Expert Systems

WILEY

Expert Systems

ORIGINAL ARTICLE

Interpretable Machine Learning for Age-at-Death Estimation From the Pubic Symphysis

Enrique Bermejo[1,2,3] | Antonio David Villegas[2] | Javier Irurita[4] | Sergio Damas[3,5] | Oscar Cordón[1,3]

[1]Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain | [2]Panacea Cooperative Research S. Coop., Ponferrada, Spain | [3]Andalusian Research Institute in Data Science and Computational Intelligence (DASCI), Granada, Spain | [4]Department of Legal Medicine, Toxicology and Physical Anthropology, University of Granada, Granada, Spain | [5]Department of Software Engineering, University of Granada, Granada, Spain

Correspondence: Enrique Bermejo (enrique.bermejo@decsai.ugr.es)

Received: 19 June 2024 | **Revised:** 19 December 2024 | **Accepted:** 4 February 2025

Funding: This work was supported by Agencia Estatal de Investigación, Junta de Andalucía, Ministerio de Ciencia, Innovación y Universidades, European Regional Development Fund.

Keywords: decision support system | genetic programming | interpretable machine learning | symbolic regression age estimation

ABSTRACT

Age-at-death estimation is an arduous task in human identification based on characteristics such as appearance, morphology or ossification patterns in skeletal remains. This process is performed manually, although in recent years there have been several studies that attempt to automate it. One of the most recent approaches involves considering interpretable machine learning methods, obtaining simple and easily understandable models. The ultimate goal is not to fully automate the task but to obtain an accurate model supporting the forensic anthropologists in the age-at-death estimation process. We propose a semi-automatic method for age-at-death estimation based on nine pubic symphysis traits identified from Todd's pioneering method. Genetic programming is used to learn simple mathematical expressions following a symbolic regression process, also developing feature selection. Our method follows a component-scoring approach where the values of the different traits are evaluated by the expert and aggregated by the corresponding mathematical expression to directly estimate the numeric age-at-death value. Oversampling methods are considered to deal with the strongly imbalanced nature of the problem. State-of-the-art performance is achieved thanks to an interpretable model structure that allows us to both validate existing knowledge and extract some new insights in the discipline.

1 | Introduction

Personal identity is associated with the preservation and defence of Human Rights and is a tool to repair their violation. This applies to both individual and multi-victim tragedies, such as accidents, fires, terrorist attacks, natural disasters, etc. All these cases are closely related to forensic anthropology (FA) as a discipline with a leading role in human identification (Ubelaker 2008). The Latin American Association of Forensic Anthropology (ALAF) and the International Committee of the Red Cross (CICR) define FA as the "application of the theories, methods and techniques of social anthropology, archaeology and biological anthropology" in the search and recovery of human remains and human identification processes, as well as to clarify the facts in support of the justice administration system and humanitarian work.

Skeleton-based forensic identification begins by estimating the biological profile (BP) (age, sex, ancestry and stature), which is a critical process to narrow down the range of possible matches during the identification process. Once the compatibilities have been established from the BP, the human identification is corroborated using DNA analysis, when the necessary resources exist and if the state of conservation of the bones allows its

application (Christensen et al. 2019). Therefore, accurate BP methods enable a significant acceleration of the identification process and its costs by reducing the number of required comparisons, which takes on particular importance considering disaster victim identification (DVI) scenarios. DVI is the method applied to identify victims of mass casualty incidents, such as aircraft crashes, natural disasters or terrorist attacks.

Age estimation, both age-at-death and in the living, is one of the most important tasks in FA. Age-at-death estimation is based on analysing certain characteristics such as appearance, morphology and ossification patterns in the skeletal remains of individuals. Among the different bony structures used, the pubic symphysis stands out due to its high reliability (Cunha et al. 2009; Dudzik and Langley 2015). Many different methods have been proposed since the pioneering proposal by Todd in 1920 (Todd 1920). We can distinguish between phase-based methods, that estimate an age-at-death range (i.e. a phase), and numerical methods, which directly provide an estimation for the specific age-at-death value. The method introduced byBrooks and Suchey (1990) extending Todd's proposal is currently the standard for phase-based methods (Schanandore et al. 2022). Such typology of methods had a great role in the mid-20th century, given the difficulty of performing complex analysis on large data sets. Dividing a problem into discrete phases was alluring due to the simplicity and ease of use through designing atlases and graphical illustrations. However, their main limitation is their great subjectivity, creating a dependence on the experience of the observer resulting in a reduced accuracy of the methods.

Meanwhile, we can also differentiate between methods obtaining the estimation from an overall analysis of the morphological characteristics associated with the different pubic symphysis changes (usually performed with a visual inspection of the bone) and methods analysing each pubic bone trait in isolation and then aggregating the partial observations to take the final decision (component-scoring or component-based methods). Gilbert and McKern's was the first proposal of such methods (Gilbert and McKern 1973). The use of component-based methods had already shown a significant reduction in both intra- and inter-observer error. This is because labeling each component separately can be done more objectively than assigning a general phase to the entire pubic symphysis.

The development of precise, fast, robust and automatic methods for age estimation is currently an area with a strong interest in FA (Ubelaker et al. 2020). In recent years, there have been several studies that attempt to automate age estimation from bone remains. Some of them rely on advanced computer vision and machine learning (ML) approaches to avoid subjective bias in the procedure, currently being the state-of-the-art methods in the area (Kotěrová et al. 2022). Despite achieving acceptable estimations, the results of these methods are too complex to support forensic anthropologists in improving and refining the technique. We refer the interested reader to Appendix A for a detailed analysis on the methods currently used for skeleton-based age estimation from the pubic symphysis.

A previous study (Gámez et al. 2022) presented a tentative design for the automation of Todd's method (Todd 1920) by considering an ordinal classification problem tackled by means of a rule learning algorithm. The forensic experts in the team managed to elicit the uncertain knowledge in Todd's proposal in the form of a set of nine variables (pubic bone traits) with different categorical values (representing their development/degenerative process state) associated (see Figure 1a and Table 1). The approach consisted of an interpretable rule-based method yielding a collection of 34 rules that leverage the values of nine linguistic variables to ascertain one of the 10 phases following Todd's methodology.

In the current study, the pubic bone traits will be used to learn an age-at-death estimator following Gilbert and McKern's component-scoring approach (see Figure 1b and Section A.2) to compute the numerical age-at-death value from the combination of the individual features. These variables will be considered as the input of a symbolic regression procedure, developed by genetic programming (GP) (Koza 1992) and applied on an oversampled data set. As other transparent ML models (Barredo Arrieta et al. 2020) like decision trees (DT) and rule sets (both of them considered in Gámez et al. (2022)), the symbolic nature of GP has an inherent interpretability but also a higher flexibility allowing it to learn more accurate models by searching in a richer model space (Mei et al. 2022). In this way, the interpretable model structure will be chosen according to the kind of relations we intend to find in the specific domain, a mathematical expression mapping morphological characteristics categorised in the nine variables to estimate a numeric age-at-death value, making use of the prior knowledge available within a theory-guided data science approach (Karpatne et al. 2017). In addition, we consider different approaches to improve the intrinsic GP interpretability (Mei et al. 2022) as the use of: (i) simple mathematical operators (lower [non-size] GP model complexity); (ii) reduced tree sizes by establishing a priori size limits (smaller GP model size) and (iii) advanced methods to reduce the model size as hybrid coefficient learning by genetic algorithm-programming (GA-P) (Howard and D'Dangelo 1995) and explicit tree simplification by GP-desired approximation (GP-DA) (Nguyen and Chu 2020). A sound experimental setup will be designed to obtain the most accurate and transparent model, which will be benchmarked against the state-of-the-art approaches in the field.

In summary, our contribution stands out for incorporating the key aspects of two distinct trends: (i) using the expert knowledge compiled in the analysis of morphological characteristics defined by traditional phase-based approaches, and (ii) estimating the age-at-death as a numerical value following a component-based approach. Such a novel design will be exploited by the symbolic regression model to learn the structure of the expressions and enhance the flexibility of the model. In addition, the empirical data-supported models obtained will allow forensic experts to validate them as well as reason and uncover new knowledge to be complemented with the existing one, composing a trustworthy human-centric decision support system designed in cooperation with the domain experts.

The remaining of this manuscript is structured as follows. Section 2 introduces some preliminaries of GP and describes the different GP methods considered in the current contribution. Section 3 provides a detailed description of the experimental setup, the series of experiments developed incrementally to succeed generating an accurate and interpretable model, and the analysis of results. Finally, Section 4 reports some concluding remarks and future works.

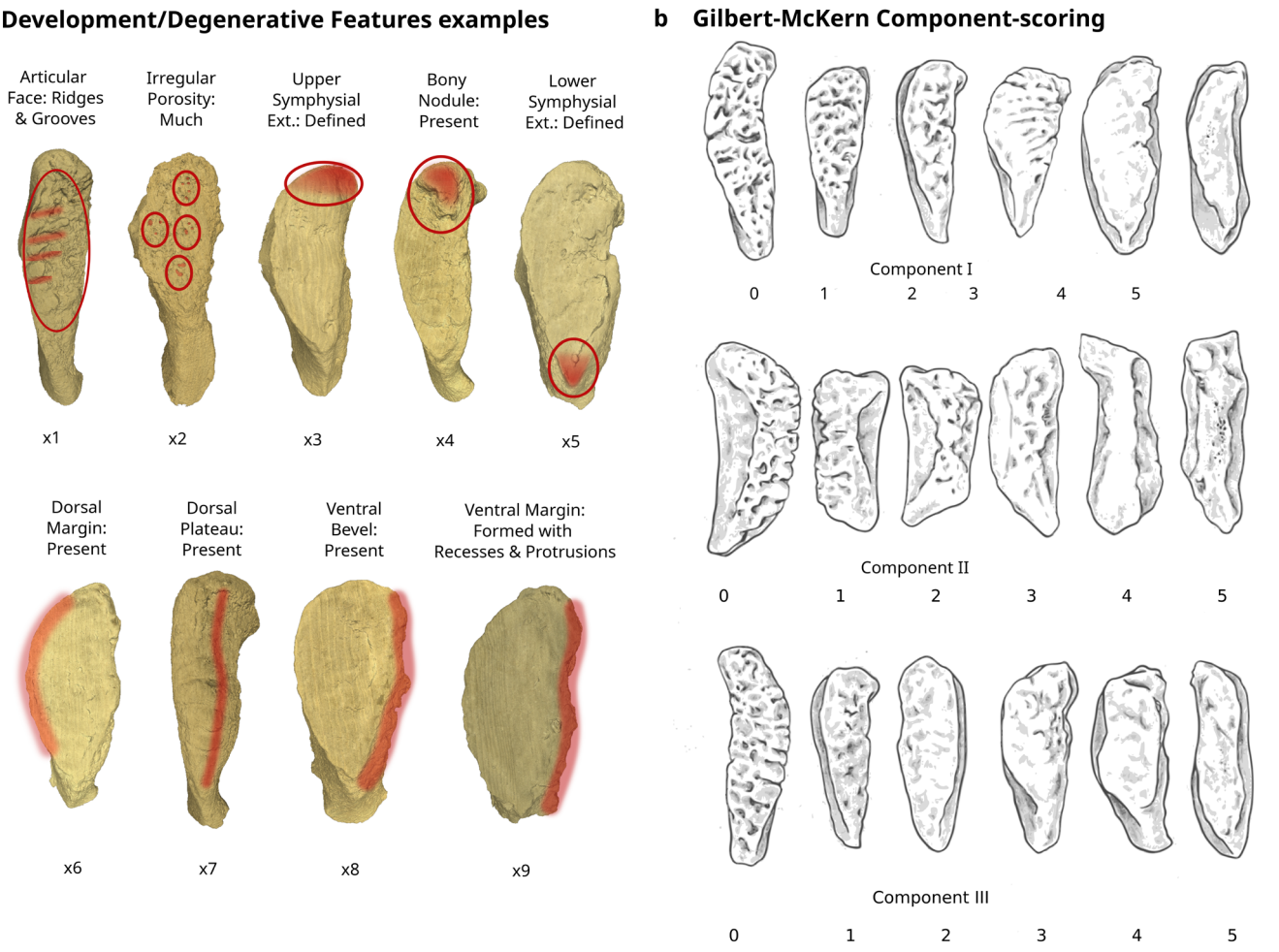

FIGURE 1 | Visualisation of relevant traits categorised for pubic symphysis in the literature and the components considered in our proposal. (a) Common development and degenerative morphological changes in the pubic symphysis during growth (relevant traits marked in red). (b) Gilbert–McKern component scoring for age-at-death estimation (reproduced from Gilbert and McKern (1973)).

