Table 2. The strong imbalance ratio of the minority class (Ph04-25-26) can be clearly observed.

We first develop in Sections 4.3 and 4.4 two different ML experiments (with and without oversampling) by considering the dataset with instances of one laterality as training set and that of the other laterality as test set and *vice versa*. Apart from being sensible from a ML perspective, this separation of the analysis by laterality is justified by possible differences in the development and degenerative processes of the right and left pubic symphyses due to genetic determinants, biomechanical factors, and environmental stress, as observed in previous studies [38].

Later, in Section 4.5 we develop another experiment with the whole original dataset and some oversampled ones in a 5×2-cross validation (5×2-cv) [39]. This test is based on randomly partitioning the dataset into two halves of the same size five different times, ensuring a balanced partition of the instances for each phase. Each half is respectively used for training and test in two different experiments, and the test results from the 10 experiments are averaged.

We have used the NSLVOrd implementation available at https://isg.ugr.es/descargas/ with the default parameters defined in [12]. Three extended ML approaches are considered for benchmarking. The Weka [40] implementation of the C4.5 (J48) algorithm [17] is used to derive different decision tree-based classifiers, considering tree pruning. Several random forests [18] are also generated using Weka with 10, 50, and 100 trees, each one built using 4 random features. Notice that, a small number of trees (10) has been considered to reduce the resulting number of rules in order to preserve interpretability. A deep neural network [19] is also considered to estimate a quality threshold for the problem. We have followed the *one-hot* or *dummy encoding* [41], where each column of the input dataset was transformed into as many columns as categorical values that pubis characteristic has associated (see Section 3.2). A -1 value is assigned to every categorical value but the correct one that gets a $+1$. Likewise, the output class is transformed into an integer output in $\{0, \dots, 9\}$. A fully connected network is considered with two hidden layers of size 64 without bias using the LeakyReLU activation function with a negative slope of 0.01. The network is trained for 64 epochs with Adam as optimizer [19] (chapter 8), $\beta_1 = 0.9, \beta_2 = 0.999$, and learning rate of 10^{-3}. The cost function directly takes the MAE metric, thus solving a regression problem. Once the network is trained, its output is approximated to the nearest integer in $\{0, \dots, 9\}$ to calculate the different metric values. The Pytorch deep learning library in Python (https://pytorch.org/) has been used.

The system accuracy is evaluated using ordinal (CCR, OMAE, F1-Score) metrics (see Section 3.3). For the interpretability, we analyze the complexity of the rules (apart from the final expert analysis in Section 5). The decision trees derived by random forests are transformed into the corresponding sets of rules. The ensemble nature of random forests provides additional

Table 2
Final distribution of the dataset considered, including the available pubic bone lateralities.

	Ph1	Ph2	Ph3	Ph4	Ph5	Ph6	Ph7	Ph8	Ph9	Ph10	Tot.
Left	14	10	11	9	32	29	75	56	65	138	*439*
Right	12	10	13	10	35	31	75	57	69	141	*453*
Both	26	20	24	19	67	60	150	113	134	279	*892*

degrees of freedom to the model, increasing its accuracy but significantly reducing its interpretability with respect to our approach.

4.3. Todd's automatic age assessment learning task with independent pubis laterality and not considering oversampling

Table 3 reports the results obtained by NSLVOrd considering one laterality to generate the rule set and the other to test the accuracy of the system. The best values in each table are highlighted in bold. The performance of the automatic age estimation systems derived from each laterality is similar, with a slight advantage for the left one. Both rule bases show a small complexity, 17 and 24 rules, allowing an easy interpretation by the forensic anthropologists. Table 4 and Table 5 respectively include the results obtained by C4.5 and random forests with 10 trees over the two datasets. Training with right laterality shows a better test performance in those cases. The accuracy of C4.5 is similar to that of NSLVOrd, being worse in the left side training dataset and slightly better in the right side training dataset. The number of rules is also similar: 17 and 24 rules in the NSLVOrd classifiers vs. 21 and 13 rules for C4.5. As expected, a higher test accuracy is obtained by random forests at the cost of generating more complex classifiers comprised by a significantly larger number of rules (936 and 837 rules, between 40 and 50 times higher than NSLVOrd). Even so, the accuracy of the NSLVOrd is still very competitive with respect to that of random forests.

Nevertheless, a more detailed analysis shows that the error metrics are leading us to draw a deceptive conclusion. The obtained rule bases are not covering some of the 10 Todd's method phases. This issue is clearly recognized in the confusion matrix reported in Table 6 for phases Ph04-25–26, Ph06-31–34, and Ph08-40–44. It is a consequence of the dataset imbalance that the learning method tries to compensate both ignoring some of the phases and including a default rule on the majority class. Note that the decision tree-based classifier in Table 7 is even less reliable as it does not cover 4 of the 10 phases: Ph03-22–24, Ph04-25–26, Ph08-40-44, and Ph09-44-50. Finally, the random forest behaves almost in the same way (see Table 8) as, even if it does not show any "empty" phase thanks to having around 900 rules, several phases are strongly under-represented and the obtained rule-based system is also extremely biased towards the last phase. The same situation stands for the other laterality. Hence, we confirm this undesired behavior is not a consequence of the learning method used and the number of rules derived but of the dataset composition.

In summary, the global error could be competitive, but the obtained rules are not reliable since they do not properly cover the input space due to the class imbalance problem.

4.4. Todd's automatic age assessment learning task with independent pubis laterality and incorporating oversampling

The oversampling methods described in Section 3.4 have been run with the same random seed and the default parameters to obtain balanced datasets. Tables 9 and 10 include the results obtained by applying NSLVOrd on the oversampled datasets for the left and right laterality, respectively, considering the original dataset of the other laterality for validation.

As expected, each of the 10 age-at-death phases is now covered by at least one rule for both lateralities in every rule base. For the case of the left laterality, the number of rules generated is around the double for most of the oversampling methods. In the right laterality, the increase is a little bit smaller but still significant.

All the oversampling methods improve the original values in every metric for both laterality datasets in Table 3. Borderline-SMOTE and SVMSMOTE achieve very similar, highly accurate results in the three metrics. Random oversampling performs remarkably well as it even achieves the best value in the F1-Score for both lateralities overall.

However, the accuracy does not seem to be as good in the test sets. Every oversampling variant reports low values of the CCR and OMAE metrics. Although this could imply the presence of some overfitting, it is probably due to the use of the default rule. The ranking of methods is pretty similar to the one obtained in training for both laterality datasets. SVMSMOTE is again the best performing method, followed by Borderline-SMOTE, while the basic oversampling shows the worst results.

Nevertheless, the situation changes in the third metric, F1-Score, for which the original classifiers are outperformed by almost every oversampling method. This results from the fact that this measure clearly accounts for the wrong decisions taken when applying the default rule and thus it uncovers the wrong behavior of the classifiers obtained from the original datasets. In the left laterality, only ADASYN obtains a worse value than the original system (0.211 vs. 0.225) while the best results are obtained by Kmeans-SMOTE and random oversampling (0.28 and 0.264). In the right laterality, the basic oversampling and SMOTE are the only two methods worsening the original system result (0.207 and 0.213 vs. 0.219) while Borderline-SMOTE and, curiously, ADASYN obtain the best values (0.264 both). Focusing on the latter method, its perfor-

Table 3
Results obtained by **NSLVOrd**.

	CCR	F1-Score	OMAE	#rules
Training: left side	**0.412**	0.307	**1.478**	17
Test: right side	**0.369**	**0.225**	**1.545**	–
Training: right side	0.402	**0.332**	1.572	24
Test: left side	0.367	0.219	1.631	–

Table 4
Results obtained by **C4.5**.

	CCR	F1-Score	OMAE	#rules
Training: left side	**0.399**	0.239	1.544	21
Test: right side	0.358	0.185	1.576	–
Training: right side	0.393	**0.282**	**1.468**	13
Test: left side	**0.383**	**0.231**	**1.54**	–

Table 5
Results obtained by **random forests** with 10 trees.

	CCR	F1-Score	OMAE	#rules
Training: left side	0.444	0.385	1.358	936
Test: right side	0.322	**0.288**	1.536	–
Training: right side	**0.45**	**0.434**	**1.278**	837
Test: left side	**0.362**	0.234	**1.506**	–

Table 6
Confusion matrix of the automatic age assessment system obtained by **NSLVOrd** on the **left laterality dataset**.

