Forensic Science International 380 (2026) 112820

Contents lists available at ScienceDirect

Forensic Science International

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

Case-specific accuracy in sex estimation from long bones in forensic anthropology: An "accuracy x-factors" approach

Siam Knecht [a,*], Gabriele Krüger [b], Leandi Liebenberg [b], Yann Ardagna [a], Marie Perrin [a], Mustapha Ouladsine [c], Christophe Roman [c], Pascal Adalian [a]

[a] Aix Marseille Univ, CNRS, EFS, ADES, Marseille, France
[b] University of Pretoria, Pretoria, South Africa
[c] LIS UMR 7020 CNRS, Aix Marseille Université, Marseille, France

ARTICLE INFO

Keywords:
Forensic anthropology
Long bones
Sex estimation
Machine learning
Reliability
Accuracy per-individual

ABSTRACT

Background: Sex estimation from human skeletal remains is a cornerstone of forensic anthropological analysis. Long bones, despite exhibiting less pronounced dimorphism than pelvis, serve as invaluable substitutes. However, traditional statistical approaches for sex estimation from long bone measurements often lack the precision and case-specific reliability demanded by stringent legal standards. This study addresses these critical limitations by rigorously exploring the potential of machine learning (ML) to significantly enhance sex estimation from long bones.

Methods: We analyzed 16 osteometric measurements from the humerus, radius, femur, and tibia of 2969 individuals (1207 females, 1762 males) across eight skeletal collections. Eleven ML algorithms were trained and cross-validated, then validated on an independent South African sample. To address the common issue of incomplete remains, we developed an "accuracy x-factors" approach. This method simulates missing data scenarios and selects tailored training subsets, yielding individualized reliability assessments adapted to specific measurement availability.

Results: Linear Discriminant Analysis (LDA) consistently achieved the highest performance, with accuracies up to 93 %. The "accuracy x-factors" approach proved effective in providing per-individual confidence measures, highlighting that prediction reliability varies with data completeness. Adjusting thresholds to higher confidence levels (e.g., >0.7) substantially reduced error rates, allowing a conservative yet legally robust classification of a smaller but more reliable subset of cases.

Conclusion: ML offers a powerful framework for sex estimation from long bones. The proposed "accuracy x-factors" approach introduces a significant methodological advance by delivering transparent, case-specific confidence levels. This strengthens both the forensic applicability and the legal admissibility of long bone-based sex estimation.

1. Introduction

Forensic anthropology currently navigates an increasingly complex landscape, characterized by evolving societal demands and progressively rigorous legal frameworks. The discipline is tasked with providing robust individual identifications in diverse contexts, ranging from mass disasters and migration crises to armed conflicts [1]. Concurrently, forensic practitioners must adhere to stringent legal admissibility standards for expert testimony, most notably the Daubert criteria, which govern the admissibility of scientific evidence in court [2]. These converging pressures underscore the need for developing reliable and transparent methodologies for biological profile estimation from skeletal remains, with sex estimation serving as a foundational component [3].

Sex estimation is indeed a pivotal step in forensic anthropological analysis [4–7], as it effectively halves the pool of potential identities and provides crucial guidance for subsequent estimations of age, stature, and population affinity. While DNA analysis remains the gold standard for

* Correspondence to: Laboratoire ADES – Anthropologie bio-culturelle, Droit, Éthique et Santé, Aix-Marseille Université, 27 bd Jean Moulin, Marseille 13385, France.
E-mail address: siam.knecht@univ-amu.fr (S. Knecht).

https://doi.org/10.1016/j.forsciint.2026.112820
Received 21 August 2025; Received in revised form 12 November 2025; Accepted 8 January 2026
Available online 10 January 2026

sex determination, its applicability is often limited by the degradation or ancient nature of remains [8], rendering skeletal assessment an indispensable tool. Traditionally, the pelvis has been recognized as the most reliable skeletal element for sex estimation due to its pronounced sexual dimorphism [9]. However, in numerous forensic cases, the pelvis is either incomplete or entirely absent, necessitating recourse to alternative skeletal elements.

Among these alternatives, long bones have long been acknowledged as valuable indicators of sex, exhibiting a significant degree of sexual dimorphism second only to the pelvis. Previous research has extensively explored sex estimation using individual long bones [e.g., [10–24] or their combinations [e.g., [25–28], predominantly employing traditional statistical methods. Yet, these classical approaches often struggle to fully capture the intricate complexity of skeletal variation. Crucially, they may not meet the high standards of reliability and quantifiable error rates demanded by the Daubert criteria for admissibility in legal proceedings. This deficiency highlights a critical gap in current methodologies, particularly when dealing with the inherent variability and incompleteness of forensic evidence.

In light of these limitations, Artificial Intelligence (AI) and machine learning (ML) algorithms present a highly promising avenue for advancing forensic anthropological practices [29]. These sophisticated computational techniques excel at detecting subtle patterns within complex datasets, offering the potential to significantly enhance both the accuracy and reliability of sex estimation from long bone measurements. Furthermore, a key advantage of AI and ML methods is their inherent ability to provide objective measures of confidence in their predictions.

In this study, we adopt a broad and operational definition of “machine learning” that includes both classical statistical classifiers (e.g., Linear and Quadratic Discriminant Analysis, Logistic Regression) and more recent algorithmic approaches (e.g., Random Forests, Gradient Boosting, Support Vector Machines). Although methods such as Linear and Quadratic Discriminant Analysis originate from traditional statistics, they are here considered part of the machine learning framework because they share the same supervised learning paradigm, training models from labeled data to predict unseen cases. This inclusive perspective aligns with current computational practice in data science and forensic applications, as reflected in widely used libraries such as *scikit-learn* in python.

A critical, yet often overlooked, consideration in the application of AI to forensic anthropology is the imperative for case-specific reliability assessments. While overall model accuracy provides valuable general insights, the unique nature of each forensic case demands a more nuanced approach. It is essential to determine the accuracy specific to each individual case, considering the unique combination of available skeletal measurements. This individual-level accuracy assessment becomes particularly crucial when confronted with incomplete remains or missing data, scenarios that are regrettably common in forensic contexts.