2 | Symbolic Regression With GP

This section is devoted to briefly review the basics of the explainable machine learning (XML) approach considered in the current contribution. To do so, we will first devote a short subsection to introduce the discipline of explainable artificial intelligence (XAI) and XML, which allows us to categorise the position of GP-based symbolic regressors in the field. Then, we discuss the fundamentals of classical GP, bloat control, and GA-P as important tools to design transparent ML models.

2.1 | XAI and XML

The area of ML, based on designing intelligent systems with the ability to learn without being explicitly programmed, is currently one of the key branches of artificial intelligence (AI). Not long ago, constructing a traditional ML system required careful engineering and considerable domain expertise to detect or classify patterns. On the contrary, in the last few years, deep learning methods have directly exceeded human capacity and accuracy on an increasing number of complex activities by learning data representations directly from raw data (LeCun et al. 2015). However, that impressive performance is generally achieved at the cost of a loss of model interpretability.

Obtaining interpretable and/or XAI systems is a real need in most (human-centric) decision support systems across different application fields, especially in those where decisions ultimately affect humans' lives such as medicine (FA, in our case), law, finance, security, social sciences and engineering. Humans are reluctant to adopt techniques that are not directly interpretable and tractable, given the increasing demand for trustworthy AI (Díaz-Rodríguez et al. 2023).

Trustworthy AI involves a series of principles to be adopted when deploying AI in real applications: fairness, accountability, transparency and ethics, and of course accuracy in problem-solving. Its effective implementation requires a detailed study of these requirements and of the AI technologies to be used to achieve better trustworthiness of the model outputs by the targeted audience in the specific application domain. Not all kinds of audiences understand model explainability in the same way, thus having different profiles to be covered in its design (Langer et al. 2021). As an example, the FA experts from our research team were more used to handling mathematical expressions while our AI experts preferred to work with rule-based

TABLE 1 | Pubic symphysis' features and categorical values (all the information but the numerical values come from Gámez et al. (2022)).

Variable name	Categorical values	Numerical values
x1 Articular face	Regular porosity (1)	Ridges formation (2)
	Ridges and grooves (3)	Grooves shallow (4)
	Grooves remains (5)	No grooves (6)
x2 Irregular porosity	Absence (1)	Medium (2)
	Much (3)	
x3 Upper symphysial ext.	Not specified (1)	Defined (2)
x4 Bony nodule	Absent (1)	Present (2)
x5 Lower symphysial ext.	Not specified (1)	Defined (2)
x6 Dorsal margin	Absent (1)	Present (2)
x7 Dorsal plateau	Absent (1)	Present (2)
x8 Ventral bevel	Absent (1)	In process (2)
	Present (3)	
x9 Ventral margin	Absent (1)	Partially formed (2)
	Formed without bony outgrowths (3)	Formed with few bony outgrowths (4)
	Formed with recesses and protrusions (5)	

descriptions. Due to that, the most recent XAI definitions incorporate the concept of audience: 'Given an audience, an XAI is one that produces details or reasons to make its functioning clear or easy to understand' (Barredo Arrieta et al. 2020).

XAI (Barredo Arrieta et al. 2020) encompasses producing ML models with a good interpretability-accuracy tradeoff via: (i) building white/grey-box ML models which are interpretable by design while achieving high accuracy or (ii) endowing black-box models with a minimum level of (post hoc) interpretability when white/grey-box models cannot achieve an admissible level of accuracy. A detailed survey is provided in Ali et al. (2023) for XML methods, software tools and recommendations within three different categories: (i) interpretable-by-design methods, (ii) model-specific post hoc methods and (iii) model-agnostic post hoc methods.

Some authors defend the use of ML models that are interpretable per se as the best approach to obtain a trustworthy, interpretable model (Rudin 2019). However, accuracy requirements can prevent their use and black-box counterparts as deep learning are mandatory. In those cases, post hoc explainability can be considered. It can include either general (explanation by simplification, feature relevance explanation and visual explanation) or ML model-specific approaches (Barredo Arrieta et al. 2020). A third alternative hybridising connectionist and symbolic paradigms (neural-symbolic learning) look very promising (Díaz-Rodríguez et al. 2022). These techniques enrich data fusion approaches both at data and knowledge level by endowing them with explainability.

The use of interpretable-by-design ML models is the approach followed in the current contribution by considering GP-based symbolic regression (Mei et al. 2022). Symbolic regression is a form of regression analysis that aims to represent the underlying relationship of data without a priori knowledge of the resulting mathematical expression, i.e. it is able to infer both the structure and the parameters used in the regression model. Most approaches commonly use evolutionary algorithms (EAs) as effective global optimization methods for exploring the complex space of mathematical expressions.

However, there are methods better suited to address the nature of symbolic regression. GP (Koza 1992) is a classic method that shares the main concepts of GAs relying on a tree-based grammar to represent individuals. Thus, genetic operators are applied to tree nodes and subtrees, encoding mathematical expressions as tree nodes and operands as terminal nodes. Alternatively, niching GA-P (Howard and D'Dangelo 1995) is a hybrid approach where the expression is handled by a GP algorithm, while the set of expression parameters are managed by the GA. The hybrid design was introduced to improve the estimation of non-numeric variables and to allow modifications to the contents of an expression. Both methods suffer from a well-known limitation named bloat. Bloat results in an excessive growth in terms of size for the individuals which leads to higher evaluation costs due to variable tree depths and overfitting the data. Commonly, bloat is handled by incorporating a parameter that controls the maximum tree depth allowed during the evolution process. Recent GP methods introduce novel bloat control mechanisms, such as GP-DA (Nguyen and Chu 2020), which is based on semantic approximation. The following subsections describe the main components of these approaches.

2.2 | Classical GP

2.2.1 | Coding Scheme

The mathematical expressions considered to estimate age-at-death are encoded in expression trees, whose terminal nodes are either one of the nine variables considered representing the pubic symphysis traits (x_1 to x_9, see Table 1) or a numeric constant, and whose inner nodes are the mathematical operators addition, subtraction, multiplication and division. Note that we have selected a small set of simple operators to preserve the interpretable nature of the obtained models for the forensic anthropologists.

2.2.2 | Selection and Replacement Scheme

It is based on the classical generational scheme, where an intermediate population is created from the current one by means

of tournament selection of size k, and the offspring population directly replaces the parents one considering elitism.

2.2.3 | Genetic Operators

The usual GP crossover is considered (Koza 1992), which is based on randomly selecting an edge in each parent and exchanging both subtrees from these edges between the both parents, as shown in Figure 2a. Two different mutation operators are considered: (a) random selection of an edge and random generation of a new subtree that substitutes the old one located in that edge, and (b) random change of the value of a node by another node of the same type: a random numeric value, another variable or another operator.

The coefficient mutation is developed using Michalewicz's non-uniform mutation operator (Michalewicz 1996) to apply strong alterations in the first stages of the algorithm and soft ones in the later stages. With $c_k \in [c_{kl}, c_{kr}]$ being the value selected for mutation in generation t, the mutated value is

$$c'_k = \begin{cases} c_k + \Delta(t, c_{kr} - c_k) & \text{if } a = 0, \\ c_k - \Delta(t, c_k - c_{kl}) & \text{if } a = 1 \end{cases}$$

where $a \in \{0, 1\}$ is a random number and the function $\Delta(t, y)$ returns a value in the range $[0, y]$ such that the probability of $\Delta(t, y)$ being close to 0 increases as the number of generations increases.

2.2.4 | Generation of the Initial Population

The whole population is composed of random expression trees with different sizes under the maximum predefined length. See Figure 2c for a representation of the maximum tree size, also considered for depth limiting as bloat control approach.

2.2.5 | Fitness Function

The fitness function is based on the mean squared error (MSE) of the encoded expression:

$$MSE = \sum_{i=1}^{N} (y_i - y'_i)^2$$

where y_i is the actual age-at-death of sample i in the data set and y'_i is the value predicted by model for such sample.

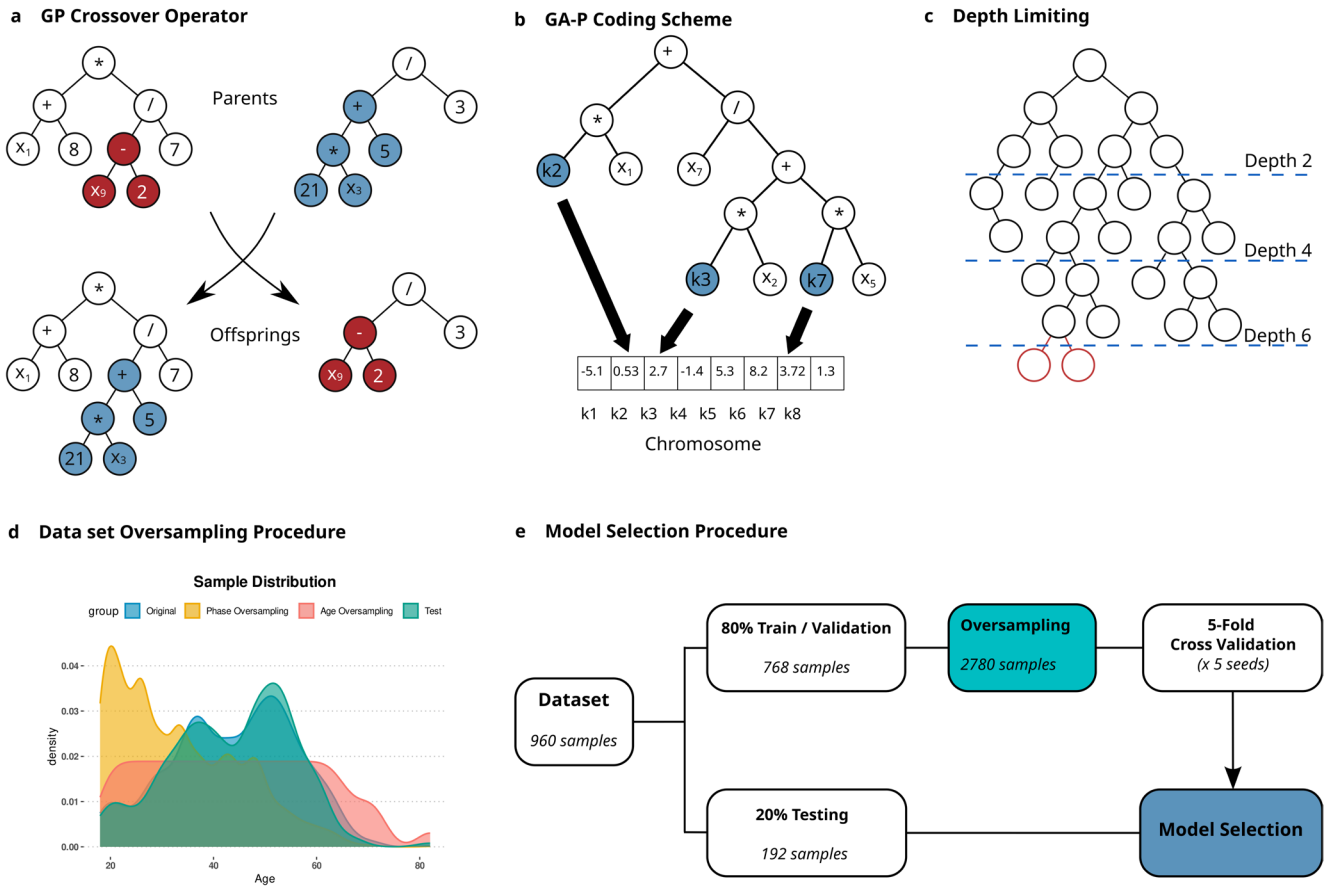

FIGURE 2 | Illustration of GP concepts and diagram of experimentation procedure. (a) Example of GP crossover operator between two solutions. (b) GA-P individual representing the expression $0.53x_1 + x_7 / (2.7x_2 + 3.72x_5)$. (c) Example of tree depth limiting approach to prevent bloat. (d) Age-at-death distribution of the pubic symphyses data set, before and after the different oversampling procedures considered. Note that the test set is extracted from the original sample and no oversampling is applied, so it maintains a similar distribution. (e) Data partition and experimental validation approach followed.

2.3 | Semantic Approximation for Bloat Control

Following the general design of the classical programming algorithm, GP-DA (Nguyen and Chu 2020) introduces a novel approach based on semantic approximation to deal with the problem of 'bloat'. Instead of limiting the maximum allowed tree depth, a new phase is included where random subtrees are replaced by smaller ones. The smaller trees are grown from a randomly generated tree until it is semantically approximate to a desired semantic vector (Pawlak et al. 2015). For a GP individual p, a semantic vector $s(p)$ is formally defined as a finite sample of outputs with respect to a set of training values:

$$s(p) = \left(p(k_1), p(k_2), \ \ldots, p(k_n)\right)$$

where $K = \{k_1, k_2, \ \ldots, k_n\}$ represents the set of fitness cases or inputs to the expression, and $p(k_i)$ is its output when applied to the fitness case k_i.