		Predicted phase										
		Ph1	**Ph2**	**Ph3**	**Ph4**	**Ph5**	**Ph6**	**Ph7**	**Ph8**	**Ph9**	**Ph10**	**Total**
Actual phase	**Ph1**	**12**	2	0	0	0	0	0	0	0	0	14
	Ph2	2	**6**	2	0	0	0	0	0	0	0	10
	Ph3	4	1	**5**	0	0	0	1	0	0	0	11
	Ph4	0	0	0	**0**	1	0	2	0	0	6	9
	Ph5	0	0	0	0	**5**	0	3	0	0	24	32
	Ph6	0	0	0	0	0	**0**	4	0	0	25	29
	Ph7	0	0	0	0	0	0	**18**	0	0	57	75
	Ph8	0	0	0	0	0	0	8	**0**	0	48	56
	Ph9	0	0	0	0	0	0	4	0	**1**	60	65
	Ph10	0	0	0	0	1	0	3	0	0	**134**	138
	Total	18	9	7	0	7	0	43	0	1	354	**439**

Table 7
Confusion matrix of the automatic age assessment system obtained by **C4.5** on the **left laterality dataset**.

		Predicted phase										
		Ph1	**Ph2**	**Ph3**	**Ph4**	**Ph5**	**Ph6**	**Ph7**	**Ph8**	**Ph9**	**Ph10**	**Total**
Actual phase	**Ph1**	**12**	2	0	0	0	0	0	0	0	0	14
	Ph2	4	**6**	0	0	0	0	0	0	0	0	10
	Ph3	5	5	**0**	0	0	0	1	0	0	0	11
	Ph4	1	0	0	**0**	0	0	2	0	0	6	9
	Ph5	1	0	0	0	**5**	0	1	0	0	25	32
	Ph6	0	0	0	0	0	**0**	3	1	0	25	29
	Ph7	1	0	0	0	1	1	**15**	0	0	57	75
	Ph8	0	0	0	0	1	2	5	**0**	0	48	56
	Ph9	0	0	0	0	2	0	2	0	**0**	61	65
	Ph10	0	0	0	0	1	1	2	0	0	**134**	138
	Total	24	13	0	0	10	7	29	0	0	356	**439**

mance appears to be highly dependent on the composition of the dataset, as it shows significantly worse results on the left laterality than on the right laterality. The same happens for SMOTE but in the opposite direction.

4.5. Todd's automatic age assessment learning task with both pubis lateralities and incorporating oversampling

We conclude that oversampling is actually required for our ML task and that the system accuracy could be competitive. To confirm this assumption, we design a final experiment with the whole pubis dataset. This makes sense from both the ML and the FA viewpoints as the application guidelines of Todd's method are independent of the specific pubic symphysis laterality. NSLVOrd, random forests, and a deep neural network [19] are run over the original dataset and two oversampled datasets generated by Borderline-SMOTE (the best performing method, also with a moderate number of rules) and random oversam-

Table 8
Confusion matrix of the automatic age assessment system obtained by **random forests** on the **left laterality dataset**.

		Predicted phase										
		Ph1	**Ph2**	**Ph3**	**Ph4**	**Ph5**	**Ph6**	**Ph7**	**Ph8**	**Ph9**	**Ph10**	**Total**
Actual phase	**Ph1**	**13**	0	1	0	0	0	0	0	0	0	14
	Ph2	4	**6**	0	0	0	0	0	0	0	0	10
	Ph3	4	2	**4**	1	0	0	0	0	0	0	11
	Ph4	0	0	0	**2**	1	0	1	0	1	4	9
	Ph5	0	0	0	0	**8**	0	2	0	2	20	32
	Ph6	0	0	0	1	0	**4**	1	0	0	23	29
	Ph7	0	0	0	1	3	1	**21**	0	2	47	75
	Ph8	0	0	0	1	1	2	4	**1**	5	42	56
	Ph9	0	0	0	1	4	0	3	0	**9**	48	65
	Ph10	0	0	0	2	2	2	2	0	3	**127**	138
	Total	21	8	5	9	19	9	34	1	22	311	**439**

pling (only considering original instances, with a good accuracy and a reasonable number of rules). C4.5 is not considered in this comparison as its performance was inferior to NSLVOrd in the previous experiment in Section 4.3 and random forests clearly showed an upper quality threshold for a decision tree-based classifier. Besides, even if we should keep the derived random forest model simple for interpretability preservation, three different results are reported for this classifier considering an incremental complexity with 10, 50, and 100 trees.

The 5 × 2-cv test results in Table 11 are alike those obtained until now, thus reinforcing the conclusions already drawn after performing a sounder experimental design. The test values for the original dataset are very similar to those in Table 3. There is a significant accuracy increase when considering oversampling. Random forests reports a slightly better accuracy than NSLVOrd but at the cost of a huge interpretability decrease: 1091 vs. 34.8 rules when considering the simplest model with 10 trees, increasing this number up to ten times more when using 100 trees. What is more, the complexity increase in the random forest classifiers does not result in any significant performance improvement, even showing a slight overfitting in some cases. This makes sense as we are applying a ML model with a high number of freedom degrees on a small dataset, as analyzed in Section 3.1. Meanwhile, the differences in the metric values between using Borderline-SMOTE and random oversampling for NSLVOrd with respect to random forests are also small, in the ranges of 6.28% for CCR, 5.88% for F1-Score, and 3.89% for OMAE. The network performs similarly to random forests, slightly better in OMAE (as expected since its learning process is guided by MAE) and slightly worse in the other two metrics. NSLVOrd keeps on being competitive, even improving the deep neural network in CCR and F1-Score in several cases (of course not in OMAE), despite being the model with less freedom degrees and thus the most explainable as desired.

We can thus conclude that NSLVOrd with oversampling shows a good accuracy and provides an explainable model. Since random oversampling works with real-world instances, being more confident for the forensic experts, we will take this as the final design for our automatic age assessment system as a trade-off between accuracy, interpretability, and reliability.

5. Expert validation and knowledge extraction from the automatic Todd's age assessment system

In this section we will develop a human-based reliability study of the intelligent system derived using our ML approach. To do so, we design three different variants using NSLVOrd over three oversampled datasets obtained with random oversampling (left laterality, right laterality, and both lateralities instances). Since the guidelines of Todd's method are independent from the specific laterality of the pubic symphysis, the three options are valid from the FA viewpoint.

The automatic age assessment system providing the best accuracy has been selected. That corresponds to the right laterality dataset (OMAE = 1.214), with slight difference with the left laterality one (OMAE = 1.254) and a larger difference with the whole dataset (OMAE = 1.408). The expert validation has been developed by two different forensic anthropologists from the Physical Anthropology lab at the University of Granada. They first made a thorough analysis of the obtained rules independently and later discussed their individual conclusions to reach the common agreements presented in this section.

Note that an alternative procedure to obtain a rule-based description of Todd's method operation would be to provide the forensic anthropologists with the pubic bone characteristics in Section 3.2 and ask them to directly explicit their expert knowledge. That was done in [33] and it did not work since it was difficult for the forensic experts to define the rules. They tended to build complete rules considering every variable and the obtained system was thus complex. What is more, the intelligent system did not properly cover all the instances, thus becoming deceptive and not reliable. The approach provided in the current contribution, making use of explainable ML to derive an initial rule base and involving the expert in the loop to refine it [10], is significantly more effective.