Moreover, the establishment of appropriate prediction thresholds for each case is paramount not only to minimize errors [30] but also to ensure compliance with legal standards. By defining a confidence threshold below which predictions are deemed unreliable, practitioners are empowered to make more informed decisions regarding the application or withholding of sex estimations based on the available evidence. This nuanced approach not only bolsters the overall reliability of sex estimation but also provides a clear, quantifiable measure of the potential error rate.

In response to these critical considerations, this study aims to validate and significantly extend previous research on sex estimation from long bones by leveraging advanced AI techniques. We employ a range of machine learning classifiers to develop robust predictive models based on combined long bone measurements. Our comprehensive approach is meticulously designed to address several key considerations: reliability, generalizability across diverse populations, transparency of results, and practical utility in real-world forensic scenarios. By harnessing the power of AI while rigorously adhering to scientific and legal standards, we endeavor to develop a reliable and versatile tool for sex estimation in forensic anthropology. This research holds substantial potential to significantly enhance the field's capacity to provide accurate, legally admissible biological profiles in both ongoing forensic investigations and archaeological contexts. Ultimately, our overarching goal is to equip forensic anthropologists with advanced, transparent tools that not only effectively meet the complex challenges of modern casework but also rigorously satisfy the stringent requirements of the legal system.

2. Materials and methods

2.1. Data

This study examined four long bones (humerus, radius, femur and tibia) from a total sample of 2969 individuals, comprising 1207 females and 1762 males. The data was compiled in a common database used for model training, incorporating eight distinct collections while adhering to ethical standards. An independent test sample from Pretoria (South Africa) was utilized for model validation (Table 1).

Among these 7 collections composing our model training sample, three originate from French archaeological sites: Laudun, La Ciotat (LC-HOP) and Marseille Cimetière des Crottes (MPC). These collections are housed at the UMR ADES osteological library in Marseille and the "Ostéothèque DRAC PACA." The Goldman dataset, accessible online through Dr. Benjamin Auerbach, encompasses postcranial measurements from various geographical locations and historical periods. The Robert J. Terry Anatomical Collection, housed at the Smithsonian Institute's National Museum of Natural History in Washington, D.C., was also included. Detailed information on its history and composition can be found in the publication by Hunt and Albanese (2005) [31]. Nice collection includes 40 individuals who donated their bodies to the medical school, in accordance with French legislation permitting donations for teaching and research purposes. The Olivier collection is housed at the National Museum of Natural History (MNHN) in Paris and serves as a pivotal anthropological reference collection. A notable feature of this collection is that the majority of the specimens are accompanied by detailed biological profiles typically encompassing essential data such as the individual's sex, age at the time of death, and stature. The Pretoria Bone Collection (PBC) validation sample was a stratified random set of 360 individuals, comprising equal representation across sex and ancestry groups [32]. The PBC was established in 1942 with the foundation of the medical school and reorganized in 2000 into a research-oriented resource with comprehensive demographic documentation. All remains derive from whole-body donations or unclaimed bodies, in accordance with the South African National Health Act, and include metadata such as sex, age, cause of death, and population affinity. For the present study, 360 postcranial skeletons were selected to ensure equal representation across sex and ancestry, including equal numbers of Black and White South African males and females.

The methods used to determine sex in each collection are specified here. In the Goldman dataset, sex was estimated using the morphological methods of Bruzek (2002) [33] and Phenice (1969) [34]. In the French archaeological collections (Laudun, LC-HOP, and MPC), sex estimation followed the morphological approach of Bruzek (2002) [33]. For the Nice, Olivier, Terry and PBC collections, biological sex was documented. While some training data relied on estimated-sex collections, all model validation and performance evaluations were conducted exclusively on known-sex individuals from the Pretoria Bone Collection, ensuring independent assessment.

2.2. Measurements

A total of 16 long bones measurements were selected for this study

Table 1
Collections selected for this study.

	Collection	Location	Time period	Sex	Females	Males	Total
	Goldman	Europe	Antiquity to the present day	Estimated	543	985	1528
	Laudun	France	5th – 13th century	Estimated	54	113	167
	LC-HOP	France	16th-19th	Estimated	105	109	214
Training sample	MPC	France	18th – 20th century	Estimated	57	47	104
	Nice	France	21st century	Known	21	19	40
	Terry	USA	20th century	Known	207	247	454
	Olivier	France	20th century	Known	40	62	102
Test sample	PBC	South-Africa	21st century	Known	180	180	360
Total	Common database		Antiquity to the present day	Estimated and known	1207	1762	2969

Table 2
Long bones measurements selected for this study.

Anatomical localization	Bone	Abbreviation	Measurement
Upper limb	Humerus	humxln	Maximum length
		humebr	Distal epicondylar breadth
		humhhd	Maximum vertical diameter of the head
		hummxd	Maximum diameter midshaft
		hummwd	Minimum diameter midshaft
	Radius	radxln	Maximum length
		radtvd	Transverse diameter at midshaft
		radapd	Antero-posterior diameter at midshaft
Lower limb	Femur	femxln	Maximum length
		fembln	Physiological length
		femebr	Epicondylar breadth
		femhhd	Maximum diameter of the head
		femmtv	Medio-lateral diameter of the midshaft
		femmap	Antero-posterior diameter of the midshaft
	Tibia	tibxln	Lateral condyle - medial malleolus length
		tibpeb	Maximum proximal epiphyseal breadth

(Table 2). These measurements followed Martin & Saller (1957) [35] protocols and included maximum lengths, measurements taken at the proximal and distal epiphyses, and at the midshaft of long bones, using sliding calipers or osteometric boards. Data were collected on the left side of individuals, and if absent, the right side was utilized.

2.3. Classification

In this study, we employed both traditional statistical techniques and modern machine learning algorithms for classification. Classical discriminant analysis represents a parametric statistical method that assumes specific data distributions and uses analytical decision boundaries derived from estimated parameters. In contrast, more recent Machine Learning approaches (such as Random Forests or Support Vector Machines) are non-parametric and rely on computational learning from data patterns without requiring explicit model assumptions.

Throughout the manuscript, we therefore use the term “Machine Learning” to encompass both classical methods like LDA and contemporary algorithmic approaches, emphasizing their shared data-driven, predictive nature while distinguishing them from purely descriptive statistical methods.