Thus, a subtree can be replaced with a semantically approximate one by minimising the squared distance between the target semantics and the semantics of the smaller tree, which involves optimising the function:

$$f(\theta) = \sum_{i=1}^{N} \left(\theta \cdot q_i - s_i\right)^2$$

where $S = (s_1, s_2, \ \ldots, s_N)$ corresponds to the target semantics vector and $Q = (q_1, q_2, \ \ldots, q_N)$ to the semantics vector of the smaller tree, while θ is the scaling factor that allows to align both semantics.

GP-DA is then based on the concept of finding the desired semantics for the subtree using backpropagation over the computed semantic approximation (Pawlak et al. 2015). The desired approximation method enhances diversity by introducing structurally distinct smaller trees, enabling exploration of new regions in the search space while restricting code bloat. These internal mechanisms of GP-DA contribute to reducing model complexity while maintaining performance.

2.4 | Niching GA-P

Most GA-P components are the same as in any of the traditional EAs. The GA-P and GP perform selection and offspring generation in a similar manner, except that the structure of the GA-P requires separate crossover and mutation operators for the expression and coefficient parts.

2.4.1 | Coding Scheme

The expression part (GP part) encodes the considered formula, i.e. variables and mathematical operators, and the coefficient string (GA part) represents the numerical values used on it, as shown in Figure 2b. We consider a real coding scheme for the GA part.

2.4.2 | Selection, Niching and Replacement Scheme

Alternatively to the GP variant, the GA-P selection is based on the steady-state approach (Smith and Eiben 2010). At each generation, two parents are selected using tournament selection, two offspring are generated by crossover and mutation and they compete with the two worst individuals in the current population for these two positions.

A niche-based GA-P implementation is considered to account for the strong relationship that exists between the expression and the coefficient parts. The algorithm performs better when applying coefficient crossover on two individuals having a similar GP part (Sánchez Ramos and Corrales 2000). Thus, we consider each individual with the same expression part to belong to the same population niche.

The GA-P algorithm incorporates two different crossover operators. The *intra-niche crossover* is applied when both individuals belong to the same niche, and in this case, only the GA parts are crossed over to exploit the search space around the learned expression encoded in the GP part. Alternatively, the *inter-niche crossover* is applied for individuals encoding different expressions, i.e. belonging to different niches. In this case, both the GP and GA parts are crossed over, thus leading to an exploration of the entire search space to introduce diversity in the GA-P population.

In each generation, a single individual is randomly selected using the universal stochastic procedure—the usual roulette wheel, as we prefer to induce a high selective pressure. Then, the crossover type is randomly selected according to a probability P_{intra_cross}. On the one hand, to perform an intra-niche crossover, another individual is selected from the same niche, crossover is applied on the GA parts and both GA parts are also mutated with probability $P_{\mathrm{m}}^{\mathrm{GA}}$. Then, a competition is established between the two parents and the two offspring to take the place of the former in the population. If there is a draw in the fitness value, simpler expressions are preferred. On the other hand, if an inter-niche crossover is chosen, another individual belonging to a different niche is selected. Crossover is performed on the GA and GP parts and both parts are mutated with probabilities $P_{\mathrm{m}}^{\mathrm{GP}}$ and $P_{\mathrm{m}}^{\mathrm{GA}}$, respectively. In this case, the two offspring compete with the two worst individuals in the population.

2.4.3 | Genetic Operators

BLX-α is considered for the GA parts in the intra-niche crossover (Eshelman and Schaffer 1993). This operator generates an offspring, $C = (c_1, \ \ldots, c_n)$, from two parents, $X = (x_1, \ \ldots, x_n)$ and $Y = (y_1, \ \ldots, y_n)$, where c_i is a number chosen at random (uniformly) from the interval $[min_i - I \cdot \alpha, max_i + I \cdot \alpha]$, where $max_i = max\{x_i, y_i\}$, $min_i = min\{x_i, y_i\}$ and $I = max_i - min_i$ ($[min_i, max_i]$) are the interval where the ith gene is defined. The operator is applied twice to obtain two offspring. Michalewicz's mutation is again considered to perform mutation in the GA part.

3 | Results and Analysis

3.1 | Data Set Description and Pre-Processing

The collection of skeletonized pubic symphyses acquired and preserved at the Physical Anthropology lab at the University of Granada (Spain), among the largest worldwide, was considered as data set for this experimentation. It consists of autopsy studies compiled since 1991 in collaboration with the Institute of Legal Medicine and Forensic Sciences. The total sample includes 837 individuals from 17 to 82 years. However, the sample was limited to 600 individuals after a careful filtering of cases with unreliable antemortem information or deficient preservation conditions developed by the forensic experts. Both left and right lateralities are considered, as each side develops independently and can exhibit asymmetry in degenerative changes due to factors such as biomechanical loading, injury, or lifestyle. After accounting for cases missing one side, the final number of samples is 960 (473 of the left laterality and 487 of the right one).

In his pioneering contribution, Todd (1920) developed a detailed description of the development and degenerative processes of the pubic symphysis, along with the morphological aspects of the bone associated with them. Nevertheless, these descriptions were very generic and relied on the forensic anthropologists' expertise to evaluate them through visual inspection of the samples and comparison with prototypical bones. To systematise those descriptions and to reduce subjectivity in their evaluation, the forensic experts in the team developed a pubic symphysis atlas where the nine traits associated with the development and degenerative changes in Todd's studies were extracted and clearly identified. Each of those traits (i.e. variables) was associated with two or more categorical values representing the possible morphological changes of the specific pubic bone characteristic, resulting in the information collected in Table 1 (the interested reader is referred to the work (Gámez et al. 2022) for additional information on the knowledge elicitation process followed).

Using this pubic symphysis atlas, the sample was annotated by two forensic anthropologists who blindly labelled the considered traits in the pubic bone in random order without any knowledge regarding the actual age-at-death. To design a component-scoring age-at-death estimation method based on symbolic regression, there is a need to transform these categorical values into numerical values, as done in Gilbert–McKern's method (see Section A.2). To do so, we will simply assign a numerical value to the categorical label of each trait between one and the maximum number of possible values for that feature (see numbers inside brackets in Table 1). In particular, a feature with two possible values will be assigned 1 or 2, depending on which of the two possible states the pubic bone shows for that trait.

The age-at-death distribution of the annotated data set is displayed in Figure 2d. The curve named *Original* corresponds to a strongly imbalanced data set, as the number of annotated samples notably varies across the age-at-death range, particularly at older ages (as of course expected). This situation often happens in real-world problems and it increases their complexity. A specific treatment (data pre-processing) of the imbalanced nature of the problem is required since the direct use of a transparent ML method over the original data set would result in deceptive models. The interested reader is referred to an extensive and detailed experimental study to analyse the influence of this issue in problem solving (Gámez et al. 2022).

In ML, a common approach to deal with imbalanced data is oversampling techniques (He and Ma 2013; López et al. 2013). Specifically, we used a random replication approach, which was selected based on its good performance in the prior study. This approach was also favoured by forensic experts due to its ability to avoid introducing artificial instances into the data set. Figure 2e summarises the data partition procedure followed to separate the data used for the learning process and the test set. As depicted, the final test will be performed with original samples only, unseen until the last stage. Meanwhile, the learning data was augmented by using random oversampling to a rebalanced data set composed of 2780 samples (1331 of the left laterality and 1449 of the right one). The age-at-death distribution of the oversampled data set is shown in Figure 2d, labelled as *Phase Oversampling* as this process balances the data across age groups. Note that the oversampled instances will be only considered for training and validation while the test set maintains a distribution similar to the original set of samples.

3.2 | Experimental Setup

Next, we describe the procedure considered for a quantitative analysis of our proposal and to perform model selection by following a standard strategy to avoid over-fitting the data (Kuhn and Johnson 2013). It should be noted that we are not only interested on an experimental test allowing us to validate the applied transparent ML approaches but also to select a faithful final model to be used by a forensic anthropologist in her usual work. The original data set is partitioned by extracting a randomised 20% of the samples as test set, which will remain unseen until the model selection has been performed. The remaining 80% is considered to build a series of models maximising their generalisation ability. A 5-fold cross-validation (5-CV) (Mosteller and Tukey 1968) procedure is then applied over this second data set as follows: five random partitions at a 20% are obtained and combined to compose five different training-validation partitions at 80%–20% (i.e., four folds for training and one for validation in each partition) within the 5-CV strategy. In addition, to provide an overview of the methods' robustness and ensure reproducibility, five different runs for each GP method are executed by considering different seeds. As a consequence, 25 different runs are made for each GP method in each experiment and thus 25 different models are learned, resulting from five runs on the five 80%–20% training-validation partitions. Finally, the model obtaining the best validation error from those 25 models is selected for its final validation over the unseen test data set composed of 20% of the original samples.

Following the described experimental setup, the performance of the considered methods (GP, GP-DA and GA-P) are analysed in similar conditions. The specific parameters for GP-DA are a maximum depth of 20 nodes. Different configurations have been considered, by restricting the maximum tree depth at 20, 40 and 60 nodes for GP and GA-P. The parameter values considered are the probability to generate a variable in the initial expressions is 0.3 for GP, the crossover and mutation probabilities are 0.75

and 0.05 (both for the GP, GP-DA and the GP part in GA-P), the intra-niche crossover probability for GA-P is 0.03, and the tournament size is 100. Finally, the population size for all the methods is set at 1000 individuals and the stopping criterion for each individual experiment is set at 1,000,000 evaluations. The experiments have been developed in an Intel Core i7-10870H CPU @ 2.20GHz computer with 8 cores.

The resulting models are analysed by considering two complementary performance metrics to measure the magnitude of the errors, commonly considered in the area:

- Rootmeansquarederror(*RMSE*):$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - y_i\right)^2}$
- Mean absolute error (*MAE*): $MAE = \sum_{i=1}^{N} \mid x_i - y_i \mid$

3.3 | Model Validation

A 5-CV approach has been performed to train and validate a set of expressions to estimate age-at-death. Results for five executions of 5-CV with different seeds are reported in Table 2, also partitioned according to three different configurations for maximum tree depths (limited to 20, 40 and 60 nodes). Note that GP-DA, because of its bloat control, internally restricts the tree growth and its results are under 20 nodes. Validation results are shown for two error metrics, i.e. RMSE and MAE.

Given the similar accuracy of the results across methods and tree depths (see Figure 3a), the characteristics of the resulting models in terms of expression complexity become particularly relevant. It can be argued that despite tree sizes of 60 nodes obtaining better results across the cross-validation experimentation, the advantages in terms of metric errors are insufficient when considering the execution times are tripled. Furthermore, the resulting models are certainly more complex and less interpretable than limiting the depth at 20 nodes.

In contrast, GP-DA stands out for achieving a slight advantage in performance as shown in Figure 3b, providing better validation results than both GP and GA-P with a smaller tree depth. GP-DA is able to restrict tree size between 12 and 20 nodes thanks to its bloat control approach.

3.4 | Model Selection

We now rely on the 5-CV procedure to select the seven best-performing models corresponding to each combination of method and tree depth. Table 3 reports the results of the comparison among the best models so far, which have been evaluated using an unseen test data set. In addition, the table includes the results of classical approaches, such as linear regression (LR), support-vector machines (SVM), DT and random forests (RF), to provide a broader evaluation of model performance within the context of established ML techniques. Overall, classical methods like LR and RF provide reasonable performance but fall short compared to SR methods. Regarding interpretability, LR sets the baseline for simplicity, providing an explicit regression formula that assigns weights to all the variables. The structures of DT and RF models are illustrated in Figures B1 and B2 for comparison. This highlights the potential of GP and GA-P for complex tasks requiring flexible representations. According to the validation results, both GP and GA-P achieve comparable outcomes for each tree depth, with GA-P standing out in two of the tree comparisons. However, GP-DA is not able to translate its initial advantage

TABLE 2 | Averaged 5-CV results and best models found for each method and tree depth.