Table 9
Results of the **oversampling methods with NSLVOrd** for the **left laterality** dataset.The metric values are shown following a color scale (green = best values, red = worst values) Orig = Original; Bas = Basic; Rand = Random; SMT = SMOTE; SVM-S = SVM-SMOTE, BL-S = BorderLine-SMOTE; AD = ADASYN; K-S = KMeans-SMOTE

Metric	Orig	Bas	Rand	SMT	SVM-S	BL-S	AD	K-S
Numb. rules	17	35	40	39	29	33	31	31
Training errors over the same laterality oversampled dataset								
CCR Train	0.412	0.414	0.496	0.484	0.522	0.509	0.452	**0.56**
F1-Score Train	0.307	0.356	**0.384**	0.375	0.357	0.366	0.328	0.347
OMAE Train	1.478	1.254	1.097	1.139	1.052	1.057	1.213	**0.927**
Test errors over the other laterality not oversampled dataset								
CCR Test	**0.369**	0.141	0.216	0.223	**0.263**	0.252	0.234	0.208
F1-Score Test	0.225	0.238	0.264	0.243	0.242	0.247	0.211	**0.280**
OMAE Test	**1.545**	2.179	1.909	1.863	**1.742**	1.808	1.883	1.799

Table 10
Results of the **oversampling methods with NSLVOrd** for the **right laterality** dataset.The metric values are shown following a color scale (green = best values, red = worst values) Orig = Original; Bas = Basic; Rand = Random; SMT = SMOTE; SVM-S = SVM-SMOTE, BL-S = BorderLine-SMOTE; AD = ADASYN; K-S = KMeans-SMOTE.

Metric	Orig	Bas	Rand	SMT	SVM-S	BL-S	AD	K-S
Numb. rules	24	35	35	44	30	35	36	38
Training errors over the same laterality oversampled dataset								
CCR Train	0.402	0.428	0.460	0.443	0.468	**0.474**	0.452	0.464
F1-Score Train	0.331	0.338	**0.370**	0.342	0.351	0.366	0.357	0.344
OMAE Train	1.572	1.214	1.204	1.166	1.245	1.166	1.196	**1.15**
Test errors over the other laterality not oversampled dataset								
CCR Test	**0.367**	0.191	0.239	0.205	0.257	0.251	0.255	0.203
F1-Score Test	0.219	0.207	0.234	0.213	0.255	**0.264**	**0.264**	0.252
OMAE Test	**1.631**	2.144	2.098	2.071	2.055	2.089	2.055	2.062

Table 11
Results of the 5 × 2-cv test.

Original	**CCR**	**F1-Score**	**OMAE**	**#rules**
NSLVOrd	**0.352**	0.212	1.576	27.6
Random forests 10	0.337	0.222	1.580	915
Random forests 50	0.336	0.229	1.557	4749.4
Random forests 100	0.332	0.224	1.571	9433.3
Deep neural network	0.245	**0.231**	**1.282**	–
Random	**CCR**	**F1-Score**	**OMAE**	**#rules**
NSLVOrd	0.373	0.336	1.337	**39**
Random forests 10	**0.432**	**0.414**	1.271	1134
Random forests 50	**0.432**	0.413	1.280	5771.1
Random forests 100	**0.432**	0.412	1.270	11593.8
Deep neural network	0.340	0.334	**1.148**	–
BorderLine	**CCR**	**F1-Score**	**OMAE**	**#rules**
NSLVOrd	0.398	0.357	1.285	**34.8**
Random forests 10	0.437	**0.422**	1.156	1091
Random forests 50	**0.439**	0.421	1.147	5415.9
Random forests 100	**0.439**	0.420	1.146	9433.3
Deep neural network	0.363	0.358	**1.063**	–

5.1. Human expert intervention to derive and validate Todd's automatic age assessment method

The final rule base is derived from an iterative human loop in the XAI process both involving the AI and forensic experts. The original rule base obtained from NSLVOrd was composed of 35 rules. The forensic anthropologists were asked to evaluate these rules following a "traffic light" procedure involving tagging each rule as green, yellow, or red according to its coherence. The green tag stands for those rules that are fully consistent with expert knowledge. Alternatively, the yellow tag stands for those rules showing a general coherent composition but including any inappropriate or inaccurate characteristic. Finally, the red tag is for not meaningful rules.

The forensic experts identified 20 rules as green, 10 as yellow, and 4 as red, apart from the default rule (R34). Note that, even if the rules were presented grouped by age-at-death phase, the rule number (starting at R0) stands for the order in which they were generated. That becomes an additional indicator of the confidence of the rule as more important rules are generated earlier. The red rules were those obtained at the late stages of the learning process: R29, R30, R31, and R33. The first 3 of them were extremely specific and only covered 2 and 4 instances from the dataset. The latter was too generic, covering 783 instances but with a high number of negative instances (418), and acted as a modifier of the default rule. The yellow rules were mostly due to the fact that some consecutive categorical values have been selected for any variable but leaving a gap in the natural order, as in R2, R4, R8, and R23 (e.g. in R4, **Articular Face** was originally assigned *Ridges Formation* OR *Grooves Shallow* but not the intermediate value *Ridges and Grooves*). The existence of rules with these traits contradicts the continuous nature of the development and degenerative process that is being evaluated. This is a consequence that NSLVOrd does not explicitly require including the intermediate values.

The AI experts manually refined the rule base by adding/removing categorical values which make the rule more coherent but did not modify the instances covered and thus the inference process. Besides, they removed R33 as it did not show any specific influence on the classifier operation. The final rule base was evaluated again by the forensic experts resulting the 29 green and 4 yellow rules (R10, R14, R24, and R29), plus the default rule, collected in the Appendix.

We discuss the coherence of these rules in depth as follows:

- The main conclusion is that 29 rules are fully coherent with the available forensic expert knowledge about pubic symphysis development process, thus showing the good behavior of the explainable ML procedure applied.
- Every age-at-death phase is properly represented in the rule base, ranging from a minimum of 2 rules for Ph03-22–24 and Ph07-35–39 to 5 rules for Ph02-20–21.
- Rules R14 and R24 use **Dorsal Plateau** = *Present* as criterion to assign phase Ph09-45–49. This trait should appear in young individuals and disappear after the age of 30. However, the dorsal platform is a confusing characteristic since it can be observed in both young and older people (see Section 5.3). That is probably a consequence of this trait not having a concrete definition in Todd's guidelines.
- Rules R10 and R29 serve as examples in which we observe small inconsistencies according to the normal processes of the pubic bone, such as the not defined lower extremity of the symphysis in phase Ph04-25–26 (R10) or the absent ventral margin in Ph07-35–39 (R29). Rules like these are beneficial as they identify "unusual" pubic symphyses in the sample, which probably present alterations produced by pathological conditions or certain habits such as drug addiction, alcoholism, extreme values in body mass index, etc. Note that, these rules are only covering 14 and 2 instances, respectively, with a perfect matching. They can be manually reviewed in the future, under careful expert evaluation, to refine the automatic age assessment method using expert knowledge.

Table 12 reports the confusion matrix of our classifier. All the age-at-death phases are properly covered. It shows a well balanced behavior, with phases Ph05-27–30 and Ph09-44–50 being the most represented. Fig. 3 collects the number of instances classified by each rule. The default rule, associated to phase Ph05-27–30, is never fired. Rules R1, R19 and R33 are also not triggered as the instances covered by them are also covered by other rules with a higher confidence. The confusion matrix of the random forests classifier with 10 trees in Table 13 shows a similar behavior, equal in the first two phases and with the main difference of having phase Ph04-25-26 as the most populated instead of phase Ph05-27-30. This is a significant behavior keeping in mind that our system only has 34 rules while the random forests classifier has 1034 rules.

Table 14 shows very similar values in the different performance metrics with respect to the random forest (Table 15), even if the latter considers more than 1000 rules. Both classifiers achieve the same result in F1-Score, NSLVOrd is slightly worse value in CCR, and it reports a slightly better value in OMAE, the most trustworthy metric.

Table 12
Confusion matrix of the automatic age assessment system over the **original right laterality** dataset using **NSLVOrd**.