Eleven classification algorithms provided by the scikit-learn library for Python [36] were utilized: Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gaussian Process (GP), Gradient Boosting (GB), Random Forest (RF), Extra Trees (ET), Gaussian Naive Bayes (GNB), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA). Extreme Gradient Boosting (XGB) was implemented using the external *xgboost* library. The models were constructed using the training sample and the set of 16 variables, with cross-validation (5 folds) and hyperparameters optimization conducted through GridSearch.

2.4. Posterior probability

In the field of classification, determining the optimal decision threshold is a complex subject that extends beyond the simple use of a default value of 0.5 for the *a posteriori* probability. Recent research highlights the nuances and challenges associated with this approach [30,37]. In-depth analysis reveals that the uniform application of an overall accuracy rate to each individual in a sample is an oversimplification. In reality, there is a significant overlap between the distributions of male and female characteristics for various bone measures and discriminant scores [38,39]. This overlap creates a ‘zone of uncertainty’ where distinguishing between the sexes becomes particularly difficult. The scientific literature highlights the limitations of a binary approach, which treats very different *a posteriori* probability (such as 0.51 and 0.99) as equivalent, while differentiating between very similar probabilities (such as 0.49 and 0.51) [3]. This method can lead to misleading interpretations, particularly in fields such as forensic anthropology where reliability is crucial. To overcome these limitations, researchers have proposed the use of multiple thresholds (e.g. 0.5, 0.7, 0.8 and 0.95) to assess the reliability of classifications [4]. This approach allows for a more nuanced analysis, considering different levels of certainty and providing a better understanding of the proportion of individuals misclassified in each category. These considerations underscore the importance of a more sophisticated approach to the interpretation of *a posteriori* probabilities in classification, particularly in high-stakes accuracy domains.

2.5. Accuracy x-factors

In forensic anthropology, the pervasive challenge of incomplete skeletal remains necessitates a nuanced approach to sex estimation. To address this, we have developed an innovative method for calculating accuracy on a per-individual basis, meticulously considering the unique combination of available and missing data for each specimen. Our approach involves creating data subsets by starting with the complete data and systematically introducing missing data patterns that precisely reflect the actual patterns of available measurements for each individual in the test sample. We then validate our machine learning models on these tailored subsets, thereby obtaining a specific accuracy score for each unique combination of measures. This personalized accuracy score is subsequently associated with the sex estimation prediction for that individual, serving as a direct and personalized confidence measure.

This method offers significant advantages in forensic contexts. It provides a more realistic assessment of prediction reliability by directly reflecting the available data for each specimen. Furthermore, it allows for flexible application across various forensic scenarios where the completeness of skeletal remains may vary considerably. Importantly, this approach offering a transparent, case-specific measure of potential

error rates. By implementing this x-factors accuracy calculation, we aim to enhance both the practical utility and legal admissibility of our sex estimation method. This approach not only improves the overall reliability of our predictions but also equips forensic practitioners with crucial, quantifiable information about the confidence level of each individual estimation. Ultimately, this method represents a substantial step forward in addressing the complex challenges of sex estimation in forensic anthropology, particularly in cases involving incomplete or fragmentary remains.

3. Results

3.1. Overall accuracy

Table 3 summarizes the overall accuracy of various classification models under different training and test data configurations. The accuracies range from 0.85 to 0.93, with Linear Discriminant Analysis (LDA) consistently delivering the best results. It achieves a maximum accuracy of 0.93 and remains the top-performing model across all combinations.

The most favorable results were observed when the training data included imputed values for missing cases, and the test data consisted solely of complete individuals (Configuration B). However, when test data includes imputed values for incomplete individuals (Configurations A and C), a noticeable decline in accuracy occurred, highlighting the potential negative impact of imputing missing data on model performance.

Imputing missing values in the training data proved beneficial for improving model reliability, as it mitigates biases caused by entirely excluding incomplete cases. Moreover, for the 85 incomplete individuals in the South African sample, analyzing their specific impact on model accuracy could provide additional insights into the robustness of the models under real-world conditions.

While overall accuracy offers valuable insights into general model performance, sensitivity (true positive rate) and specificity (true negative rate) provide complementary information critical for forensic contexts. Tables 4 and 5 show that LDA not only achieves high overall accuracy but also maintains a good balance between sensitivity (0.91) and specificity (0.91–0.95), indicating that it reliably identifies both male and female individuals across configurations. Other models show variable trade-offs between sensitivity and specificity, highlighting the importance of considering these metrics alongside accuracy, especially in high-stakes individual assessments where errors can have significant consequences.

Table 3
Accuracy of classification models under different training and test data configurations.

Models	Accuracy with PP = 0.5			
	A Train = 2609 Test = 360	**B** Train = 2609 Test = 275	**C** Train = 1678 Test = 360	**D** Train = 1678 Test = 275
RF	0.90	0.91	0.88	0.88
GBC	0.90	0.90	0.87	0.87
SVM	0.91	0.93	0.89	0.91
LR	0.91	0.93	0.89	0.91
DT	0.86	0.86	0.87	0.87
GPC	0.89	0.91	0.87	0.89
ADA	0.89	0.90	0.88	0.89
ET	0.89	0.91	0.87	0.88
GNB	0.86	0.87	0.85	0.86
LDA	0.91	0.93	0.89	0.92
XGB	0.89	0.91	0.86	0.87
QDA	0.89	0.91	0.88	0.89

A: Training data = all individuals – Test data = all individuals
B: Training data = all individuals – Test data = complete individuals only
C: Training data = complete individuals only – Test data = all individuals
D: Training data = complete individuals only – Test data = complete individuals only

Table 4
Sensitivity of classification models under different training and test data configurations.