		Depth 20		Depth 40		Depth 60	
Method	**Seed**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
GP	8324	7.69	5.71	7.64	5.69	7.65	5.71
	12,345	7.68	5.74	7.62	5.63	7.75	5.74
	34,634	7.71	5.71	7.64	5.68	7.68	5.69
	34,679	7.70	5.79	7.66	5.69	7.63	5.68
	92,034	7.71	5.80	7.70	5.71	7.65	5.65
GA-P	8324	7.72	5.73	7.74	5.74	7.64	5.64
	12,345	7.81	5.83	7.64	5.65	7.63	5.63
	34,634	7.77	5.75	7.64	5.68	7.64	5.68
	34,679	7.82	5.78	7.69	5.72	7.66	5.76
	92,034	7.75	5.77	7.66	5.65	7.64	5.65
GP-DA	8324	8.27	5.59	—	—	—	—
	12,345	7.39	5.27	—	—	—	—
	34,634	7.58	5.56	—	—	—	—
	34,679	7.44	5.41	—	—	—	—
	92,034	7.38	5.23	—	—	—	—

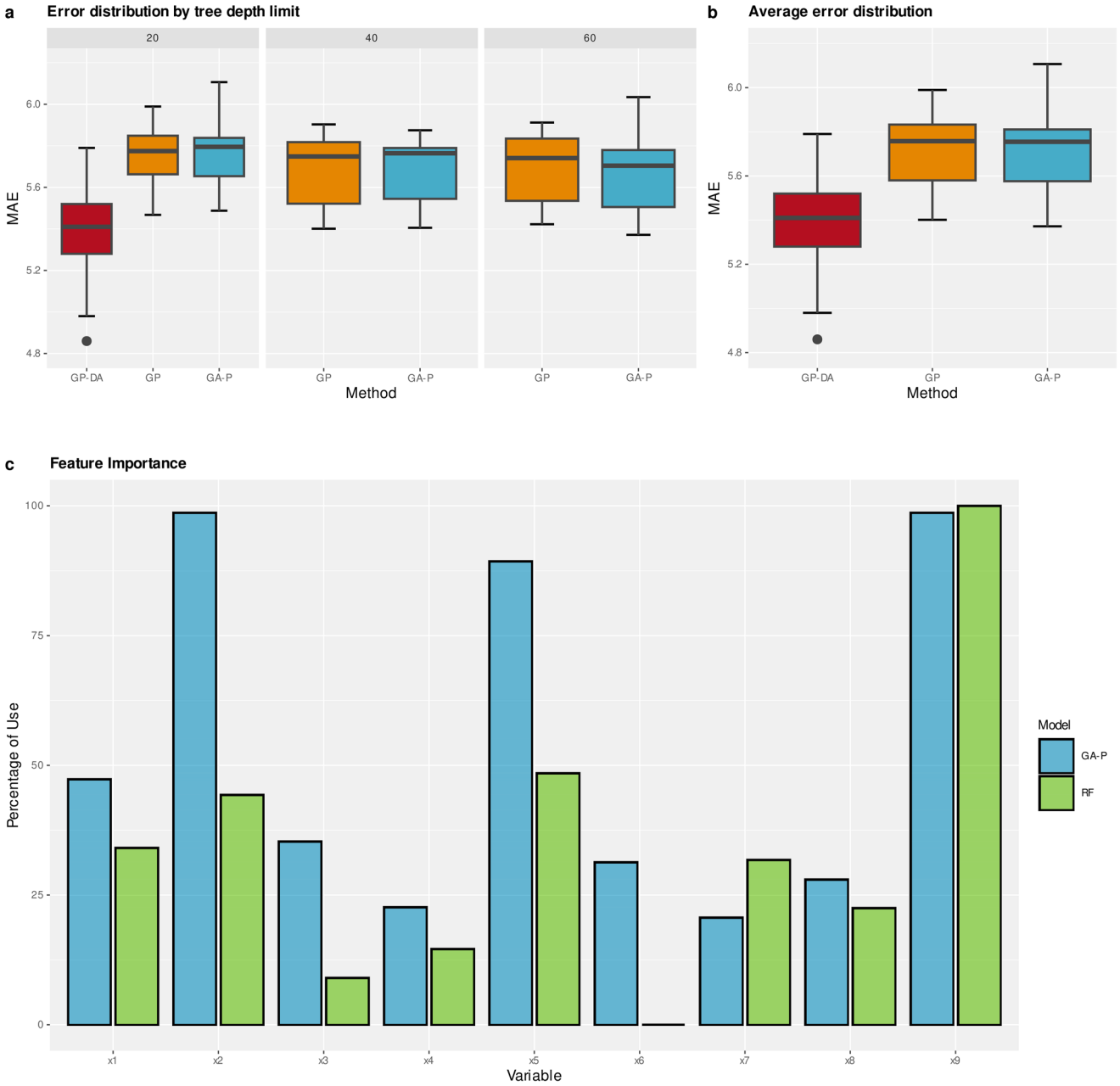

FIGURE 3 | Visual analysis of the results. (a) Distribution of errors according to the three tree depth limits considered. (b) Average distribution error (MAE in years) for GP-DA, GP and GA-P. (c) Frequency histogram of model variables occurrence in the 175 experiments performed during the 5-CV procedure. Box plots represent the interquartile range (IQR) highlighting the median value, and whiskers extend to $\pm 1.5 \times$ IQR beyond the box.

when generalising to the unseen data set. The best outcomes are achieved at a tree depth of 20 nodes, with GA-P taking the lead by a negligible difference. Precisely, the best test results are 10.81 and 8.55 years according to RMSE and MAE, respectively. This is a confirmation of the usual assumption that simpler models have a better generalisation ability. Therefore, in our case, the lowest configuration of tree depth not only offers more readable models but also performs better than the remaining models with a larger tree size thanks to dealing with a lower number of parameters. Nevertheless, such assumption does not hold true for GP-DA, whose best result is comparably more complex than the rest. While maintaining a smaller tree depth, the semantic approximation approach tends to result in a more convoluted combination of features, which negatively impacts its readability.

Statistical analysis was conducted to assess performance differences among ML models. Specifically, the Wilcoxon signed-rank test was employed to compare paired results across models, ensuring a nonparametric and robust assessment of statistical significance, considering a standard confidence level of 95%. Table 4 summarises the results of the Wilcoxon test. It reveals significant differences among most of the models, but LR and DT with a similar performance. Notably, while most comparisons among the GP and GA variants showed significant differences, a subset of comparisons, particularly between variants at adjacent depth levels, did not reach statistical significance. These results underscore the superior performance of SR approaches over traditional ML methods, while also highlighting nuanced performance variations within GP and GA-P approaches.

TABLE 3 | Final test results for each method and depth considering the best model found during validation.

Depth	Method	RMSE	MAE	Best model
—	LR	10.86	8.60	$7.6 + -0.99x_1 + 5.11x_2 + 0.34x_3 + 0.22x_4 + 4.14x_5 + 1e^{-17}x_6 + 0.71x_7 + 0.47x_8 + 4.99x_9$
—	SVM	11.67	9.13	2082 support vectors
20	DT	10.97	8.66	see Figure B1
20	RF	10.90	8.63	see Figure B2
20	GP	10.82	8.56	$\frac{-7.13x_2 - x_3 + (x_5 - 3.56)(x_3 + 7.13x_9 + 8.45)}{x_5 - 3.56}$
	GA-P	10.81	8.55	$6.06x_2 + x_3 + 6.06x_5 + 5.06x_9 + \frac{x_8}{x_2}$
	GP-DA	10.84	8.55	$\frac{x_6(x_1^4x_2^5x_9(x_3+x_4) + x_1^4x_2^5(2.2x_2 + 2x_5 + x_6^2 + 0.1x_6x_7) - 0.9x_1^4x_4 + 0.8x_1^2x_2^6(x_5 - x_9) + 0.8x_2^5(x_1^3(-x_3 + 1.5x_4) + 0.4x_8))}{x_1^4x_2^5}$
40	GP	10.93	8.63	$\frac{1.48(0.67x_2x_6(x_2 - x_5)(7.52x_2 + x_4 + 7.52x_9) + 6.33x_2(x_2 - x_5) - x_6(0.8x_8 + 1.96x_9(x_2 - x_5) + 0.8x_9 + 2.96))}{x_2x_6(x_2 - x_5)}$
	GA-P	10.83	8.55	$\frac{x_1(x_3 - 1.96)(5.51x_2 + 4.82x_5 + 5.69x_9 + 5.29) + 0.1}{x_1(x_3 - 1.96)(1.74x_6 - 2.41)}$
60	GP	10.87	8.61	$0.74x_1x_2 - 2x_1 + 2x_2 - \frac{x_2}{\frac{x_2}{8.99 - \frac{x_5x_9}{x_4}} - 1.29} + 4x_5 + x_8 - 8.66 - \frac{26.4}{0.12x_9 - 1.14}$
	GA-P	10.90	8.66	$-x_1 + 4x_2 + 3.96x_5 + x_7 + x_9 + \frac{x_2 + 12.4}{\frac{0.12x_1}{x_2} + x_6 + 0.12 + \frac{x_6}{x_2}} + 6.71 + \frac{0.13}{-x_2 + x_7 + 1.95} + \frac{8.4x_9}{x_6}$

TABLE 4 | Pairwise Wilcoxon's test results between the classical ML methods compared and the SR approaches.

		—	—	20	20	20			40		60	
Depth	Method	LR	SVM	DT	RF	GP	GA-P	GP-DA	GP	GA-P	GP	GA-P
—	LR	—										
—	SVM	$< 10^{-8}$	—									
20	DT	**0.26**	$< 10^{-8}$	—								
20	RF	$< 10^{-6}$	$< 10^{-8}$	$< 10^{-8}$	—							
20	GP	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	—						
	GA-P	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	**0.43**	—					
	GP-DA	$< 10^{-4}$	$< 10^{-8}$	$< 10^{-4}$	$< 10^{-5}$	$< 10^{-3}$	$< 10^{-3}$	—				
40	GP	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	**0.06**	0.01	$< 10^{-3}$	—			
	GA-P	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	**0.98**	**0.18**	$< 10^{-3}$	**0.35**	—		
60	GP	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	**0.11**	**0.19**	$< 10^{-3}$	$< 10^{-3}$	**0.1**	—	
	GA-P	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	**0.08**	0.01	$< 10^{-3}$	**0.79**	**0.27**	$< 10^{-3}$	—

Note: p-Values are shown in exponential notation, no significant differences are highlighted in bold.

A detailed visual analysis of the prediction performance for the best regression model can be seen in Figure 4a, which shows the variability of age-at-death estimations obtained for each age value. The graph elicits a clear pattern consistent with other studies (Kotěrová et al. 2022; Schmitt et al. 2002; Djurić et al. 2007), i.e. the model tends to overestimate younger individuals and underestimate older individuals. However, the estimations are better for individuals in the range 18–22. A likely explanation relates to the easier identification of Todd's bone traits unaffected by degenerative processes. Overall, the model displays a tendency towards the mean age-at-death values, which might be attributed to high difficulty of the problem due to the complex identification of the bone's development and degenerative processes (as well as to the lack of appropriate datasets). This compensatory tendency is a well-known issue in regression-based age-at-death estimation as identified by Aykroyd et al. (1997). Figure 4a also highlights the strongly imbalanced nature of the problem at hand. Despite considering one of the largest data collections of pubic symphyses, some ages are heavily underrepresented.

From a practical point of view, the differences in the error assumed by each of the models are practically the same.

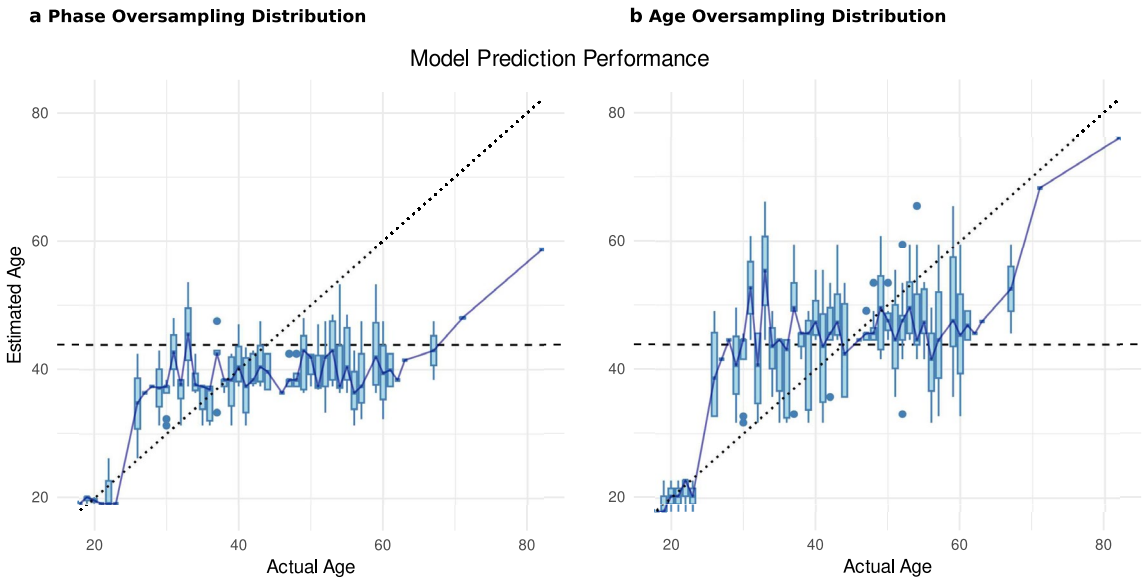

FIGURE 4 | Visual analysis of prediction performance achieved by the selected models. The image shows the distribution of age-at-death estimations obtained by GA-P (Depth 20) over the test set for the two distributions considered: Phase oversampling (a), and Age oversampling (b). For each age, a box plot shows the distribution of the model estimation error, i.e. the plot summarises the minimum, maximum and interquartile range (IQR). The blue straight line shows the prediction trend connecting median values, the horizontal dashed line delineates the mean age of the test set, while the black dotted line is used as a reference for the correct estimation.