		Predicted phase										
		Ph1	**Ph2**	**Ph3**	**Ph4**	**Ph5**	**Ph6**	**Ph7**	**Ph8**	**Ph9**	**Ph10**	**Total**
Actual phase	**Ph1**	**10**	2	0	0	0	0	0	0	0	0	12
	Ph2	3	**7**	0	0	0	0	0	0	0	0	10
	Ph3	2	2	**8**	1	0	0	0	0	0	0	13
	Ph4	0	0	0	**5**	5	0	0	0	0	0	10
	Ph5	0	0	0	5	**26**	1	0	0	2	1	35
	Ph6	0	0	0	3	19	**4**	0	0	4	1	31
	Ph7	0	0	0	13	28	0	**2**	2	24	6	75
	Ph8	0	0	0	3	28	1	0	**3**	17	5	57
	Ph9	0	0	0	0	27	0	0	0	**39**	3	69
	Ph10	0	0	0	4	47	0	0	5	57	**28**	141
	Total	15	11	8	34	180	6	2	10	143	44	**453**

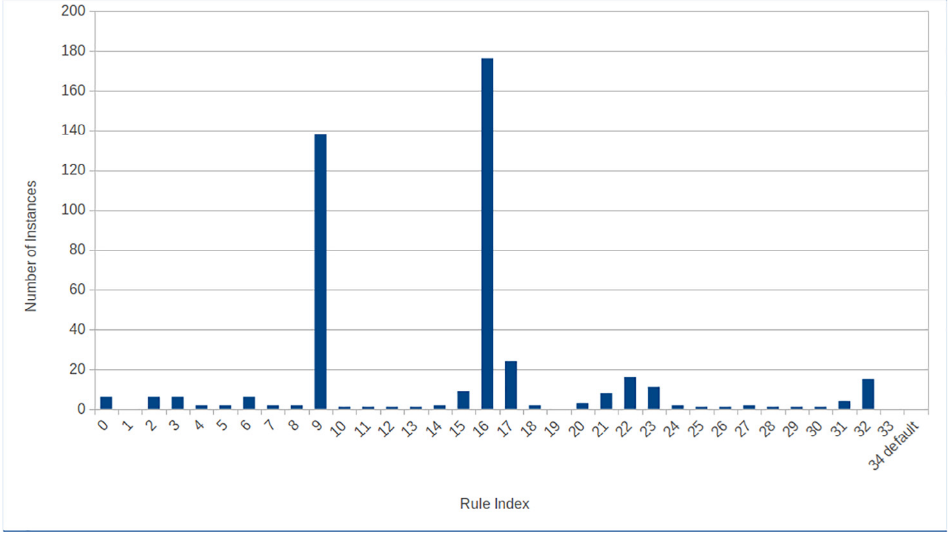

Fig. 3. Number of examples classified by each rule.

Table 13
Confusion matrix of the automatic age assessment system over the **original right laterality** dataset using **random forests with 10 trees.**

		Predicted phase										
		Ph1	**Ph2**	**Ph3**	**Ph4**	**Ph5**	**Ph6**	**Ph7**	**Ph8**	**Ph9**	**Ph10**	**Total**
Actual phase	**Ph1**	**10**	2	0	0	0	0	0	0	0	0	12
	Ph2	3	**7**	0	0	0	0	0	0	0	0	10
	Ph3	2	2	**9**	0	0	0	0	0	0	0	13
	Ph4	0	0	0	**9**	1	0	0	0	0	0	10
	Ph5	0	0	0	12	**10**	10	0	1	2	0	35
	Ph6	0	0	0	9	4	**14**	0	0	3	1	31
	Ph7	0	0	0	21	11	9	**4**	4	17	9	75
	Ph8	0	0	0	14	3	13	1	**10**	11	5	57
	Ph9	0	0	0	11	9	8	2	2	**31**	6	69
	Ph10	0	0	0	27	14	15	3	7	32	**43**	141
	Total	15	11	9	103	52	69	10	24	96	64	**453**

Table 14
Metric values for the automatic age assessment system over the **original right laterality** dataset (**NSLVOrd**).

	Ph1	**Ph2**	**Ph3**	**Ph4**	**Ph5**	
	Ph6	**Ph7**	**Ph8**	**Ph9**	**Ph10**	**Global**
CCR/Recall	0.833	0.7	0.615	0.5	0.743	
	0.129	0.027	0.053	0.565	0.199	0.291
F1-Score	0.741	0.667	0.762	0.227	0.242	
	0.216	0.052	0.09	0.368	0.303	0.265
OMAE	0.167	0.3	0.538	0.5	0.543	
	1.323	2.173	2.193	1.609	2.312	1.77

Table 15
Metric values for the automatic age assessment system over the **original right laterality** dataset (**random forests with 10 trees**).

	Ph1	**Ph2**	**Ph3**	**Ph4**	**Ph5**	
	Ph6	**Ph7**	**Ph8**	**Ph9**	**Ph10**	**Global**
CCR/Recall	0.833	0.7	0.692	0.9	0.286	
	0.452	0.053	0.175	0.449	0.305	0.325
F1-Score	0.741	0.667	0.818	0.159	0.23	
	0.28	0.094	0.247	0.376	0.42	0.265
OMAE	0.167	0.3	0.462	0.1	0.943	
	1.129	2.12	1.982	1.841	2.461	1.823

Making a deeper expert analysis of the performance of the classifier, we can observe that it correctly classifies 68% of the cases when focusing on individuals under 30 years of age (phases Ph01-18–19 to Ph05-27–30), with suitable F1-Score and OMAE values per class (ranging from 0.762 to 0.227 and from 0.167 to 0.5, respectively). This shows that our automatic classifier has a consistent behavior with respect to the existing knowledge in the discipline. Nevertheless, above this age the percentage drops to 19.5%. The loss of efficacy in older people has been widely discussed in the context of FA. As early as 1986, Katz and Suchey found that the error increases much after the age of 40 [42], a conclusion that other authors have also reached later [23]. There is also a tendency to underestimate the older subjects as clearly seen in the confusion matrix from phase Ph06-30–35 on. This effect is all the more pronounced the older the age of the subject, being more balanced in phase Ph07-35–39 and more extreme in Ph09-44–50 and Ph10-50+. Once again, this trend is shared by most of the validation studies of other authors [22,43].

Underestimation of age-at-death in older people could be attributed to factors as the greater morphological variability of the pubis in advanced ages or the ineffectiveness of the morphological characteristics used for this age group, as proposed by recent studies [32]. The appearance of specific degenerative features in the pubic symphysis, such as bone erosion, irregularities in the surface of the symphysis, sclerosis, and cyst formation, are conditioned by numerous biochemical and biomechanical factors and different pathologies [44]. The influence of these factors will be cumulative with age, so many of the traits analyzed by Todd may lose their usefulness in advanced ages due to increased variability. The main solution to this problem will be to analyze each trait in detail, as well as to propose new morphological analysis methods, until finding those that show a more stable relationship with age.

Brooks and Suchey proposed a new method considering the fusion of Todd's age-at-death phases Ph01-18-19, Ph02-20-21 and Ph03-22-24; Ph04-25-26 and Ph05-27-30; and Ph07-35-39 and Ph08-39-44, thus reducing the final number of phases to 6 [8]. Our explainable ML method provides us with objective information for a possible restructuring of Todd's phases (with an alternative grouping) and criteria, which will be treated in detail in future works. However, by way of example, we can observe that phases Ph01-18-19, Ph02-20-21 and Ph03-22-24 show especially acceptable percentages of correct classification, so merging these phases would only cause a reduction in the accuracy of the automatic age estimation system. On the contrary, phase Ph05-27-30 houses the majority of individuals in phases Ph04-25-26 and Ph06-30-35, so the performance could be improved by merging these phases.

Likewise, for older people, the classifier is able to distinguish two clearly differentiated groups: one in phase Ph09-44-50, which we could attribute to middle-aged adults, characterized mainly by pubic symphyses with irregular porosity of medium degree and the beginning of the formation of bony outgrowths in the ventral margin (rule R9); and another in phase Ph10-50+, which we could attribute to older adults and in which the abundant presence of irregular porosity and bony outgrowths on the ventral margin stands out (rules R22 and R32).