Models	Sensitivity with PP = 0.5			
	A Train = 2609 Test = 360	**B** Train = 2609 Test = 275	**C** Train = 1678 Test = 360	**D** Train = 1678 Test = 275
RF	0.88	0.87	0.87	0.86
GBC	0.86	0.85	0.85	0.83
SVM	0.91	0.92	0.88	0.87
LR	0.91	0.92	0.91	0.92
DT	0.84	0.82	0.87	0.86
GPC	0.87	0.86	0.89	0.89
ADA	0.88	0.88	0.91	0.91
ET	0.90	0.89	0.89	0.89
GNB	0.94	0.95	0.96	0.96
LDA	0.91	0.91	0.91	0.91
XGB	0.86	0.85	0.82	0.80
QDA	0.91	0.91	0.89	0.88

A: Training data = all individuals – Test data = all individuals
B: Training data = all individuals – Test data = complete individuals only
C: Training data = complete individuals only – Test data = all individuals
D: Training data = complete individuals only – Test data = complete individuals only

Table 5
Specificity of classification models under different training and test data configurations.

Models	Specificitiy with PP = 0.5			
	A Train = 2609 Test = 360	**B** Train = 2609 Test = 275	**C** Train = 1678 Test = 360	**D** Train = 1678 Test = 275
RF	0.91	0.94	0.88	0.90
GBC	0.93	0.95	0.89	0.90
SVM	0.91	0.95	0.90	0.94
LR	0.90	0.94	0.86	0.90
DT	0.87	0.89	0.87	0.87
GPC	0.91	0.94	0.85	0.89
ADA	0.89	0.92	0.86	0.88
ET	0.88	0.92	0.84	0.88
GNB	0.77	0.79	0.74	0.77
LDA	0.91	0.95	0.88	0.92
XGB	0.93	0.96	0.91	0.93
QDA	0.88	0.91	0.88	0.91

A: Training data = all individuals – Test data = all individuals
B: Training data = all individuals – Test data = complete individuals only
C: Training data = complete individuals only – Test data = all individuals
D: Training data = complete individuals only – Test data = complete individuals only

The observed variability across overall accuracy, sensitivity, and specificity underscores the need for a nuanced, case-specific evaluation of model performance, particularly when dealing with missing or imputed data, as is common in forensic anthropology.

3.2. Accuracy 16 factors – complete individuals

To provide a more nuanced evaluation of model performance, we analyzed accuracy as a function of the prediction threshold for complete individuals, ranging from 0.5 to 0.99 (Fig. 1). The results indicate that LDA accuracy consistently increases as the confidence threshold rises, reflecting a critical trade-off between the percentage of the population retained for classification and the certainty of predictions.

To illustrate this relationship, the predictive model was applied to six randomly selected individuals, with results presented in Table 6. This analysis demonstrates that two individuals were incorrectly classified, both with confidence thresholds below 0.7. This finding critically

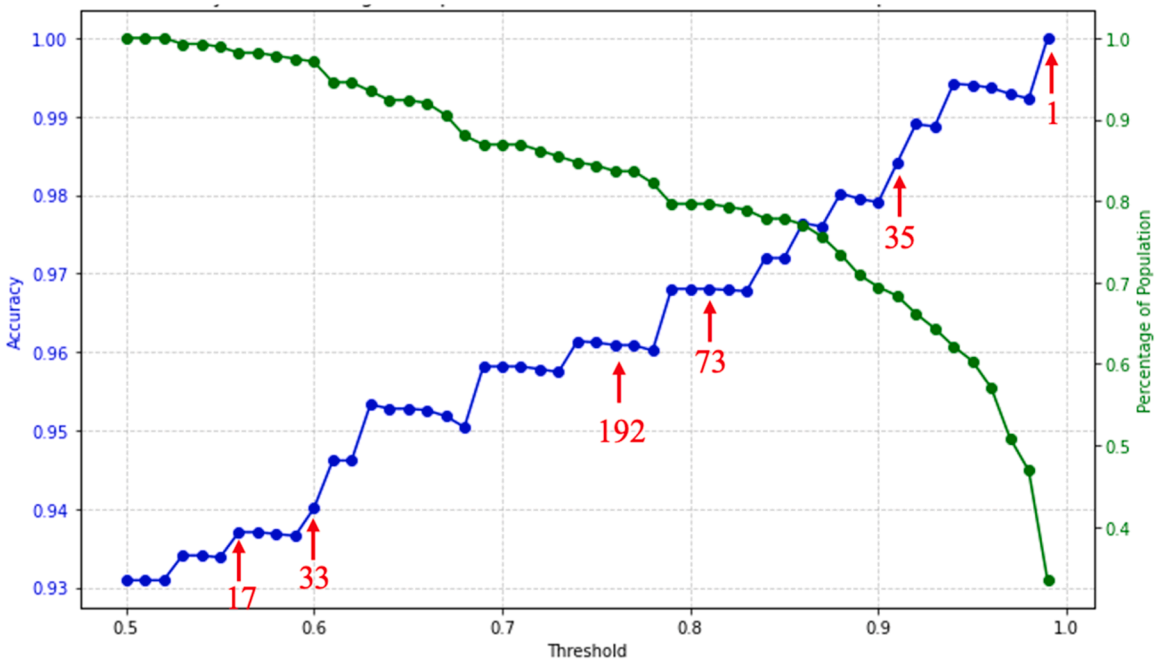

Fig. 1. Accuracy and Percentage of population as a function of threshold with LDA when test sample are complete individuals (configuration B) (⟶ Individual).

Table 6
Accuracy 16 factors in six random complete individuals.

Random case	x-factors	Missing values	Real Sex	Predicted Sex	Level of confidence	Accuracy	Sensitivity	Specificity	Percentage of population
1	16	-	Male	Male	0.99	1	1	1	33 %
35			Female	Female	0.91	0.98	0.98	0.99	68 %
73			Female	Female	0.81	0.97	0.95	0.98	80 %
31			Female	Male	0.60	0.94	0.92	0.96	97 %
17			Male	Female	0.56	0.98	0.92	0.95	98 %
189			Female	Female	0.77	0.96	0.95	0.98	99 %

highlights the importance of considering both overall model accuracy and the percentage of individuals retained at varying confidence thresholds, particularly in forensic contexts where false positives or negatives carry significant weight.