However, the first and second models (Table 3) allow us to obtain very good estimates by only using four and five features, respectively, which may be an advantage for estimating the age-at-death of degraded or incomplete remains, a frequent situation in contexts of FA. In this sense, the choice of one or the other model as final method for age-at-death estimation may depend on the state of conservation of the pubis rather than on the error assumed with each one, selecting those variables that are better defined or easier to identify (except for variables $x2$ (irregular porosity), $x5$ (lower symphysial extremity) and $x9$ (ventral margin), which are considered by every model). This gives greater versatility to our method compared with classical phase-based methods, which commonly require the use of all variables, as already mentioned.

The finally selected model can be expressed by the following equation:

$$Age = 6.06 \cdot I_P + U_{SE} + 6.06 \cdot L_{SE} + 5.06 \cdot V_M + \frac{V_B}{I_P} \quad (1)$$

where I_P represents the irregular porosity ($x2$), U_{SE} the upper symphysial extremity ($x3$), L_{SE} the lower symphysial extremity ($x5$), V_M the ventral margin ($x9$) and V_B the ventral bevel ($x8$). We will select this model due to its simplicity, which consequently helps to make it more interpretable. As said, only 5 different pubic symphysis variables are considered. In addition, the expression is simple as in all the cases but one the variables are simply multiplied by a factor and added together. The only compound trait is a ratio between the ventral bevel and the irregular porosity values, that is directly incorporated into the remaining expression. Besides, we can acknowledge that the most influential features are the irregular porosity ($x2$) and the lower symphysial extremity ($x5$), according to the coefficients in the expression. Notably, the former plays a double role, as it also acts as a modifier of the ventral bevel ($x8$) trait in the denominator of the $\frac{V_B}{I_P}$ variable.

3.5 | Analysis of Feature Importance

In addition, we analyse the mathematical expressions derived from extensive experimentation, aiming to gain insights into the relevance of specific bone traits. To this end, we conduct a detailed comparison of GA-P, the best-performing method, with RF, which serves as the baseline for classical ML methods. As seen in Figure 3c, in the case of GA-P, three variables stand out with more than 75% usage, which correspond to the bone traits irregular porosity (variable $x2$), lower symphysial extremity ($x5$) and ventral margin ($x9$). We can state that these three traits are particularly relevant for age-at-death estimation from the pubic symphysis. In contrast, traits such as bony nodule ($x4$) and dorsal plateau ($x7$), are the least used by GA-P with fewer than a 25% usage.

These results are in agreement with those reported in the previous work (Gámez et al. 2022). In that contribution, the same nine variables (bone traits) were considered to learn a rule-based system comprising an automatic phase-based age-at-death estimation method. A variable-use study was also developed for the considered methods. Even though it was more segmented, as the use of the traits was measured at phase level, that study showed that traits dorsal margin ($x6$), upper symphysial extremity ($x3$), bony nodule ($x4$), lower symphysial extremity ($x5$) and dorsal plateau ($x7$) were the least used, in that order. In fact, the dorsal margin ($x6$) was never used in any model obtained from the two different rule learning methods used, including all the trees in RF. The main conclusion of that study is that certain traits were better suited for estimating young or elderly individuals, while others were applicable across the entire age range.

The plot in Figure 3c illustrates a similar idea, aligning with the authors' findings, as RF excludes the use of the dorsal margin trait ($x6$). A similar pattern is observed with LR, which also assigns no value to this variable. RF trees give less importance to the upper symphysial extremity ($x3$), bony nodule ($x4$) and the dorsal margin ($x6$). However, the models heavily rely on the ventral margin ($x9$) to guide the estimation process, while GA-P reports a higher balance between the relevant traits, which in turn translates into better overall performance than RF models.

The main point of agreement between both studies is the relevance of the ventral margin for age estimation. This result also coincides with the one published by Castillo et al. (2021), where authors evaluated the performance of a similar scoring system designed using different ML methods. However, we found two important differences. On the one hand, we have concluded that irregular porosity ($x2$) is the most used variable in the current study according to the results in Figure 3c, while both in the previous study (Gámez et al. 2022) and Castillo et al. (2021), the use of this variable is less frequent and is mainly considered to estimate the age-at-death of older subjects. On the other hand, the opposite occurs with the articular face ($x1$) and with the ventral bevel ($x8$), while they are two of the most used variables (for all ages) in the two previous studies (Gámez et al. 2022; Castillo et al. 2021), when we apply regression equations, as in this work, their use is greatly reduced.

3.6 | Benchmarking and Overview

As a final comparison, Table 5 summarises the relation of the best models found in this contribution for GP, GA-P and GP-DA with the age estimation methods reviewed earlier. Though the proposed benchmarking is a rough and generic comparison, it will allow us to draw a valuable overview on the accuracy of the methods. First of all, the comparison involves methods of different typology (phase-based and numerical, as well as global and component-scoring) which were tested by using a different validation methodology (leave-one-out cross validation [LOOCV], single 50% training-test partition, use of the samples of one pubic symphysis laterality for training and those of the other laterality for test, and 5-fold CV). Moreover, the size, age range and distribution according to sex or ethnic groups of the samples are key differences. Setting aside such differences, our proposal stands out in the comparison.

Up to now, the two ML methods proposed by Kotěrová (SASS and AANNESS) were the current state-of-the-art methods solely based on the pubic symphysis. SAAS, being an interpretable method (as the one proposed in this study) based on multi-LR, provided an RMSE of 14.3 years and an MAE of 11.7 years. Meanwhile, AANNESS, a black box model that does not directly permit human interpretation, achieved a RMSE of 12.9 years and a MAE of 10.6 years. In addition, the reported error corresponds to a 5-CV procedure, and not to a single, finally selected age-at-death estimation model whose expression is actually validated in an unseen data set. In our experiment, following a sounder statistical validation procedure, GA-P proved to be an accurate approach by achieving an RMSE of 10.81 years and an MAE of 8.55 considering a transparent regression model structure, surpassing both the interpretable and black box reference proposals by more than 2 years according to both metrics.

In any case, we do not consider our method as a competitor to the black-box AANNESS approach proposed by Kotěrová et al. as ours requires an expert labeling of the considered pubic symphysis traits, while theirs permits fully automatic estimation of those traits, thus reducing subjectivity and error-prone human activity. Thus, both proposals can be considered totally complementary approaches to address the difficult task of age-at-death estimation from the pubic symphysis.

TABLE 5 | Comparison between the best proposed methods, and the state-of-the-art results.

Method	Type	Exp setup	# Samples	Age range	RMSE	MAE
Slice and Algee-Hewitt (2015)	N	LOOCV	41	19–96	17.15	—
Stoyanova et al. (2015)	N	LOOCV	56	16–100	19	—
Stoyanova et al. (2017)	N	50%–50% split	93	16–90	13.7–16.5	—
Kotěrová et al. (2018)	CS,N	5-CV (w/o test)	941	19–100	12.1	9.7
Kotěrová et al. (2022) SAAS	CS,N	5-CV (w/o test)	483	18–92	14.3	11.7
Kotěrová et al. (2022) AANNESS	N	5-CV (w/o test)	483	18–92	12.9	10.6
Gámez et al. (2022)	PB	tra: right lat.-test: left lat.	892 (439–453)	18–60	13.19	10.38
Gámez et al. (2022)	PB	tra: right lat.-test: left lat.	960 (487–473)	18–82	14.61	11.62
GP (Depth 20)	CS,N	5-CV tra-val-test	960 (614–154-192)	18–82	10.82	8.56
GA-P (Depth 20)	CS,N	5-CV tra-val-test	960 (614–154-192)	18–82	10.81	8.55
GP-DA (Depth 20)	CS,N	5-CV tra-val-test	960 (614–154-192)	18–82	10.84	8.55
GA-P (Depth 20) AD	CS,N	5-CV tra-val-test	668 (381–95-192)	18–82	9.54	7.51

Abbreviations: 50%–50% split, single 50% training-test partition; 5-CV, 5-fold cross-validation; AD, alternative distribution; CS, component-scoring; LOOCV, leave-one-out cross-validation; *N*, numeric, PB, phase-based; tra, XXX-test, XXX, use of the samples of one pubic symphysis laterality for training and those of the other laterality for test.

As expected, the proposed component-based models also outperform the phase-based method introduced in the previous study (Gámez et al. 2022). The goal of that study was to propose a transparent model replicating Todd's original method operation in the form of a set of rules, not to obtain an accurate, automatic age-at-death estimation method from the pubic symphysis to assist the forensic anthropologist. As previously discussed, the use of component-based methods had already shown a better performance in the specialised literature (Shirley and Ramirez 2015; Fojas et al. 2018).

3.7 | Designing a New Model

Our prior experiment illustrates the suitability of the proposed methodology in addressing the age-at-death estimation problem. It further demonstrates that employing a powerful GP algorithm is unnecessary. In addition, the derived models are able to adapt and select pertinent variables, yielding an intuitive expression that is straightforward to employ. However, the behaviour of the model is rather compensatory and shows a tendency towards the mean. Given this observation, we intend to capitalise on the current situation. Instead of concentrating oversampling efforts on the younger individuals to maximise the sample count, we will shift our focus to the elderly population. With this change, we aim to provide a better data distribution for the learning model and adapt the method to modern populations.

For a more thorough exploration of the tentative expression space, we will follow a different approach by combining age-targeted undersampling and oversampling strategies. This results in a new uniform distribution in which each age value is represented by 21 samples. Specifically, the undersampling step reduces the dataset to 476 samples in the middle-age range, while the oversampling step yields a total of 1113 samples for training purposes to avoid individuals over 64 years being considered outliers. This refined preprocessing leads to the distribution depicted in Figure 2d, labelled as *Age Oversampling*.

We follow the same experimental set as in our previous experiment (5-CV) considering the new data distribution. The behaviour of the algorithms is similar and GA-P (Depth 20) also achieves the best performance among the GP methods. The last row of Table 5 summarises the results for an overview comparison. In particular, the test results are 9.54 and 7.51 years according to RMSE and MAE, respectively. Hence, the refined preprocessing allows the method to improve its performance and to achieve an even lower test error. As seen in Figure 4b, while the model effectively captures the overall trend, the compensatory effect towards the mean is still present. However, the predictions for both younger and older age groups exhibit a significant improvement. Meanwhile, the resulting equation can be expressed as follows:

$$Age = V_M\left(I_P + 4.88B_N + 4.88D_M - V_M + 3.22 + \frac{I_P - V_M}{V_B}\right). \quad (2)$$

In contrast to Equation (1), both variables related to the symphysial extremities are replaced by B_N, the bony nodule ($x4$) and D_M, the dorsal margin ($x6$). The rest of the bone traits from the previous model are also present. These observations further support the notion that there exist certain traits more appropriate for specific age ranges. Symphysial extremities gained relevance in the initial oversampling approach that favoured younger ages. However, with the refined preprocessing the target focus has shifted and the resulting model is able to identify more suitable traits, selecting the bony nodule and the dorsal margin to better represent both the younger and the older age groups, respectively. These findings are also in agreement with the previous study (Gámez et al. 2022) and Castillo et al. (2021), where the dorsal margin was identified as a trait that better allowed to estimate the age-at-death for older individuals. Altogether, they suggest that the adoption of hierarchical approaches that incorporate partial decisions to effectively segment age groups might perform better than examining all features at a single level since, as we have seen, biasing the data sample to focus on a specific age range at a lower or higher degree results in the use of different traits.

4 | Conclusions

During the last two decades, in the context of FA, an exhaustive and necessary methodological revision process has been carried out regarding the estimation of the BP (Lesciotto 2015; Liversidge et al. 2015; Corron et al. 2018; Ubelaker and Khosrowshahi 2019). This process has focused on several key aspects, including the use of new technologies, improving the results in terms of precision and accuracy, and ensuring the admissibility of expert evidence through a precise calculation of assumed error. Although this updating process has gained widespread acceptance within the scientific community, a vigorous debate persists regarding the optimal design of age estimation methods. Here are some examples:

Classical or inverse calibration seems to make us choose between a smaller error or an apparently unavoidable bias, overestimating age in young individuals and underestimating it in older individuals (Prince et al. 2008). The analysis of central measures when studying categorical variables or development phases, such as the mean age or confidence intervals, seems to lead us to an inevitable bias derived from the age distribution of the study sample (Smith 1991; Konigsberg 2015). To solve this problem, a large majority of researchers propose to calculate the age of onset for each phase instead of the average age, using transition analysis and logistic regression (Fojas et al. 2018). However, this type of approach does not seem to be widely accepted among specialists with more practical than scientific dedication (Buckberry 2015). Bayesian statistics appear to be the best solution to quantify the assumed error in terms of probability of success (Sironi et al. 2017). Numerous researchers have prioritised this perspective for the development of new methods, and have even developed specific computer tools to assist other specialists in the development and validation of predictive methods (Getz 2020). Finally, we cannot forget the dilemma between phase-based methods and component-based methods. In this case, component-based methods clearly minimise the subjectivity of the method, reduce the assumed error, and depend less on the experience of the observer (Shirley and Ramirez 2015; Fojas et al. 2018).