5.2. Comparison with results from the state of the art

We will also perform a comparison with state-of-the-art results, considering the age estimation methods reviewed in Section 2.3. Notice that, those methods are not phase-based but directly estimate the age-at-death value. To compare our accuracy, a transformation has been made from a phase-based age-at-death estimation to a numerical value based on the mean of the range in each Todd's phase, which will clearly degrade the performance of our methods. This approximation is only made for comparison purposes. Besides, the existing methods are not based on automating the operation of a classical skeleton-based age estimation method but consider advanced design methodologies involving the use of 3D models and which in some cases consider different pubic symphysis traits from those used by Todd. For example, the best performing method by Koterova et al.'s [32] uses both pubic symphysis and ilium bone traits. Even so, we think that the comparison, although very generic, can allow us to benchmark the accuracy of our proposal.Table 16 collects the comparative results. Two different rows are reported both for our method and for random forests with 10 trees (i.e., the classifiers described in the previous section), corresponding to the values obtained over the right laterality dataset (the one considered to learn the current models, after the application of the oversampling technique, i.e. a training error) and the left laterality dataset (i.e. a test error). It can be seen how even if our method does not directly compute an age-at-death value but an age-at-death range (i.e. one of the Todd's phases), it is in the state-of-the-art accuracy ranges, outperforming all of them but Koterova et al.'s in the RMSE test errors and providing a MAE value rather close to that method. Besides, our method outperforms the random forests classifier both in training and test errors, with the additional advantage of its simplicity and higher interpretability. Of course, the benchmarking with the state-of-the-art methods must be taken as a rough comparison as the sample considered in each case actually has an influence on the different results obtained. The size, age range, and composition of the samples considered significantly vary in each case. The number of samples ranges between 41 and 941, with some methods dealing with a very small number of instances; there are different age ranges, with our method having the shortest range; and some experiments use a single-ethnic population while others deal with a multi-ethnic one, some consider a single sex and others the two of them, and some make use of casts representing prototypical samples while others do not. It can be seen how our method is dealing with the smallest age range among the reported samples. This decision was taken to better reproduce the original scenario tackled by Todd in his seminal paper [6] (see Section 4.2). Even so, in order to develop the fairest possible comparison with the existing approaches, we have also validated the obtained phase-based classifier using the whole available sample. To do so, we have also incorporated the pubic symphysis from those subjects whose age-at-death was over 60 years (i.e. 34 left laterality and 34 right laterality instances in the 61–82 years range). Notice that, the

Table 16
Comparison between **NSLVOrd**, **random forests with 10 trees**, and state-of-the-art results.

Method	# samples	age range	RMSE	MAE
Slice and Algee-Hewitt [29]	41	19–96	17.15	–
Stoyanova et al. [30]	56 (44 real + 12 casts)	16–100	19	–
Stoyanova et al. [31]	93 (68 real + 25 casts)	16–90	13.7–16.5	–
Koterova et al. [32]	941	19–100	**12.1**	**9.7**
NSLVOrd (train: right laterality)	453	18–60	12.35	9.13
NSLVOrd (test: left laterality)	439	18–60	13.19	10.38
RF 10 (train: right laterality)	453	18–60	12.65	9.19
RF 10 (test: left laterality)	439	18–60	13.82	10.85

Table 17
Confusion matrix of the **NSLVOrd** automatic age assessment system over the **whole original right laterality dataset: 18–82 years age range.**

		Predicted phase										
		Ph1	**Ph2**	**Ph3**	**Ph4**	**Ph5**	**Ph6**	**Ph7**	**Ph8**	**Ph9**	**Ph10**	**Total**
	Ph1	**10**	2	0	0	0	0	0	0	0	0	12
	Ph2	3	**7**	0	0	0	0	0	0	0	0	10
	Ph3	2	2	**8**	1	0	0	0	0	0	0	13
	Ph4	0	0	0	**5**	5	0	0	0	0	0	10
Actual phase	**Ph5**	0	0	0	5	**26**	1	0	0	2	1	35
	Ph6	0	0	0	3	19	**4**	0	0	4	1	31
	Ph7	0	0	0	13	28	0	**2**	2	24	6	75
	Ph8	0	0	0	3	28	1	0	**3**	17	5	57
	Ph9	0	0	0	0	27	0	0	0	**39**	3	69
	Ph10	0	0	0	4	54	1	0	5	70	**41**	175
	Total	15	11	8	34	187	7	2	10	156	57	**487**

Table 18
Metric values for the **NSLVOrd** automatic age assessment system over the **whole original right laterality dataset: 18–82 years age range.**

	Ph1	Ph2	Ph3	Ph4	Ph5	
	Ph6	Ph7	Ph8	Ph9	Ph10	Global
CCR/Recall	0.833	0.7	0.615	0.5	0.743	
	0.129	0.027	0.053	0.565	0.234	0.298
F1-Score	0.741	0.667	0.762	0.227	0.234	
	0.211	0.052	0.09	0.347	0.353	0.282
OMAE	0.167	0.3	0.538	0.5	0.543	
	1.323	2.173	2.193	1.609	2.16	1.754

rule base considered is the same that has been derived from the 18–60 years sample (i.e. no new training process has been developed) and the new instances have only been considered to validate the already designed classifier.

Table 17 reports the confusion matrix of our classifier over the whole right laterality sample. Of course, the main difference with respect to that in Table 12 is in the last phase, Ph10-50+, now including 34 more samples which are mainly classified as Ph9-44-50 (13) and Ph10-50+ (13) although 7 are also classified as Ph5-27-30 and 1 as Ph6-30-35.

Table 18 shows very similar values in the different performance metrics with respect to the original ones in Table 14. The CCR value increases from 0.291 to 0.298, as a consequence of a better classification of the Ph10-50 + instances thanks to the increase of samples (of course, the other nine phases stay on the same classification percentage). As regards F1-Score, there is a slight decrease in the values for age-at-death phases Ph5-27-30, Ph6-30-35, and Ph9-44-50 (as a consequence of the said 21 new instances wrongly classified) that is compensated by the right classification of 13 new instances in phase Ph10-50+. The overall value of the measure slightly increases. Finally, the OMAE value slightly increases as well as a result of the error decrease in the last phase. Finally, we also make the approximate computation of the numeric age-at-death values to benchmark again our automatic estimation methods with those from the literature. Table 19 reports the obtained values for NSLVOrd and random forests in the 18–82 age range. To get them, the representative numerical value associated to the last Todd's phase is now considered as 66 (the average in the new age range 50–82) instead of the previous 55 (average in the previous 50–60 age range). This is very harmful for the numerical errors of our methods as now the last Todd's phase covers a very large age-at-death range, 50–82 years, and the mean value is very far from the interval extents. Of course, that is a consequence of the original Todd's method definition, that was less influential in its performance by the time the technique was defined. Anyway, we should recall again this computation is just an approximation for a rough comparison and must be treated in such way. As expected, the obtained test errors are higher than the previous ones: 1.43 and 1.24 years higher in RMSE

Table 19
Comparison between **NSLVOrd**, **random forests**, and state-of-the-art results **in the 18–82 years age range**.

Method	# samples	age range	RMSE	MAE
Slice and Algee-Hewitt [29]	41	19–96	17.15	–
Stoyanova et al. [30]	56 (44 real + 12 casts)	16–100	19	–
Stoyanova et al. [31]	93 (68 real + 25 casts)	16–90	13.7–16.5	–
Koterova et al. [32]	941	19–100	**12.1**	**9.7**
NSLVOrd (train: right laterality)	487	18–82	13.89	10.51
NSLVOrd (test: left laterality)	473	18–82	14.61	11.62
RF 10 (train: right laterality)	487	18–82	14.66	11.15
RF 10 (test: left laterality)	473	18–82	15.98	12.9

Table 20
Relation between pubic bone traits and age-at-death phase identification in the derived rule base (**NSLVOrd**) AF = Articular Face; IP = Irregular Porosity; USE = Upper Symphysial Extremity; BN = Bony Nodule; LSE = Lower Symphysial Extremity; DM = Dorsal Margin; DP = Dorsal Plateau; VB = Ventral Bevel; VM = Ventral Margin.

	AF	IP	USE	BN	LSE	DM	DP	VB	VM	NumRules
Ph1	2		3	2	2		1		2	4
Ph2	10		1	2	3		3			5
Ph3	3			1	1				1	2
Ph4	6		2	1	1		1	4	5	4
Ph5		2			1			2	4	3
Ph6	2	2					1	4	3	3
Ph7	1	1						2	3	2
Ph8	1	2						4	4	3
Ph9	4	1					2	4	3	4
Ph10		3					1	1	4	4

Table 21
Relation between pubic bone traits and age-at-death phase identification in the derived rule base (**random forests**) AF = Articular Face; IP = Irregular Porosity; USE = Upper Symphysial Extremity; BN = Bony Nodule; LSE = Lower Symphysial Extremity; DM = Dorsal Margin; DP = Dorsal Plateau; VB = Ventral Bevel; VM = Ventral Margin.