3.3. Accuracy x-factors – incomplete individuals

This analysis was further extended to cases with imputed missing values. Among the 85 incomplete individuals in our South African test sample, we identified 39 unique combinations of missing variables. To rigorously assess the impact of these missing data patterns, we artificially introduced these 39 combinations of missing data into our 275 complete individuals from the test sample. This approach allowed us to simulate a wide range of realistic forensic scenarios while maintaining a consistent baseline for comparative analysis. It is important to note that our models were initially trained on the complete dataset, resulting in a single model for each classification method. The observed variation in accuracy, therefore, stems from the validation process, where different combinations of missing data were applied to a population. This methodology enables us to comprehensively evaluate how each model performs across various patterns of missing data, providing a more comprehensive and realistic assessment of their reliability in forensic applications. By doing so, we gain a deeper understanding of the limitations and strengths of each model when confronted with incomplete skeletal remains, a pervasive challenge in forensic anthropology.

To illustrate this detailed analysis, we randomly selected six individuals with $x + 1$ missing variables, representing a common missing data pattern. Fig. 2 demonstrates that across all combinations, accuracy tends to improve with higher confidence thresholds, albeit with a reduction in the population percentage retained. We observe that depending on the combination of missing imputed values, the starting accuracy (threshold set at 0.5) is variably impacted; for example, combination 4 shows an initial accuracy of 86 %. Table 7 provides a detailed breakdown for these six randomly selected individuals. These results clearly demonstrate that the confidence threshold required for reliable classification varies significantly depending on the combination of missing values. Additionally, as illustrated in Fig. 2, this variability directly impacts the accuracy x-factors assigned to each prediction, underscoring the individualized nature of reliability.

We utilized these six example cases to illustrate the calculation of these accuracy x-factors, but the underlying principle dictates that this process should be applied to every individual, particularly each of the 85 incomplete cases. This comprehensive approach ensures that the sex predicted for each individual is rigorously supported by a reliable performance measure, precisely reflecting the exact availability of data and its direct impact on model performance for that particular case.

4. Discussion

Our study provides a comprehensive analysis of sex estimation using long bone measurements and various machine learning algorithms, with a particular focus on addressing the pervasive challenges of incomplete skeletal remains and the critical need for case-specific accuracy assessments in forensic anthropology. The findings have profound implications for both fundamental research and practical applications within the field.

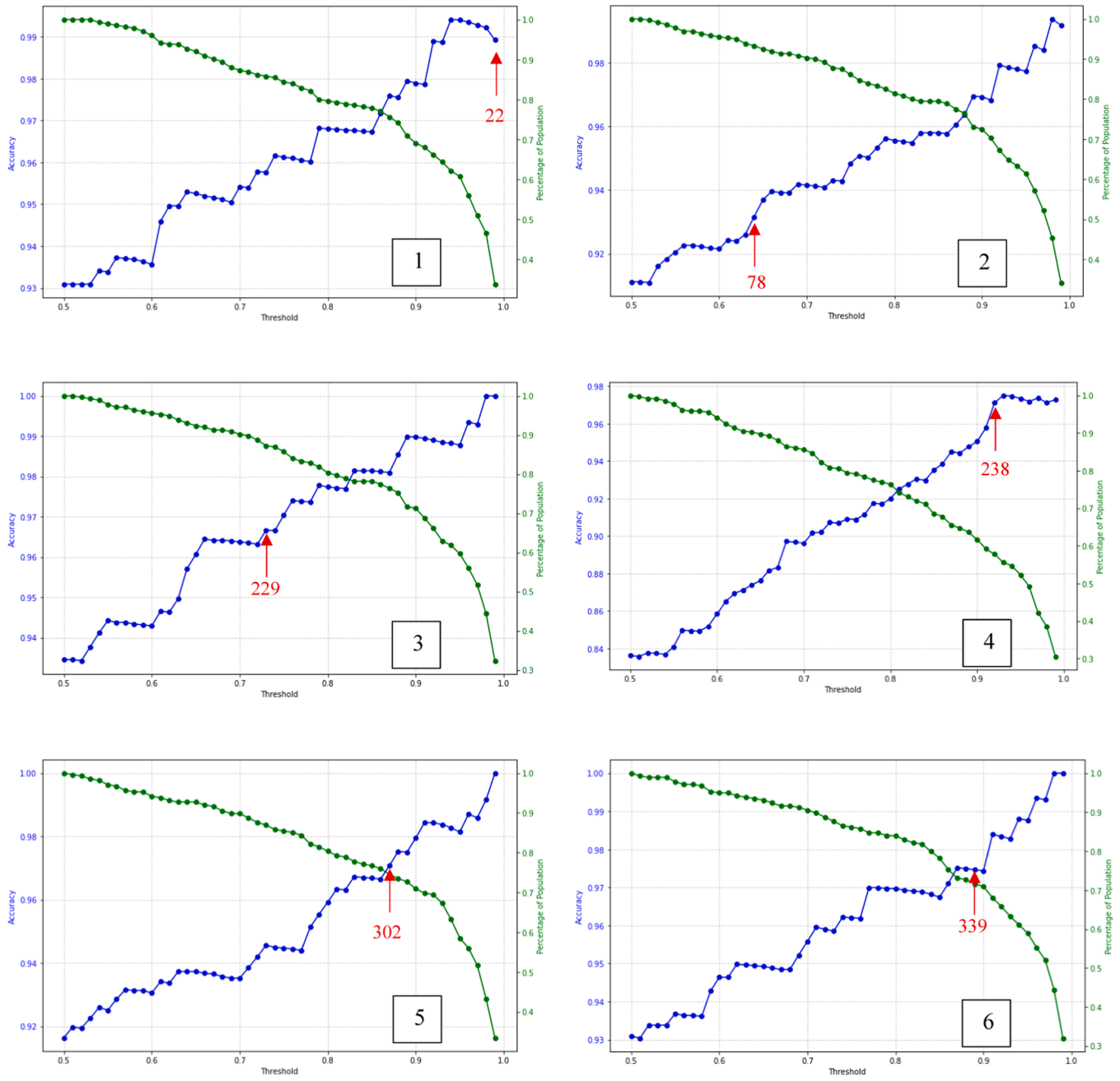

Fig. 2. Accuracy and Percentage of population as a function of threshold with LDA in six *random incomplete combinations (⟶ Individual). Missing values imputed:* **1:** hummxd. **2:** femebr, tibpeb. **3:** femebr, tibxln, tibpeb. **5:** radxln, radtvd, radapd, femmtv. **5:** humxln, humebr, humhhd, hummxd, hummwd. **6:** femxln, fembln, femebr, femhhd, femmtv, femmap.