The contribution of our study lies in addressing the latter aspect. We believe the dilemma extends beyond simply dividing

a biological process into phases or components. It also encompasses the possibility of offering age-at-death estimates as discrete phases or as a continuous variable. Most proposals based on component analysis for age-at-death estimation through the pubic symphysis continue to offer estimates based on phases, with the consequent limitations in terms of procedural complexity and the inevitable loss of accuracy (Dudzik and Langley 2015; Gilbert and McKern 1973; Castillo et al. 2021). Nevertheless, according to our review, the present study is the first proposal to estimate age-at-death through the analysis of the classic components of the pubic symphysis while offering an estimated age as a continuous variable. This innovative approach has been shown to considerably improve the accuracy of the method.

The results obtained in this study are not only robust, but they also represent the best of the current state of the art. The use of symbolic regression has proven to be an effective choice, as it has led to improved precision and accuracy of age estimates, both considering classical and novel GP approaches. The meticulous validation process carried out in this work is a sufficient endorsement to propose it as a reliable alternative for the design of novel methods for age-at-death estimation in FA. Even so, the use of a transparent model structure is another additional value of our proposal. While artificial intelligence (AI) experts are very comfortable dealing with rule-based models, forensic anthropologists are more used to dealing with mathematical expressions (Cameriere et al. 2006; Irurita and Alemán 2017). As already stated in the area, EAI must consider the type of audience of the intelligent model to account for its transparency and interpretability (Barredo Arrieta et al. 2020). This is also a pillar of theory-guided data science, a very interesting paradigm to be integrated with EAI when expert knowledge and theories about the domain are available (Karpatne et al. 2017). The structure of the intelligent model should be chosen according to the kind of relations we intend to find in the specific domain and the prior knowledge can help in its design. Besides, the interpretable output of the model allows domain experts (in our case, forensic anthropologists) to reason and uncover new knowledge to be complemented with the existing one, as we have done in the current contribution.

In summary, we identify various benefits that make our proposal particularly compelling. First, the mathematical expression derived from the regression model has the advantage of not requiring all the bone traits identified in Todd's studies, which eases the real applicability of the model. As highlighted earlier, the proposed model is transparent and offers a solution easily understandable for the forensic anthropologist. Finally, the overall simplicity of our proposal can be effectively implemented in a real-world scenario due to its remarkable speed of execution and the ease of implementation of the final model. For example, advanced deep learning models tend to be computationally intensive, while a mathematical expression can be easily applied on edge devices. This feature opens up the possibility to create an app for its use on site, specially relevant in DVI scenarios. Future studies aimed at making the proposed method an easy-to-use tool, as well as checking its performance in other different populations, should be carried out to confirm its potential.

Despite the contributions of this study, certain limitations should be acknowledged. Age-at-death estimation inherently faces structural challenges due to the reliance on osteological collections, which are often not openly accessible for research, limiting the reproducibility of findings. Another limitation lies in the uncertainty introduced by human labeling, as the age estimates depend on subjective assessments by forensic experts. In addition, the model treats all variables at the same level, potentially overlooking hierarchical relationships or interactions that could more accurately reflect the underlying biological processes. Future study should focus on developing hierarchical models to capture the multi-level relationships among variables and incorporating methods to handle labeling uncertainty, such as probabilistic approaches or ensemble expert assessments, to enhance the robustness and reliability of age-at-death estimation. In this regard, we have been exploring the use of ensembles to integrate multiple expert opinions and to study the influence of biased human labeling in a recent work (Bermejo et al. 2024).

Author Contributions

S.D. and O.C. conceived the experiments, J.I. acquired and processed the materials, E.B. and A.D.V. conducted the experiments, E.B., J.I. and O.C. analysed and interpreted the results. All authors reviewed the manuscript.

Acknowledgements

This work was supported by grant CONFIA (PID2021-122916NB-I00) funded by MCIN/AEI/10.13039/501100011033, funded by 'ERDF A way of making Europe'. Additionally, E.B.'s work has been supported by the Regional Government of Andalusia as postdoctoral fellow (DOC_01130). A.D.V.'s work has been supported by the Regional Government of Andalusia under the Recovery, Transformation and Resilience Plan (GR/INV/0004/2022).

Ethics Statement

The institutional Ethics Committee of the University of Granada reviewed and approved the research study without issuing a reference number. The skeletal samples were anonymized according to the European Data Protection Directive. Consequently, the present study was not assigned a reference number following the committee's assessment. All the partners in the study obey national laws and European directives of privacy and data protection by ensuring that their employees and collaborators fulfil the appropriate procedures to handle sensitive data.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author on reasonable request.

References

Ali, S., T. Abuhmed, S. El-Sappagh, et al. 2023. "Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence." *Information Fusion* 99: 101805. https://doi.org/10.1016/j.inffus.2023.101805.

Aykroyd, R. G., D. Lucy, A. M. Pollard, and T. Solheim. 1997. "Technical Note: Regression Analysis in Adult Age Estimation." *American Journal of Physical Anthropology* 104, no. 2: 259–265. https://doi.org/10.1002/(SICI)1096-8644(199710)104:2<259::AID-AJPA11>3.0.CO;2-Z.

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion* 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Berg, G. 2008. "Pubic Bone Age Estimation in Adult Women." *Journal of Forensic Sciences* 53: 569–577. https://doi.org/10.1111/j.1556-4029.2008.00712.x.

Bermejo, E., O. Cordón, J. Irurita, I. Alemán, and Á. R. Salvador. 2024. "Age-At-Death Estimation Based on Symbolic Regression Ensemble Learning From Multiple Annotations." In 2024 IEEE Congress on Evolutionary Computation (CEC) (1–8). IEEE.

Brooks, S., and J. M. Suchey. 1990. "Skeletal Age Determination Based on the Os Pubis: A Comparison of the Acsádi-Nemeskéri and Suchey-Brooks Methods." *Human Evolution* 5, no. 3: 227–238. https://doi.org/10.1007/bf02437238.

Buckberry, J. 2015. "The (Mis)use of Adult Age Estimates in Osteology." *Annals of Human Biology* 42, no. 4: 323–331.

Cameriere, R., L. Ferrante, and M. Cingolani. 2006. "Age Estimation in Children by Measurement of Open Apices in Teeth." *International Journal of Legal Medicine* 120, no. 1: 49–52. https://doi.org/10.1007/s00414-005-0047-9.

Campanacho, V., A. T. Chamberlain, P. Nystrom, and E. Cunha. 2020. "Degenerative Variance on Age-Related Traits From Pelvic Bone Articulations and Its Implication for Age Estimation." *Anthropologischer Anzeiger* 77, no. 3: 259–268. https://doi.org/10.1127/anthranz/2020/1184.

Castillo, A., I. Galtés, S. Crespo, and X. Jordana. 2021. "Technical Note: Preliminary Insight Into a New Method for Age-At-Death Estimation From the Pubic Symphysis." *International Journal of Legal Medicine* 135, no. 3: 929–937. https://doi.org/10.1007/s00414-020-02434-6.

Christensen, A., N. Passalacqua, and E. Bartelink. 2019. *Forensic Anthropology: Current Methods and Practice*. Academic Press.

Corron, L., F. Marchal, S. Condemi, and P. Adalian. 2018. "A Critical Review of Sub-Adult Age Estimation in Biological Anthropology: Do Methods Comply With Published Recommendations?" *Forensic Science International* 288: 328.e1–328.e9. https://doi.org/10.1016/j.forsciint.2018.05.012.

Cunha, E., E. Baccino, L. Martrille, et al. 2009. "The Problem of Aging Human Remains and Living Individuals: A Review." *Forensic Science International* 193, no. 1: 1–13. https://doi.org/10.1016/j.forsciint.2009.09.008.

Demirjian, A., H. Goldstein, and J. Tanner. 1973. "A New System of Dental Age Assessment." *Human Biology* 5, no. 2: 211–217.

Smith, B. H. 1991. "Standards of Human Tooth Formation and Dental Age Assessment." In *Advances in Dental Anthropology*, edited by M. A. Kelley, C. S. Larsen, and K. Larsen, 143–168. Wiley-Liss, Inc.

Díaz-Rodríguez, N., J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma, and F. Herrera. 2023. "Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation." *Information Fusion* 99: 101896. https://doi.org/10.1016/j.inffus.2023.101896.

Díaz-Rodríguez, N., A. Lamas, J. Sanchez, et al. 2022. "EXplainable Neural-Symbolic Learning (X-NeSyL) Methodology to Fuse Deep Learning Representations With Expert Knowledge Graphs: The MonuMAI Cultural Heritage Use Case." *Information Fusion* 79: 58–83. https://doi.org/10.1016/j.inffus.2021.09.022.

Djurić, M., D. Djonić, S. Nikolić, D. Popović, and J. Marinković. 2007. "Evaluation of the Suchey–Brooks Method for Aging Skeletons in the Balkans." *Journal of Forensic Sciences* 52, no. 1: 21–23. https://doi.org/10.1111/j.1556-4029.2006.00333.x.

Dudzik, B., and N. Langley. 2015. "Estimating Age From the Pubic Symphysis: A New Component-Based System." *Forensic Science International* 257: 98–105. https://doi.org/10.1016/j.forsciint.2015.07.047.

Eshelman, L., and J. Schaffer. 1993. "Real-Coded Genetic Algorithms and Interval-Schemata." In *Foundations of Genetic Algorithms*, edited by L. D. Whitley, 2nd ed, 2, 187–202. Elsevier.

Fojas, C., J. Kim, J. Minsky-Rowland, and B. Algee-Hewitt. 2018. "Testing Inter-Observer Reliability of the Transition Analysis Aging Method on the William M. Bass Forensic Skeletal Collection." *American Journal of Physical Anthropology* 165: 183–193. https://doi.org/10.1002/ajpa.23342.

Gámez, J. C., D. García, A. González, and R. Pérez. 2016. "Ordinal Classification Based on the Sequential Covering Strategy." *International Journal of Approximate Reasoning* 76: 96–110.

Gámez, J. C., J. Irurita, A. González, S. Damas, I. Alemán, and O. Cordón. 2022. "Automating the Decision Making Process of Todd's Age Estimation Method From the Pubic Symphysis With Explainable Machine Learning." *Information Sciences* 612: 514–535. https://doi.org/10.1016/j.ins.2022.08.110.

Getz, S. M. 2020. "The Use of Transition Analysis in Skeletal Age Estimation." *WIREs Forensic Science* 2, no. 6: e1378. https://doi.org/10.1002/wfs2.1378.

Gilbert, B., and T. W. McKern. 1973. "A Method for Aging the Female Os Pubis." *American Journal of Physical Anthropology* 38, no. 1: 31–38.

Godde, K., and S. Hens. 2012. "Age-At-Death Estimation in an Italian Historical Sample: A Test of the Suchey-Brooks and Transition Analysis Methods." *American Journal of Physical Anthropology* 149: 259–265. https://doi.org/10.1002/ajpa.22126.

Hanihara, K., and T. Suzuki. 1978. "Estimation of Age From the Pubic Symphysis by Means of Multiple Regression Analysis." *American Journal of Physical Anthropology* 48, no. 2: 233–239. https://doi.org/10.1002/ajpa.1330480218.

He, H., and Y. Ma. 2013. *Imbalanced Learning: Foundations, Algorithms, and Applications*. John Wiley & Sons.

Howard, L. M., and D. J. D'Dangelo. 1995. "The GA-P: A Genetic Algorithm and Genetic Programming Hybrid." *IEEE Expert-Intelligent Systems and Their Applications* 10, no. 3: 11–15. https://doi.org/10.1109/64.393137.

Irurita, J., and I. Alemán. 2017. "Proposal of New Regression Formulae for the Estimation of Age in Infant Skeletal Remains From the Metric Study of the Pars Basilaris." *International Journal of Legal Medicine* 131, no. 3: 781–788. https://doi.org/10.1007/s00414-016-1478-1.

Karpatne, A., G. Atluri, J. H. Faghmous, et al. 2017. "Theory-Guided Data Science: A New Paradigm for Scientific Discovery From Data." *IEEE Transactions on Knowledge and Data Engineering* 29, no. 10: 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168.