	AF	IP	USE	BN	LSE	DM	DP	VB	VM	NumRules
Ph1	425	**261**	69	66	80		187	**336**	394	444
Ph2	51		31	38	26		39	**22**	26	53
Ph3	53	**5**	18	34	43		17	**31**	43	62
Ph4	55	36	5	11	31		25	52	52	58
Ph5	56	45			17		32	57	58	59
Ph6	40	49			5		26	47	48	51
Ph7	56	52	2		4		23	52	63	63
Ph8	55	51			9		31	55	57	57
Ph9	71	65			4		48	7	69	73
Ph10	88	106	1				5	10	107	114

and MAE for NSLVOrd, and 2.16 and 2.05 years higher in RMSE and MAE for random forests. Nevertheless, the Todd-based classifier obtained by NSLVOrd stays competitive with respect to the state of the art when the age range is increased 22 years. It is only clearly surpassed by Koterova et al.'s method even our approach is automating an existing a phase-based age estimation method and it considers a much larger number of instances in the considered samples than the remaining methods. Again, it clearly outperforms the random forests solution.

5.3. Knowledge extraction from the Todd's automatic age assessment method designed

One of the most interesting contributions of our study is that it allows us to know which pubic symphysis traits are more/less useful to estimate age-at-death, as well as to uncover the hidden relation between characteristics and phases.

Table 20 provides a global vision of the use of each variable in the derived rule base. This table allows us to segment different groups of variables. The **Dorsal Margin** is never used, thus showing up as an irrelevant trait in Todd's method, an extremely remarkable finding. Besides, there are some variables which are considered to estimate age-at-death along almost the whole individual's age-at-death range: **Articular Face**, **Dorsal Plateau**, and **Ventral Margin**. Finally, there is another group of variables which are mainly focused on the identification of either young individuals: **Upper Symphysial Extremity**, **Bony Nodule** and **Lower Symphysial Extremity**; or older individuals: **Irregular Porosity** and **Ventral Bevel**.

Table 21 collects the statistics on the use of the variables for the random forests classifier. The behavior is very similar even using 30 times more rules. In particular, the **Dorsal Margin** is never used again, confirming the previous conclusion. The only two differences, highlighted in bold font, are on the use of **Irregular Porosity** and **Ventral Bevel** in the first three phases (especially in the first one). However, we should remind the performance of both classifiers was the same in those phases, hence that use can be considered as redundant as the same knowledge is being captured by other traits.

As a practical example, we can focus on the importance of the **Bony Nodule**, a discarded characteristic in the adaptation of Todd's method developed by Brooks and Suchey [8]. According to them, it is an inconstant trait that may or may not appear. However, our results show that, in those cases in which it is observed, it is actually a useful morphological characteristic to identify the early stages of development.

The uncovered information can be so useful for the design of new age estimation methods from the pubic symphysis, regardless their manual or automatic nature. The fact that different subsets of pubic symphysis traits are used to estimate age-at-death for different age groups seems to recommend the design of hierarchical methods considering partial decisions to segment those age groups instead of examining all the features at the same level as in Todd's method and its variants.

6. Conclusions and future works

Age-at-death estimation is a fundamental task in FA. The pubic symphysis is extensively used due to its reliability. Todd defined 10 age-at-death phases characterized by different pubic bone traits such as appearance, ossification patterns, and morphology, which has been later used by a family of methods. In this contribution, we propose the automation of Todd's age assessment method. This involves an especially complex explainable ML task: an imbalanced ordinal classification problem with a small sample size and a high dimension. We apply an ordinal classification method (NSLVOrd) over a good collection of pubic symphysis. A thorough experimental study is developed considering the whole pubis dataset and both subsets with a single laterality, the use or not of oversampling, and different training-test partitions. A final experiment considered 3 datasets, 3 ML methods (NSLVOrd, random forests, and a deep neural network), and a 5×2-cv test.

A compact set of 34 rules with state-of-the-art performance is derived from an iterative XAI process loop with the forensic anthropologists. Thanks to the explainable nature of the intelligent system, a human-based reliability study is faced allowing them to identify which pubic symphysis traits are more useful to estimate age-at-death, as well as to uncover the hidden relation between characteristics and phases. As a consequence, we consider that the proposed explainable ML design can be a good alternative for the automation of the decision process in many other phase-based methods frequently used in forensic anthropology.

Despite the successful automation of the decision-making process of the method, it can still be ineffective for accurately estimating age-at-death in forensic contexts, mainly due to the limitation derived from the narrow definition of the ranges for each phase. We thus plan to exploit the extracted knowledge to design new age estimation methods from the pubic symphysis, considering new age ranges and the longer life expectancy of the current population. We also aim to design new component-based age estimation methods based on Todd's pubic symphysis traits from the problem modeling presented in Section 3.2. To do so, we will make use of genetic programming methods [45] to develop symbolic regression over those traits, also developing feature selection in an automatic way. We also aim to study the impact of intra and interobserver error in our future age estimation method proposals. Finally, we will explore the automatic extraction of the morphological traits to facilitate the fast and objective characterization of this crucial bone for age-at-death estimation.

CRediT authorship contribution statement

Juan Carlos Gámez-Granados: Data curation, Formal analysis, Software, Investigation, Methodology, Methodology, Writing - original draft. **Javier Irurita:** Data curation, Resources, Methodology, Formal analysis, Investigation, Writing - review & editing. **Raúl Pérez:** Methodology, Formal analysis, Investigation, Writing - original draft. **Antonio González:** Funding acquisition, Methodology, Formal analysis. **Sergio Damas:** Funding acquisition, Methodology, Formal analysis, Writing - review & editing. **Inmaculada Alemán:** Data curation, Resources, Formal analysis. **Oscar Cordón:** Funding acquisition, Methodology, Formal analysis, Supervision, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Spanish Ministry of Science and Innovation (MCIN), Agencia Estatal de Investigación (AEI), under grant CONFIA (PID2021-122916NB-I00), by the Spanish Ministry of Science, Innovation and Universities under grant EXASOCO (PGC2018-101216-B-I00), and by the Regional Government of Andalusia and the University of Granada under grants EXAISFI (P18-FR-4262) and FORAGE (B-TIC-456-UGR20), including European Regional Development Funds (ERDF). Funding for open access charge: Universidad de Granada/CBUA.