Table 7
Accuracy x- factors in six random incomplete individuals.

Random Case	x-factors	Missing values	Real Sex	Predicted Sex	Level of confidence	Accuracy	Sensitivity	Specificity	Percent-age of population
22	1	hummxd	Male	Male	0.99	0.99	0.97	1	34 %
78	2	femebr tibpeb	Male	Female	0.64	0.96	0.94	0.98	93 %
229	3	femebr tibxln tibpeb	Female	Female	0.73	0.97	0.96	0.98	86 %
238	4	radxln radtvd radapd femmtv	Female	Male	0.92	0.97	0.98	0.96	58 %
302	5	humxln humebr humhhd hummxd hummwd	Female	Female	0.87	0.97	0.97	0.97	75 %
339	6	femxln fembln femebr femhhd femmtv femmap	Male	Male	0.89	0.97	0.97	0.98	72 %

4.1. Machine learning and robust performance

The high accuracies achieved by the models across different data configurations underscore the significant potential of machine learning approaches in sex estimation, understood here broadly to include both classical statistical classifiers and modern non-linear or ensemble algorithms. Notably, the models demonstrated remarkable robustness when tested on an independent and geographically remote population, reinforcing their applicability across diverse forensic contexts. Linear Discriminant Analysis (LDA) consistently outperformed other models, achieving accuracies as high as 93 %, depending on the data setup. While this highlights the robustness of LDA, it also illustrates that more complex modern machine learning algorithms such as random forests, gradient boosting, or SVMs do not necessarily outperform classical linear classifiers in sex estimation. This observation is consistent with prior studies [40–43] suggesting that the choice of algorithm should consider interpretability, simplicity, and context-specific performance rather than assuming that more complex methods are inherently superior [43].

4.2. Impact of missing data

The observed variability in model performance across different data configurations highlights the critical importance of robust validation procedures. Our results demonstrate that accuracy can reach 91 % when considering all individuals and up to 93 % when restricting predictions to complete cases. However, the noticeable decline in accuracy when test data includes imputed values for incomplete individuals vividly illustrates the challenges posed by missing data in forensic contexts. This finding underscores the importance of data quality and the careful selection of imputation methods when developing and applying predictive models in forensic anthropology.

To address this challenge, we meticulously simulated various combinations of missing data using the South African dataset, enabling us to evaluate model reliability under realistic conditions. By artificially introducing unique patterns of missing variables, we demonstrated that while some models exhibit greater robustness to missing data, the choice of approach should always be guided by the specific attributes and available data for each case. This approach fills an important gap in previous research, which has often relied on ideal conditions with complete skeletal remains or lacked detailed assessments of imputation accuracy— a gap precisely addressed by our "accuracy x-factors" methodology.

These "accuracy x-factors" provide forensic practitioners with a powerful tool to reinforce predictions on a case-by-case basis, improving the reliability of sex estimates by tailoring them to the specific data completeness and quality of each individual case.

4.3. Reliability of sex labels and training data composition

Another potential limitation concerns the inclusion of several training collections in which sex was estimated anthropologically rather than documented. Although this introduces a small degree of label uncertainty, its impact on model reliability appears negligible. Importantly, all validation and performance assessments were conducted on independent samples with known biological sex, thus avoiding any circularity between training and testing phases. When estimated-sex datasets such as Goldman were excluded from training, overall accuracy decreased (from 93 % to about 90 % on the South African test sample), indicating that the gain in generalization from including these collections outweighs the minor potential bias. The possible error rate in the estimated labels is likely below 5 %, which is more than compensated by the increased sample diversity and representativeness. This interpretation is further supported by the independent study of Knecht et al. (2025) [43], which applied a similar model trained on same mixed collections to an Italian sample of known sex, achieving 95 % accuracy. Together, these findings suggest that the inclusion of high-quality estimated-sex data can enhance model robustness without compromising forensic reliability.

4.4. Threshold-based predictions

One of the most notable contributions of our study is the emphasis on adjusting prediction thresholds to reduce error rates, a topic recently highlighted by Koterova et al. (2024) [30]. Our findings reveal that setting a higher prediction threshold (e.g., 0.95 instead of 0.5) dramatically reduces errors, with profound implications for practical applications. By implementing a more conservative approach, predictions are only made when the model demonstrates a high degree of confidence, thereby providing additional guidance relevant to considerations of error rate under the Daubert criteria [2]. This threshold-based strategy introduces a quantifiable measure of uncertainty that may support considerations of reliability and admissibility of sex estimation evidence in legal contexts. Given the high stakes in forensic anthropology, where misclassifications can have profound legal, ethical, and social implications, avoiding errors must remain the paramount concern. Moreover, our findings confirm that adjusting the prediction threshold increases accuracy, but as an inherent trade-off, reduces the sample size retained. While this compromise aligns with recent studies emphasizing the impact of confidence thresholds on error rates, we strongly recommend setting a high threshold—at least 0.7 in forensic contexts—depending on the specific model used. This flexibility allows for the adjustment of the model's accuracy, ensuring it does not provide answers in uncertain cases, thereby enhancing the overall reliability of the prediction. This conservative approach minimizes error risk while maintaining robust confidence in predictions.

While the default threshold of 0.5 may be appropriate in some research or archaeological contexts, we strongly advise against its use in forensic settings, as it can lead to an unacceptably high error rate. By empowering practitioners to customize the threshold based on their specific needs, our approach enables experts to prioritize either accuracy or the quantity of predictions, depending on the specific case requirements. For example, archaeologists may opt for lower thresholds to maximize sample size for broader population studies, whereas forensic scientists must unequivocally prioritize accuracy in sensitive forensic applications. This flexibility allows field experts to adapt the tool to their expertise and case requirements while ensuring that accuracy remains a top priority in sensitive forensic applications.