Kimmerle, E., L. Konigsberg, R. L. Jantz, and J. P. Baraybar. 2008. "Analysis of Age-At-Death Estimation Through the Use of Pubic Symphyseal Data." *Journal of Forensic Sciences* 53, no. 3: 558–568.

Konigsberg, L., N. Herrmann, D. Wescott, and E. Kimmerle. 2008. "Estimation and Evidence in Forensic Anthropology: Age-At-Death." *Journal of Forensic Sciences* 53, no. 3: 541–557.

Konigsberg, L. W. 2015. "Multivariate Cumulative Probit for Age Estimation Using Ordinal Categorical Data." *Annals of Human Biology* 42, no. 4: 368–378.

Kotěrová, A., D. Navega, M. Štepanovský, Z. Buk, J. Brůžek, and E. Cunha. 2018. "Age Estimation of Adult Human Remains From Hip Bones Using Advanced Methods." *Forensic Science International* 287: 163–175. https://doi.org/10.1016/j.forsciint.2018.03.047.

Kotěrová, A., M. Štepanovský, Z. Buk, J. Brůžek, N. Techataweewan, and J. Velemínská. 2022. "The Computational Age-At-Death Estimation From 3D Surface Models of the Adult Pubic Symphysis Using Data

Mining Methods." *Scientific Reports* 12, no. 1: 10324. https://doi.org/10.1038/s41598-022-13983-8.

Koza, J. R. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press.

Kuhn, M., and K. Johnson. 2013. *Applied Predictive Modeling. Chapter Over-Fitting and Model Tuning*. Springer.

Langer, M., D. Oster, T. Speith, et al. 2021. "What Do We Want From Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research." *Artificial Intelligence* 296: 103473. https://doi.org/10.1016/j.artint.2021.103473.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature* 521: 436–444. https://doi.org/10.1038/nature14539.

Lesciotto, K. M. 2015. "The Impact of Daubert on the Admissibility of Forensic Anthropology Expert Testimony." *Journal of Forensic Sciences* 60, no. 3: 549–555. https://doi.org/10.1111/1556-4029.12740.

Liversidge, H. M., J. Buckberry, and N. Marquez-Grant. 2015. "Age Estimation." *Annals of Human Biology* 42, no. 4: 299–301. https://doi.org/10.3109/03014460.2015.1089627.

López, V., A. Fernández, S. García, V. Palade, and F. Herrera. 2013. "An Insight Into Classification With Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics." *Information Sciences* 250: 113–141. https://doi.org/10.1016/j.ins.2013.07.007.

McKern, T. 1976. *Sexual Dimorphism in the Maturation of the Human Public Symphysis*. Harvard University Press.

McKern, T., and T. Stewart. 1957. *Skeletal Age Changes in Young American Males: Analysed From the Standpoint of Age Identification*. Headquarters, Quartermaster Research & Development Command.

Mei, Y., Q. Chen, A. Lensen, B. Xue, and M. Zhang. 2022. "Explainable Artificial Intelligence by Genetic Programming: A Survey." *IEEE Transactions on Evolutionary Computation* 27, no. 3: 621–641. https://doi.org/10.1109/TEVC.2022.3225509.

Michalewicz, Z. 1996. *Genetic Algorithms + Data Structures = Evolutionary Programs*. Springer Berlin.

Milner, G. R., and J. L. Boldsen. 2012. "Transition Analysis: A Validation Study With Known-Age Modern American Skeletons." *American Journal of Physical Anthropology* 148, no. 1: 98–110. https://doi.org/10.1002/ajpa.22047.

Mosteller, F., and J. W. Tukey. 1968. *Data Analysis, Including Statistics*. Addison-Wesley.

Nguyen, Q. U., and T. H. Chu. 2020. "Semantic Approximation for Reducing Code Bloat in Genetic Programming." *Swarm and Evolutionary Computation* 58: 100729. https://doi.org/10.1016/j.swevo.2020.100729.

Pawlak, T. P., B. Wieloch, and K. Krawiec. 2015. "Semantic Backpropagation for Designing Search Operators in Genetic Programming." *IEEE Transactions on Evolutionary Computation* 19, no. 3: 326–340. https://doi.org/10.1109/TEVC.2014.2321259.

Prince, D., E. Kimmerle, and L. Konigsberg. 2008. "A Bayesian Approach to Estimate Skeletal Age-At-Death Utilizing Dental Wear." *Journal of Forensic Sciences* 53, no. 3: 588–593. https://doi.org/10.1111/j.1556-4029.2008.00714.x.

Rudin, C. 2019. "Stop Explaining Black Box ML Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1: 206–215.

Sánchez Ramos, L., and G. J. Corrales. 2000. "A Niching Scheme for Steady State GA-P and its Application to Fuzzy Rule Based Classifiers Induction." *Mathware & Soft Computing* 7, no. 2–3: 337–350.

Schanandore, J. V., M. Wolden, and N. Smart. 2022. "The Accuracy and Reliability of the Suchey–Brooks Pubic Symphysis Age Estimation Method: Systematic Review and Meta-Analysis." *Journal of Forensic Sciences* 67, no. 1: 56–67. https://doi.org/10.1111/1556-4029.14911.

Schmitt, A., P. Murail, E. Cunha, and D. Rougé. 2002. "Variability of the Pattern of Aging on the Human Skeleton: Evidence From Bone Indicators and Implications on Age at Death Estimation." *Journal of Forensic Sciences* 47, no. 6: 1203–1209. https://doi.org/10.1520/JFS15551J.

Shirley, N., and M. P. Ramirez. 2015. "Age Estimation in Forensic Anthropology: Quantification of Observer Error in Phase Versus Component- Based Methods." *Journal of Forensic Sciences* 60, no. 1: 107–111.

Sinha, A., and V. Gupta. 1995. "A Study on Estimation of Age From Pubic Symphysis." *Forensic Science International* 75, no. 1: 73–78. https://doi.org/10.1016/0379-0738(95)01772-B.

Sironi, E., J. Vuille, N. Morling, and F. Taroni. 2017. "On the Bayesian Approach to Forensic Age Estimation of Living Individuals." *Forensic Science International* 281: e24–e29. https://doi.org/10.1016/j.forsciint.2017.11.007.

Slice, D., and B. Algee-Hewitt. 2015. "Modeling Bone Surface Morphology: A Fully Quantitative Method for Age-At-Death Estimation Using the Pubic Symphysis." *Journal of Forensic Sciences* 60, no. 4: 835–843. https://doi.org/10.1111/1556-4029.12778.

Smith, J. E., and A. E. Eiben. 2010. *Introduction to Evolutionary Computing*. Springer Berlin.

Stoyanova, D., B. Algee-Hewitt, and D. Slice. 2015. "An Enhanced Computational Method for Age-At-Death Estimation Based on the Pubic Symphysis Using 3D Laser Scans and Thin Plate Splines." *American Journal of Physical Anthropology* 158, no. 3: 431–440. https://doi.org/10.1002/ajpa.22797.

Stoyanova, D. K., B. F. B. Algee-Hewitt, J. Kim, and D. E. Slice. 2017. "A Computational Framework for Age-At-Death Estimation From the Skeleton: Surface and Outline Analysis of 3D Laser Scans of the Adult Pubic Symphysis." *Journal of Forensic Sciences* 62, no. 6: 1434–1444. https://doi.org/10.1111/1556-4029.13439.

Todd, T. W. 1920. "Age Changes in the Pubic Bone." *American Journal of Physical Anthropology* 3, no. 3: 285–328.

Ubelaker, D. H. 2008. "Forensic Anthropology: Methodology and Diversity of Applications." In *Biological Anthropology of the Human Skeletonch*. John Wiley & Sons, Ltd.

Ubelaker, D. H., Q. R. Cordero, and N. F. Linton. 2020. "Recent Research in Forensic Anthropology." *European Journal of Anatomy* 24: 221–227.

Ubelaker, D. H., and H. Khosrowshahi. 2019. "Estimation of Age in Forensic Anthropology: Historical Perspective and Recent Methodological Advances." *Forensic Sciences Research* 4, no. 1: 1–9. https://doi.org/10.1080/20961790.2018.1549711.

Appendix A

Skeleton-Based Age Estimation Methods From the Pubic Symphysis

Brief Survey of Classical and Artificial Intelligence-Based Skeleton-Based Age Estimation Methods From the Pubic Symphysis

Currently, one of the most recommended methods to estimate the age of adults in forensic anthropology contexts (Cunha et al. 2009) is that introduced by Suchey and Brooks (Brooks and Suchey 1990) based on the analysis of morphological changes of the pubic symphysis. This method is an adaptation of Todd's pioneering proposal in 1920 (Todd 1920), who established 10 age-at-death phases, each with an age interval, to describe the degenerative changes of the pubis. Suchey and Brooks (Brooks and Suchey 1990) reduced the number of phases to 6 and modified the age intervals, now broader and overlap each other. However,

the way of applying the method remained practically the same as the one proposed by Todd: the subjective evaluation of the morphological characteristics of the pubic symphysis and the manual assignment of a developmental age-at-death phase by the forensic anthropologist.

Both Todd's and Suchey-Brooks' methods involve great difficulty and subjectivity in evaluating morphological traits and assigning a specific phase (Dudzik and Langley 2015). For example, it is common to observe morphological features corresponding to different age-at-death phases in the same pubic symphysis, so the forensic expert must decide, based on her/his experience and expertise, which pubic bone traits are more or less important to assign a specific phase. Other important limitations are the absence of a statistical analysis in Todd's method to obtain the results, or the excessive width of the age-at-death intervals proposed by Suchey and Brooks.

These kinds of methods are called *phase-based* as age estimation is treated as a classification problem, providing a phase (i.e. an age-at-death range) as final output. Smith (Smith 1991) argued that phase-based methods are highly dependent on the age distribution of the study sample used to create them, making them less reproducible. To solve this problem, he proposed to calculate the starting age of each phase using logistic regression, also called *transition analysis*. With this same idea, other authors introduced the use of Bayesian statistics combined with transition analysis as the best system to design age estimation methods (in this case based on teeth) (Prince et al. 2008). Since then, many different methods of this kind have been proposed (Berg 2008; Kimmerle et al. 2008; Konigsberg et al. 2008; Godde and Hens 2012; Milner and Boldsen 2012). Despite the fact that this Bayesian analysis is theoretically much more appropriate, recent studies suggest that the results in the estimations do not improve considerably with respect to traditional methods (Campanacho et al. 2020).

In 2022, a research team designed an explainable, rule-based age-at-death estimation method from the pubic symphysis, eliciting Todd's method knowledge and operation mode (Gámez et al. 2022). The method considered nine pubic bone traits extracted by forensic anthropologists from Todd's descriptions. The machine learning task to be solved was an an ordinal classification problem that aimed to assign one of the ten Todd's phases to each pubic bone. Both the problem statement and the considered data sample (the pubic symphysis collection of the University of Granada) are strongly imbalanced. The use of data oversampling techniques and an ordinal classification rule learning algorithm (NSLVOrd) (Gámez et al. 2016), which outperformed random forests and was competitive with a non-explainable deep neural network was thus considered. The learned model reported really good error rates (root mean square error, RMSE, and mean absolute error, MAE, of 12.34and 10.38 years, respectively) in a data sample that included 892 pubic bones. This is, to date, the best result of the state of the art on semi-automatic phase-based age-estimation methods. In addition, a human-readable description of the subjective Todd's method was obtained in the form of a list of 34 rules. Thanks to considering explainable machine learning and following a theory-guided data science approach, new knowledge was acquired in the form of a study on the importance of the pubic bone traits considered by Todd. That study concluded that certain traits are never used for age-at-death estimation, others are considered only for some (young or old subjects) phases, while others are kept along the whole age range.

As an alternative to phase-based methods, Gilbert and McKern (1973) were possibly the first to consider an independent analysis of each of the components (*component-scoring methods*). Their proposal was based on a simple regression system, which used a continuous variable for the age-at-death obtained by accumulating a value associated with each component (i.e. pubic symphysis trait) separately as represented in Figure 1a. This numeric value is assigned by the forensic anthropologist after visual inspection of the pubic bone and represents the possible states of each trait. Nevertheless, using this system they were unable to improve the accuracy of the age-at-death estimation method considerably, probably because of the use of a very rudimentary statistical analysis. Although Gilbert-McKern's method did not perform properly, it became the starting point of the increasing criticism raised by many authors over phase-based methods in the following decades. The argument was that the assumed error is lower when components are analysed separately than when the phases are analysed (Shirley and Ramirez 2015; Fojas et al. 2018). One of the most popular examples of methods using a similar approach is the Demirjian's dental age estimation method (Demirjian et al. 1973).