Appendix A. Rule base of the Todd’s age assessment system

.

```
                 Phase01
R0: IF "BonyNodule" IS ’Absent’ AND "LowerSymphysialExtremity" IS ’NotDefined’ AND "
    VentralMargin" IS ’Absent’
 THEN Phase IS Ph01-18-19 with Weight = 0,506
R5: IF "UpperSymphysialExtremity" IS ’Defined’ AND "BonyNodule" IS ’Absent’ AND "
    LowerSymphysialExtremity" IS ’NotDefined’ AND "VentralMargin" IS ’Absent’
 THEN Phase IS Ph01-18-19 with Weight = 0,632
R6: IF "ArticularFace" IS ’RidgesAndGrooves’ AND "UpperSymphysialExtremity" IS ’NotDefined
    ’
 THEN Phase IS Ph01-18-19 with Weight = 0,580
R25: IF "ArticularFace" IS ’RidgesAndGrooves’ AND "UpperSymphysialExtremity" IS ’
    NotDefined’ AND "DorsalPlateau" IS ’Present’
 THEN Phase IS Ph01-18-19 with Weight = 1,000

                 Phase02
R1: IF "ArticularFace" IS (’RidgesAndGrooves’ OR ’GroovesShallow’) AND "
    UpperSymphysialExtremity" IS ’Defined’ AND "LowerSymphysialExtremity" IS ’NotDefined’
 THEN Phase IS Ph02-20-21 with Weight = 0,629
R3: IF "ArticularFace" IS ’GroovesShallow’ AND "LowerSymphysialExtremity" IS ’NotDefined’
 THEN Phase IS Ph02-20-21 with Weight = 0,731
R4: IF "ArticularFace" IS ’RidgesFormation’ AND "DorsalPlateau" IS ’Absent’
 THEN Phase IS Ph02-20-21 with Weight = 0,700
R7: IF "ArticularFace" IS (’RegularPorosity’ OR ’RidgesFormation’ OR ’RidgesAndGrooves’ OR
     ’GroovesShallow’) AND "BonyNodule" IS ’Present’ AND "DorsalPlateau" IS ’Absent’
 THEN Phase IS Ph02-20-21 with Weight = 0,718
R11: IF "ArticularFace" IS (’RidgesAndGrooves’ OR ’GroovesShallow’) AND "BonyNodule" IS ’
    Absent’ AND "LowerSymphysialExtremity" IS ’NotDefined’ AND "DorsalPlateau" IS ’
    Present’
 THEN Phase IS Ph02-20-21 with Weight = 0,707

                 Phase03
```

```
R2: IF "BonyNodule" IS ’Present’
 THEN Phase IS Ph03-22-24 with Weight = 0,674
R8: IF "ArticularFace" IS (’RidgesFormation’ OR ’RidgesAndGrooves’ OR ’GroovesRemains’)
    AND "LowerSymphysialExtremity" IS ’Defined’ AND "VentralMargin" IS ’Absent’
 THEN Phase IS Ph03-22-24 with Weight = 1,000

                  Phase04
R10 : IF "LowerSymphysialExtremity" IS ’NotDefined’ AND "VentralBevel" IS ’Absent’ AND "
    VentralMargin" IS ’PartiallyFormed’
 THEN Phase IS Ph04-25-26 with Weight = 1,000
R12: IF "ArticularFace" IS (’GroovesShallow’ OR ’NoGrooves’) AND "UpperSymphysialExtremity
    " IS ’Defined’ AND "BonyNodule" IS ’Absent’ AND "DorsalPlateau" IS ’Present’
 THEN Phase IS Ph04-25-26 with Weight = 0,412
R17: IF "ArticularFace" IS (’GroovesShallow’ OR ’GroovesRemains’) AND "VentralBevel" IS (’
    Absent’ OR ’InProcess’) AND "VentralMargin" IS (’Absent’ OR ’PartiallyFormed’)
 THEN Phase IS Ph04-25-26 with Weight = 0,149
R21: IF "ArticularFace" IS (’GroovesShallow’ OR ’NoGrooves’) AND "UpperSymphysialExtremity
    " IS ’Defined’ AND "VentralBevel" IS ’Absent’ AND "VentralMargin" IS (’Absent’ OR ’
    PartiallyFormed’)
 THEN Phase IS Ph04-25-26 with Weight = 0,118

                  Phase05
R16: IF "IrregularPorosity" IS ’Absence’ AND "LowerSymphysialExtremity" IS ’Defined’ AND "
    VentralMargin" IS (’PartiallyFormed’ OR ’FormedWithoutBonyOutgrowths’)
 THEN Phase IS Ph05-27-30 with Weight = 0,103
R31: IF "IrregularPorosity" IS ’Medium’ AND "VentralBevel" IS (’Absent’ OR ’InProcess’)
    AND "VentralMargin" IS (’Absent’ OR ’PartiallyFormed’)
 THEN Phase IS Ph05-27-30 with Weight = 0,364
R34 : IF
 THEN Phase IS Ph05-27-30 with Weight = 0,038

                  Phase06
R13: IF "ArticularFace" IS ’NoGrooves’ AND "DorsalPlateau" IS ’Present’ AND "VentralBevel"
     IS (’InProcess’ OR ’Present’)
 THEN Phase IS Ph06-31-34 with Weight = 0,556
R18: IF "ArticularFace" IS ’GroovesRemains’ AND "IrregularPorosity" IS ’Absence’ AND "
    VentralBevel" IS ’Present’ AND "VentralMargin" IS (’Absent’ OR ’PartiallyFormed’)
 THEN Phase IS Ph06-31-34 with Weight = 0,556
R20: IF "IrregularPorosity" IS ’Medium’ AND "VentralBevel" IS ’Absent’ AND "VentralMargin"
     IS ’PartiallyFormed’
 THEN Phase IS Ph06-31-34 with Weight = 0,692

                  Phase07
R26: IF "IrregularPorosity" IS ’Much’ AND "VentralMargin" IS (’Absent’ OR ’PartiallyFormed
    ’)
 THEN Phase IS Ph07-35-39 with Weight = 1,000
R29 : IF "ArticularFace" IS ’NoGrooves’ AND "VentralBevel" IS (’InProcess’ OR ’Present’)
     AND "VentralMargin" IS ’Absent’
 THEN Phase IS Ph07-35-39 with Weight = 1,000

                  Phase08
R15: IF "ArticularFace" IS ’GroovesRemains’ AND "VentralBevel" IS (’InProcess’ OR ’Present
    ’) AND "VentralMargin" IS ’FormedWithFewBonyOutgrowths’
 THEN Phase IS Ph08-40-44 with Weight = 0,263
R19: IF "IrregularPorosity" IS ’Medium’ AND "VentralBevel" IS ’Absent’ AND "VentralMargin"
     IS (’PartiallyFormed’ OR ’FormedWithFewBonyOutgrowths’)
 THEN Phase IS Ph08-40-44 with Weight = 0,235
```

(*continued on next page*)

```
R28: IF "IrregularPorosity" IS 'Medium' AND "VentralBevel" IS 'Present' AND "VentralMargin
     " IS 'PartiallyFormed'
 THEN Phase IS Ph08-40-44 with Weight = 1,000

                  Phase09
R9: IF "IrregularPorosity" IS 'Medium' AND "VentralMargin" IS ('
     FormedWithoutBonyOutgrowths' OR 'FormedWithFewBonyOutgrowths')
 THEN Phase IS Ph09-45-49 with Weight = 0,200
R14 : IF "ArticularFace" IS 'NoGrooves' AND "DorsalPlateau" IS 'Present' AND "
     VentralBevel" IS ('Absent' OR 'InProcess')
 THEN Phase IS Ph09-45-49 with Weight = 0,667
R24 : IF "ArticularFace" IS ('GroovesShallow' OR 'GroovesRemains') AND "DorsalPlateau"
     IS 'Present' AND "VentralBevel" IS 'Present'
 THEN Phase IS Ph09-45-49 with Weight = 0,667
R30: IF "ArticularFace" IS 'GroovesRemains' AND "VentralBevel" IS 'Absent' AND "
     VentralMargin" IS 'FormedWithFewBonyOutgrowths'
 THEN Phase IS Ph09-45-49 with Weight = 1,000

                  Phase10
R22: IF "IrregularPorosity" IS 'Much'
 THEN Phase IS Ph10-50+ with Weight = 0,228
R23: IF "VentralBevel" IS 'Absent' AND "VentralMargin" IS 'FormedWithFewBonyOutgrowths'
 THEN Phase IS Ph10-50+ with Weight = 0,714
R27: IF "IrregularPorosity" IS ('Absence' OR 'Medium') AND "VentralMargin" IS '
     FormedWithRecessesAndProtrusions'
 THEN Phase IS Ph10-50+ with Weight = 1,000
R32: IF "DorsalPlateau" IS 'Absent' AND "VentralMargin" IS ('FormedWithFewBonyOutgrowths'
     OR 'FormedWithRecessesAndProtrusions')
 THEN Phase IS Ph10-50+ with Weight = 0,198
```

References

[1] D.H. Ubelaker, Forensic Anthropology: Methodology and Diversity of Applications, John Wiley & Sons, Ltd, 2008, Ch. 2, pp. 41–69. doi:10.1002/9780470245842.ch2.

[2] D.H. Ubelaker, Q.R. Cordero, N. Linton, Recent research in forensic anthropology, Eur. J. Anatomy 24 (3) (2020) 221–227.

[3] D. Ubelaker, H. Khosrowshahi, Estimation of age in forensic anthropology: historical perspective and recent methodological advances, Forens. Sci. Res. 4 (2019) 1–9.

[4] E. Cunha, E. Baccino, L. Martrille, F. Ramsthaler, J. Prieto, Y. Schuliar, N. Lynnerup, C. Cattaneo, The problem of aging human remains and living individuals: A review, Forensic Sci. Int. 193 (1) (2009) 1–13, https://doi.org/10.1016/j.forsciint.2009.09.008.

[5] B. Dudzik, N. Langley, Estimating age from the pubic symphysis: A new component-based system, Forensic Sci. Int. 257 (2015) 98–105, https://doi.org/10.1016/j.forsciint.2015.07.047.

[6] T.W. Todd, Age changes in the pubic bone, Am. J. Phys. Anthropol. 3 (3) (1920) 285–328.