4.5. Contributions of accuracy x-factors

The development of “accuracy x-factors” represents a significant advancement in forensic anthropology. By providing tailored accuracy measures for each case based on available skeletal measurements, this approach offers a realistic and transparent evaluation of prediction reliability. Furthermore, verifying predictions against accuracy levels derived from simulated missing data scenarios establishes a robust framework for assessing model reliability in real-world cases. This methodology enhances the scientific rigor of forensic anthropology while directly aligning with the growing emphasis on quantifiable measures of certainty in forensic science a core tenet of legal admissibility.

4.6. Limitations and future research

Despite the promising results, our study has some limitations that warrant further investigation. One key limitation is the imperative to validate the utilization of these “accuracy x-factors” across diverse populations. Population-specific variations in sexual dimorphism remain a significant challenge in forensic anthropology, and future studies should evaluate these methods in a broader range of global populations [28,44]. For instance, the work of Kranioti et al. (2017)

demonstrated substantial differences in sex-related traits across populations, emphasizing the need for population-specific standards [45]. Similarly, Spradley and Jantz (2011) highlighted significant variability in the accuracy of sex estimation methods when applied to different population groups [5]. In addition, in the context of the South African validation sample, it is important to consider findings from Kruger et al. (2017) [26], who reported variable levels of sexual dimorphism between different population groups within this sample. This variation may affect the performance and generalizability of predictive models and underscores the necessity of population-specific validation, even within a single geographic region.

These observations are consistent with the recent findings of Zeng et al. (2024) [46], who showed a decline in accuracy when models trained on European-derived datasets were applied to South African samples. While such results reinforce the need for population-specific evaluation, they also illustrate a fundamental challenge in forensic casework, namely, that the ancestry or population affinity of skeletal remains is often unknown in practice. For this reason, our approach prioritizes the development of a generalizable model capable of robust performance across diverse populations, rather than one optimized for a single ancestry group.

To mitigate the uncertainty introduced by population variability, we incorporated a posterior probability thresholding approach, allowing the model to produce classifications only when the confidence level is sufficiently high. This conservative strategy minimizes misclassification risk and supports cautious, case-by-case interpretation, thereby enhancing the forensic and legal reliability of the results.

While our models demonstrated a degree of robustness to population variability, the establishment of truly universal methods remains a significant challenge. As global migratory flows increase and skeletal remains become more diverse, forensic scientists require models capable of addressing this inherent complexity. This study contributes to that overarching goal but highlights the continued need for further research to confirm the broad applicability of these methods in highly heterogeneous contexts.

Additionally, future research should focus on developing more sophisticated imputation techniques for handling missing data. Advanced machine learning-based imputation methods such as those leveraging deep learning, could potentially further enhance the accuracy of sex estimations in cases with incomplete skeletal remains [47]. This avenue of research is crucial for maximizing the utility of available evidence in challenging forensic scenarios.

5. Conclusion

Our study demonstrates the substantial potential of machine learning approaches to improve sex estimation in forensic anthropology, even in the face of incomplete data. By directly addressing key challenges such as missing data, the judicious application of prediction thresholds, and the critical need for case-specific accuracy assessments, we provide a robust and transparent framework for practical applications. The introduction and utilization of "accuracy x-factors" represents a significant methodological leap forward, empowering forensic practitioners to tailor predictions to individual cases while maintaining unparalleled transparency and reliability. As forensic science increasingly demands quantifiable measures of certainty and rigorous validation, our approach makes a substantial contribution to the development of more rigorous, reliable, and ethically sound methods for sex estimation, thereby strengthening the scientific foundation of forensic evidence in legal proceedings.

CRediT authorship contribution statement

Gabriele Krüger: Writing – review & editing, Validation, Resources, Data curation. **Knecht Siam:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Pascal Adalian:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Christophe Roman:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Mustapha Ouladsine:** Supervision. **Marie Perrin:** Writing – original draft, Validation, Resources, Data curation. **Yann Ardagna:** Writing – review & editing, Validation, Resources, Data curation. **Leandi Liebenberg:** Writing – original draft, Validation, Resources, Data curation.

Ethical approval

All research procedures received institutional ethical approval and followed current European guidelines for research ethics on human remains.

Consent for publication

Not applicable.

Consent to participate

Not applicable.

Funding

Not applicable.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to thank the Laënnec Institute for supporting this project.

References

[1] H.H. De Boer, Z. Obertová, E. Cunha, P. Adalian, E. Baccino, T. Fracasso, E. Kranioti, P. Lefévre, N. Lynnerup, A. Petaros, A. Ross, M. Steyn, C. Cattaneo, Strengthening the role of forensic anthropology in personal identification: position statement by the Board of the forensic anthropology society of europe (FASE), Forensic Sci. Int. 315 (2020) 110456, https://doi.org/10.1016/j.forsciint.2020.110456.
[2] V. Daubert, Merrell Dow Pharmaceuticals, Inc. 509 U.S. 579, 589, (1993).
[3] D.H. Ubelaker, Recent advances in forensic anthropology, Forensic Sci. Res 3 (2018) 275–277, https://doi.org/10.1080/20961790.2018.1466384.
[4] L. Scheuer, Application of osteology to forensic medicine, Clin. Anat. 15 (2002) 297–312, https://doi.org/10.1002/ca.10028.
[5] M.K. Spradley, R.L. Jantz, Sex estimation in forensic anthropology: skull versus postcranial elements, J. Forensic Sci. 56 (2011) 289–296, https://doi.org/10.1111/j.1556-4029.2010.01635.x.
[6] D. Ferembach, I. Schwidetzky, M. Stloukal, Recommandations pour déterminer l'âge et le sexe sur le squelette, bmsap 6 (1979) 7–45, https://doi.org/10.3406/bmsap.1979.1945.
[7] P.L. Walker, Greater sciatic notch morphology: sex, age, and population differences, Am. J. Phys. Anthr. 127 (2005) 385–391, https://doi.org/10.1002/ajpa.10422.
[8] R.M. Thomas, C.L. Parks, A.H. Richard, Accuracy rates of sex estimation by forensic anthropologists through comparison with DNA typing results in forensic casework, J. Forensic Sci. 61 (2016) 1307–1310, https://doi.org/10.1111/1556-4029.13137.
[9] P. Murail, J. Bruzek, J. Braga, A new approach to sexual diagnosis in past populations. Practical adjustments from Van Vark's procedure, Int. J. Osteoarchaeol 9 (1999) 39–53, https://doi.org/10.1002/(SICI)1099-1212(199901/02)9:1<39::AID-OA458>3.0.CO;2-V.
[10] V. Alunni-Perret, P. Staccini, G. Quatrehomme, Sex determination from the distal part of the femur in a French contemporary population, Forensic Sci. Int. 175 (2008) 113–117, https://doi.org/10.1016/j.forsciint.2007.05.018.
[11] F. Curate, F. Mestre, S.J. Garcia, Sex assessment with the radius in Portuguese skeletal populations (late 19th – early to mid 20th centuries), Leg. Med. 48 (2021) 101790, https://doi.org/10.1016/j.legalmed.2020.101790.