One of the first works that applied the component-scoring approach was developed by Hanihara and Suzuki (1978). The authors considered multiple LR since they used seven pubic symphysis traits and a constant to obtain a formula that is capable of estimating age-at-death. After the different experiments developed they managed to achieve a good accuracy although, as indicated, the age-at-death range considered was rather limited (18–38 years) and the number of samples was not very significant (135). Years later, in 1995, other researchers tested Hanihara-Suzuki component-based method over a sample of 41 male individuals in the 12–75 age range, comparing the results obtained with Todd's phase-based method (Sinha and Gupta 1995). The authors reported significantly better results for the former method and claimed again against the distribution of the phases proposed by Todd, as done by other authors Brooks and Suchey (1990). They observed that some of the variables tend to areas lying between two Todd phases. They also found that some of the pubic symphysis traits were irrelevant when estimating the age-at-death in many of the phases.

The most recent proposals for age-at-death estimation from the pubic symphysis are also based on component analysis, considering different regression and artificial intelligence techniques to develop new methods. One of the most relevant is a study developed by Slice and Algee-Hewitt (2015), in which the pubic bones were scanned to sample the development and degenerative process of the pubic symphysis and a LR model was used to estimate the numerical age-at death and some candidate age ranges. The data set used in this experiment consisted of 41 skeletons of Americans and the authors managed to obtain a RMSE of around 17.15 years when the intermediate point of the age range interval of the phases is considered. The same authors introduced an improvement of their method in Stoyanova et al. (2015) by considering a single feature automatically extracted from 3D models of the pubic bone, the flexion of a plane so that such a plane coincides with the surface of the bone. A LR model directly estimating a numerical age-at-death value from that single variable obtained an RMSE of around 19 years in a similar sample of 56 subjects in the range of 16–100 years. Finally, this study was extended in Stoyanova et al. (2017) including a new trait set integrating additional curvature-based features extracted from the 3D models. Based on a multi-variate regression, two models were proposed that reported an RMSE in the range 13.7–16.5 years over a sample of 93 caucasian male individuals between 16 and 90 years.

Dudzik and Langley (2015) proposed different models based on decision trees and multinomial logistic regression to design a component-scoring phase-based method considering only 5 pubic symphysis features and the first three of the six Suchey and Brooks' phases. For these experiments, they used a small sample of 47 subjects in the 18–40 year age-at-death range. They achieved very good results (94% success rate) even though they only considered the age range where the development processes can be more easily identified.

Kotěrová et al. (2018) published a study where the designed age-at-death estimation methods considered traits from two different bones: three from the pubic symphysis and four from the hip. Instead of reporting a phase, the age-at-death was directly estimated as a numerical value. Nine different models were used to aggregate the component scoring of the seven features. Based on a traditional (additive) scoring system, different kinds of regression methods were considered: K-nearest neighbours, regression trees, Bayesian models and artificial neural networks. In the experiments developed over a data set of 941 samples from subjects of different ethnics in the 19–100 years range, the best results were reported by the multinomial regression model with a RMSE of about 12.1 years and a MAE of 9.7 years. That model thus achieved the state-of-the-art results in semi-automatic age estimation from the pubic symphysis.

Castillo et al. (2021) proposed a new component-scoring, phase-based method designed using machine learning methods. With the aim of handling both interindividual variation and the subjectivity of visual scoring systems, their framework considered 16 different pubic symphysis traits, including a novel one which had not been previously used (microgrooves). All of them only showed two categorical values (present or absent) to ease the labeling task for the forensic anthropologist. A wrapper feature selection method was applied over each of the five age-at-death intervals (i.e. phases) analysed to select the best traits for it considering three different supervised machine learning algorithms (ZeroR, Naïve Bayes and Random Tree classifiers). The authors showed very promising results for some age-at-death intervals as well as discriminated between useful and useless traits.

Kotěrová et al. (2022) introduced two numerical age-at-death estimation methods from the pubic symphysis based on advanced computer vision and machine learning approaches. First, the authors proposed an explainable component-scoring method (SASS) based on multi-LR method that considered six features extracted directly from a 3D model of the pubic symphysis instead of relying on expert knowledge. SASS provided a RMSE of 14.3 years and a MAE of 11.7 years over a data set of 483 samples from subjects in the 18–92 years range. In addition, the authors introduced a method based on an ensemble of convolutional neural networks working over 41 different 2D projections of the 3D model (AANNESS). AANNESS carries the advantage of obtaining an estimation without explicitly requiring the identification of features or traits by the human expert at the cost of providing a non-explainable technique. It achieved a RMSE of 12.9 years and a MAE of 10.6 years. Hence, both methods are currently considered the state-of-the-art in terms of the accuracy of the predicted age-at-death from the pubic symphysis.

In the current study, we aim to design a new semi-automatic component-based age-at-death estimation method from Todd's proposed pubic symphysis traits. The method will rely on a mathematical expression combining the values of the different components, with these values provided by the forensic expert after visual inspection, into a numerical age-at-death value. Explainable machine learning methods (Mei et al. 2022) will be considered to derive such formulas from the data of the pubic collection of the Physical Anthropology Lab of the University of Granada. A theory-guided data science approach (Karpatne et al. 2017) will be followed combining the expertise of the forensic anthropologist and the explainable descriptions derived from the data.

Gilbert-McKern's Component-Scoring Age-at-Death Estimation Method From the Pubic Symphysis

As mentioned, the inherent limitations of Todd's method have resulted in the proposal of many different adaptations and alternative methods in the last century. In particular, McKern and Stewart (1957) and Gilbert and McKern (1973) made two fundamental changes. They assigned numerical values to each pubic bone trait and they also aggregated those values to directly estimate the numerical age-at-death. The former proposal was validated on male pubic symphysis. Then, it was adapted to female pubic bones in the latter study. To develop their methods, they considered three different components: (I) the dorsal demi-face; (II) the ventral rampart and (III) the sympheyseal rim of the pubic symphysis. A value in $\{0, \ldots, 5\}$ was assigned to each of these three components, corresponding to six possible development stages. The addition of those three values resulted in a single number in $\{0, \ldots, 15\}$ which is directly associated with an age range according to the analysed sample (see Figure 1b).

McKern summarised all the previous work in 1976 (McKern 1976) by analysing the problem, the proposed approach, the differences to take into account if the samples correspond to male or female individuals, and reporting a comparison with Todd's method. Without contradicting the work by Todd, McKern concluded that his age estimation method from the pubic symphysis outperformed Todd's pioneering method.

Appendix B

Machine Learning Model Visualisation

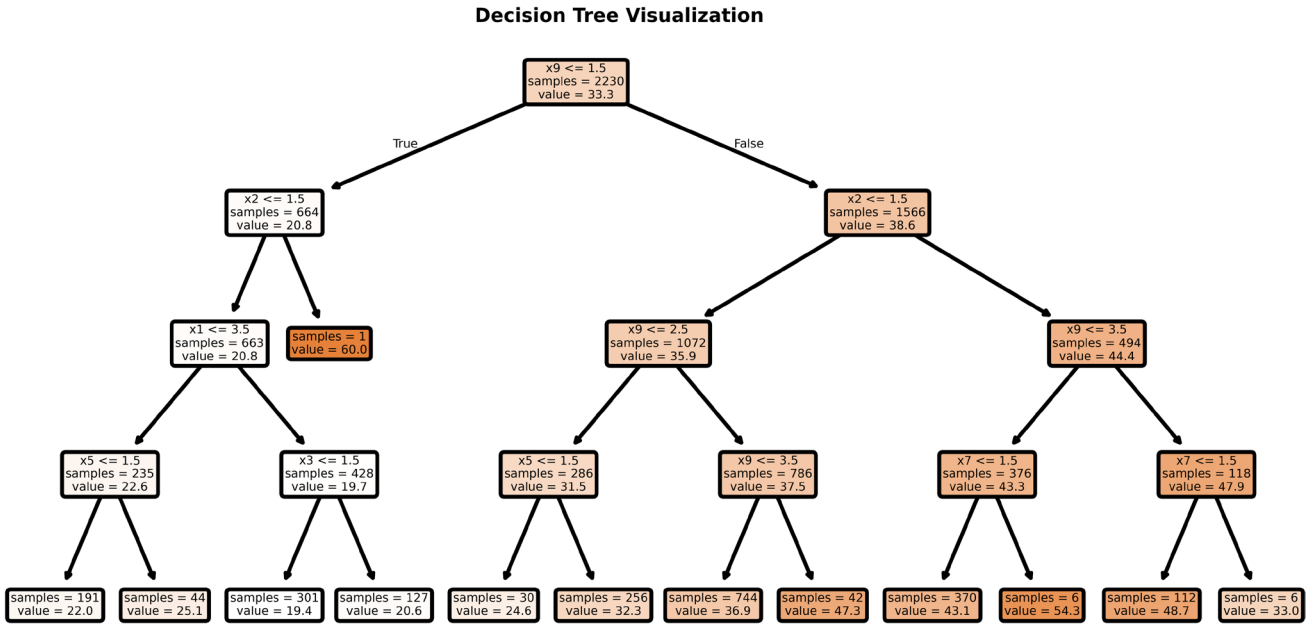

FIGURE B1 | Visualisation of the best decision tree model.

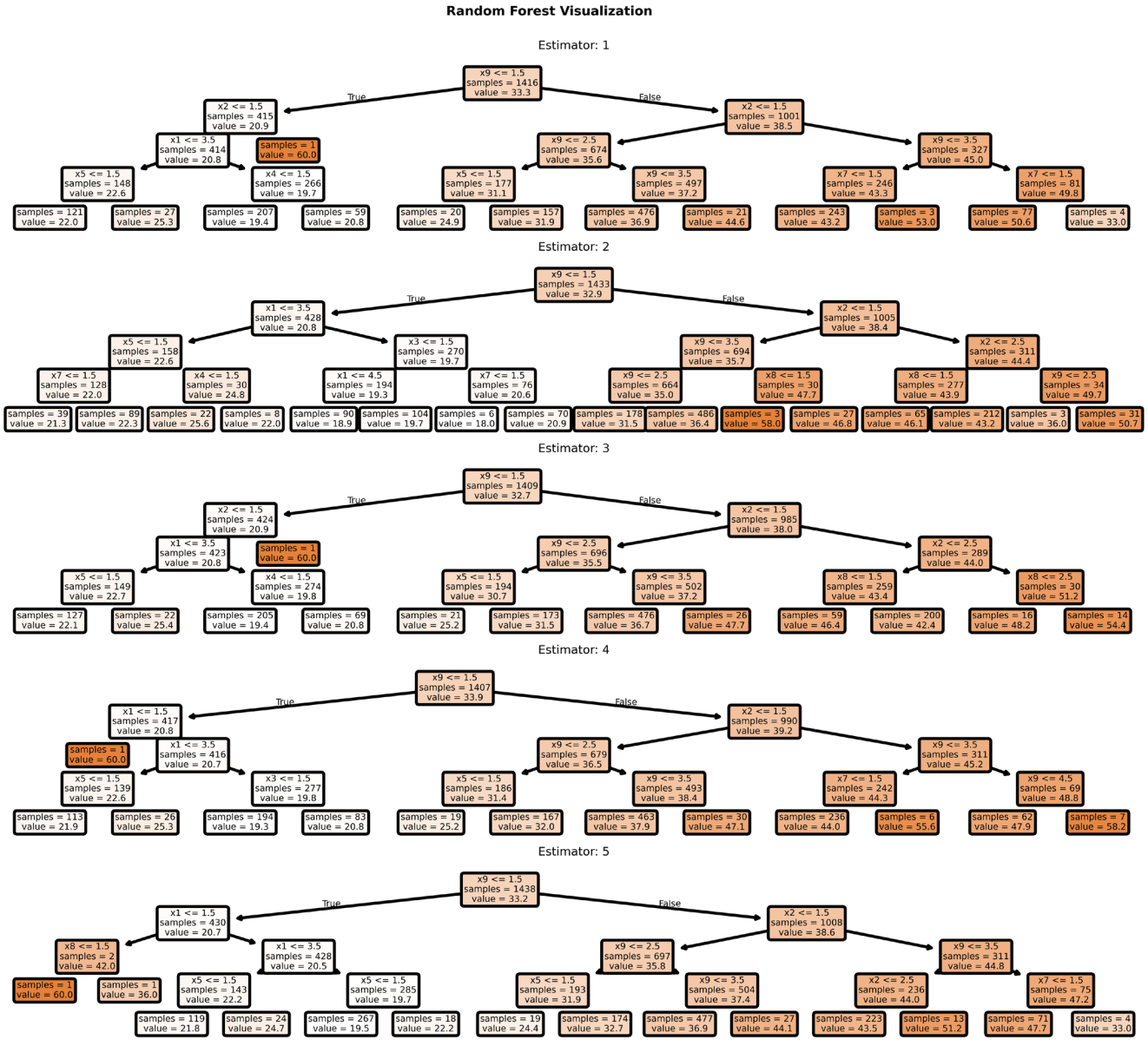

FIGURE B2 | Visualisation of random forest estimators for the best model found. Note that this visualisation includes only four of 100 weak learners.