[7] B. Gilbert, T.W. McKern, A method for aging the female os pubis, Am. J. Phys. Anthropol. 38 (1) (1973) 31–38.

[8] S. Brooks, J.M. Suchey, Skeletal age determination based on the os pubis: A comparison of the acsádi-nemeskéri and suchey-brooks methods, Human Evolution 5 (3) (1990) 227–238, https://doi.org/10.1007/bf02437238.

[9] K. Hartnett, Analysis of age-at-death estimation using data from a new, modern autopsy sample-part I: Pubic bone, J. Forensic Sci. 55 (5) (2010) 1145–1151, https://doi.org/10.1111/j.1556-4029.2010.01399.x.

[10] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inform. Fusion 58 (2020) 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

[11] J.S. Cardoso, J.F.P. da Costa, Learning to classify ordinal data: The data replication method, J. Mach. Learn. Res. 8 (50) (2007) 1393–1429, URL: http://jmlr.org/papers/v8/cardoso07a.html.

[12] J.C. Gámez, D. García, A. González, R. Pérez, Ordinal Classification based on the Sequential Covering Strategy, Int. J. Approximate Reasoning 76 (2016) 96–110.

[13] R. Caruana, Learning from imbalanced data: Rank metrics and extra tasks, in: American Association for Artificial Intelligence (AAAI) Conference (AAAI Technical Report WS-00-05), 2000, pp. 51–57.

[14] H. He, Y. Ma, Imbalanced Learning: Foundations, Algorithms, and Applications, first ed., Wiley-IEEE Press, 2013.

[15] F. Provost, T. Fawcett, Robust classification for imprecise environments, Mach. Learn. 42 (2001) 203–231, https://doi.org/10.1023/A:1007601015854.

[16] A. Fernández, S. García, F. Herrera, N.V. Chawla, SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary, J. Artif. Int. Res. 61 (1) (2018) 863–905.

[17] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.

[18] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[19] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, URL: http://www.deeplearningbook.org.

[20] J. Buckberry, A. Chamberlain, Age estimation from the auricular surface of the ilium: A revised method, Am. J. Phys. Anthropol. 119 (2002) 231–239, https://doi.org/10.1002/ajpa.10130.

[21] M. Iscan, S. Loth, R. Wright, Metamorphosis at the sternal rib end: a new method to estimate age at death in white males, Am. J. Phys. Anthropol. 65 (2) (1984) 147–156, https://doi.org/10.1002/ajpa.1330650206.

[22] A. Schmitt, P. Murail, E. Cunha, D. Rougé, Variability of the pattern of aging on the human skeleton: Evidence from bone indicators and implications on age at death estimation, J. Forensic Sci. 47 (2002) 1203–1209, https://doi.org/10.1520/JFS15551J.

[23] G. Berg, Pubic bone age estimation in adult women, J. Forensic Sci. 53 (2008) 569–577, https://doi.org/10.1111/j.1556-4029.2008.00712.x.
[24] A. Valsecchi, J. Irurita, P. Mesejo, Age estimation in forensic anthropology: methodological considerations about the validation studies of prediction models, Int. J. Legal Med. 133 (2019) 1915–1924, https://doi.org/10.1007/s00414-019-02064-7.
[25] E. Kimmerle, D. Prince, G. Berg, Inter-observer variation in methodologies involving the pubic symphysis, sternal ribs, and teeth, J. Forensic Sci. 53 (2008) 594–600, https://doi.org/10.1111/j.1556-4029.2008.00715.x.
[26] N. Shirley, P. Ramirez Montes, Age estimation in forensic anthropology: quantification of observer error in phase versus component-based methods, J. Forensic Sci. 60 (1) (2015) 107–111.
[27] C. Fojas, J. Kim, J. Minsky-Rowland, B. Algee-Hewitt, Testing inter-observer reliability of the transition analysis aging method on the william m. bass forensic skeletal collection, Am. J. Phys. Anthropol. 165 (2018) 183–193, https://doi.org/10.1002/ajpa.23342.
[28] F. Dedouit, T. No, C. Guilbeau-Frugier, D. Gainza, P. Otal, F. Joffre, D. Rougé, Virtual autopsy and forensic identification-practical application: A report of one case, J. Forensic Sci. 52 (2007) 960–964, https://doi.org/10.1111/j.1556-4029.2007.00475.x.
[29] D. Slice, B. Algee-Hewitt, Modeling bone surface morphology: A fully quantitative method for age-at-death estimation using the pubic symphysis, J. Forensic Sci. 60 (4) (2015) 835–843, https://doi.org/10.1111/1556-4029.12778.
[30] D. Stoyanova, B. Algee-Hewitt, D. Slice, An enhanced computational method for age-at-death estimation based on the pubic symphysis using 3D laser scans and thin plate splines, Am. J. Phys. Anthropol. 158 (3) (2015) 431–440, https://doi.org/10.1002/ajpa.22797.
[31] D.K. Stoyanova, B.F.B. Algee-Hewitt, J. Kim, D.E. Slice, A computational framework for age-at-death estimation from the skeleton: Surface and outline analysis of 3D laser scans of the adult pubic symphysis, J. Forensic Sci. 62 (6) (2017) 1434–1444, https://doi.org/10.1111/1556-4029.13439.
[32] A. Kotěrová, D. Navega, M. Štepanovský, Z. Buk, J. Bružek, E. Cunha, Age estimation of adult human remains from hip bones using advanced methods, Forensic Sci. Int. 287 (2018) 163–175, https://doi.org/10.1016/j.forsciint.2018.03.047.
[33] P. Villar, I. Alemán, L. Castillo, S. Damas, O. Cordón, A first approach to a fuzzy classification system for age estimation based on the pubic bone, in: 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2017, pp. 1–6, https://doi.org/10.1109/FUZZ-IEEE.2017.8015760.
[34] M. Stepanovsky, A. Ibrova, Z. Buk, J. Velemínská, Novel age estimation model based on development of permanent teeth compared with classical approach and other modern data mining methods, Forensic Sci. Int. 279. doi:10.1016/j.forsciint.2017.08.005.
[35] S. Aja-Fernández, R. de Luis-Garcia, M.A. Martin-Fernández, C. Alberola-López, A computational *TW*3 classifier for skeletal maturity assessment. A computing with words approach, J. Biomed. Inform. 37 (2) (2004) 99–107.
[36] J. Tanner, M. Healy, N. Cameron, H. Goldstein, Assessment of Skeletal Maturity and Prediction of Adult Height (TW3 Method), Assessment of Skeletal Maturity and Prediction of Adult Height, W.B. Saunders, 2001.
[37] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357, https://doi.org/10.1613/jair.953.
[38] R. Overbury, L. Cabo, D. Dirkmaat, S. Symes, Asymmetry of the os pubis: Implications for the suchey-brooks method, Am. J. Phys. Anthropol. 139 (2009) 261–268, https://doi.org/10.1002/ajpa.20999.
[39] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput. 10 (7) (1998) 1895–1923, https://doi.org/10.1162/089976698300017197.
[40] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The weka data mining software: an update, ACM SIGKDD Explorations Newsletter 11 (1) (2009) 10–18.
[41] J.T. Hancock, T.M. Khoshgoftaar, Survey on categorical data for neural networks, J. Big Data 7 (1) (2020) 1–41.
[42] D. Katz, J.M. Suchey, Race differences in pubic symphyseal aging patterns in the male, Am. J. Phys. Anthropol. 80 (2) (1989) 167–172, https://doi.org/10.1002/ajpa.1330800204.
[43] M. Djuric, D. Djonic, D. Popovic, J. Marinkovic, Evaluation of the suchey–brooks method for aging skeletons in the balkans, J. Forensic Sci. 52 (2007) 21–23, https://doi.org/10.1111/j.1556-4029.2006.00333.x.
[44] S. Mays, The effect of factors other than age upon skeletal age indicators in the adult, Ann. Hum. Biol. 42 (2015) 1–10, https://doi.org/10.3109/03014460.2015.1044470.
[45] R. Poli, W.B. Langdon, N.F. McPhee, A field guide to genetic programming, Published via http://lulu.com and freely available at http://www.gp-field-guide.org.uk, 2008, (with contributions by J.R. Koza).