[12] F. Curate, J. Coelho, D. Gonçalves, C. Coelho, M.T. Ferreira, D. Navega, E. Cunha, A method for sex estimation using the proximal femur, Forensic Sci. Int. 266 (2016) 579.e1–579.e7, https://doi.org/10.1016/j.forsciint.2016.06.011.
[13] M.Y. Işcan, D. Shihai, Sexual dimorphism in the Chinese femur, Forensic Sci. Int. 74 (1995) 79–87, https://doi.org/10.1016/0379-0738(95)01691-b.
[14] M. Slaus, Z. Bedić, D. Strinović, V. Petrovečki, Sex determination by discriminant function analysis of the tibia for contemporary Croats, Forensic Sci. Int. 226 (2013) 302.e1-4, https://doi.org/10.1016/j.forsciint.2013.01.025.
[15] M. Steyn, M.Y. Işcan, Sex determination from the femur and tibia in South African whites, Forensic Sci. Int. 90 (1997) 111–119, https://doi.org/10.1016/s0379-0738(97)00156-4.
[16] J. Albanese, A method for estimating sex using the clavicle, humerus, radius, and ulna, J. Forensic Sci. 58 (2013) 1413–1419, https://doi.org/10.1111/1556-4029.12188.
[17] E.F. Kranioti, M. Michalodimitrakis, Sexual dimorphism of the humerus in contemporary Cretans–a population-specific study and a review of the literature, J. Forensic Sci. 54 (2009) 996–1000, https://doi.org/10.1111/j.1556-4029.2009.01103.x.
[18] L. Nogueira, F. Santos, F. Castier, S. Knecht, C. Bernardi, V. Alunni, Sex assessment using the radius bone in a French sample when applying various statistical models, Int. J. Leg. Med. (2023), https://doi.org/10.1007/s00414-023-02981-8.
[19] S.D. Tallman, A.I. Blanton, Distal humerus morphological variation and sex estimation in modern thai individuals, J. Forensic Sci. 65 (2020) 361–371, https://doi.org/10.1111/1556-4029.14218.
[20] W. Jongmuenwai, M. Boonpim, T. Monum, A. Sintubua, S. Prasitwattanaseree, P. Mahakkanukrauh, Sex estimation using radius in a Thai population, Anat. Cell Biol. 54 (2021) 321–331, https://doi.org/10.5115/acb.20.319.
[21] R. Purkait, Measurements of ulna–a new method for determination of sex, J. Forensic Sci. 46 (2001) 924–927.
[22] L.S. Cowal, R.F. Pastor, Dimensional variation in the proximal ulna: evaluation of a metric method for sex assessment, Am. J. Phys. Anthr. 135 (2008) 469–478, https://doi.org/10.1002/ajpa.20771.
[23] F. Introna, M. Dragone, P. Frassanito, M. Colonna, Determination of skeletal sex using discriminant analysis of ulnar measurements, Boll. Soc. Ital. Biol. Sper. 69 (1993) 517–523.
[24] R. Srivastava, V. Saini, R.K. Rai, S. Pandey, T.B. Singh, S.K. Tripathi, A.K. Pandey, Sexual dimorphism in ulna: an osteometric study from India, J. Forensic Sci. 58 (2013) 1251–1256, https://doi.org/10.1111/1556-4029.12158.
[25] M.A. Bidmos, P. Mazengenya, Accuracies of discriminant function equations for sex estimation using long bones of upper extremities, Int. J. Leg. Med. 135 (2021) 1095–1102, https://doi.org/10.1007/s00414-020-02458-y.
[26] G.C. Krüger, E.N. L'Abbé, K.E. Stull, Sex estimation from the long bones of modern South Africans, Int. J. Leg. Med. 131 (2017) 275–285, https://doi.org/10.1007/s00414-016-1488-z.
[27] K.E. Stull, E.N. L'Abbé, S.D. Ousley, Subadult sex estimation from diaphyseal dimensions: STULL et al. Am. J. Phys. Anthr. 163 (2017) 64–74, https://doi.org/10.1002/ajpa.23185.
[28] S. Knecht, F. Santos, Y. Ardagna, V. Alunni, P. Adalian, L. Nogueira, Sex estimation from long bones: a machine learning approach, Int. J. Leg. Med. (2023), https://doi.org/10.1007/s00414-023-03072-4.
[29] A. Thurzo, H.S. Kosnáčová, V. Kurilová, S. Kosmeľ, R. Beňuš, N. Moravanský, P. Kováč, K.M. Kuracinová, M. Palkovič, I. Varga, Use of advanced artificial intelligence in forensic medicine, forensic anthropology and clinical anatomy, Healthcare 9 (2021) 1545, https://doi.org/10.3390/healthcare9111545.
[30] A. Pilmann Kotěrová, F. Santos, Š. Bejdová, R. Rmoutilová, M.H. Attia, A. Habiba, J. Velemínská, J. Brůžek, Prioritizing a high posterior probability threshold leading to low error rate over high classification accuracy: the validity of MorphoPASSE software for cranial morphological sex estimation in a contemporary population, Int. J. Leg. Med. 138 (2024) 1759–1768, https://doi.org/10.1007/s00414-024-03215-1.
[31] D.R. Hunt, J. Albanese, History and demographic composition of the Robert J. Terry anatomical collection, Am. J. Phys. Anthr. 127 (2005) 406–417, https://doi.org/10.1002/ajpa.20135.
[32] G.C. Krüger, E.N. L'Abbé, K.E. Stull, Sex estimation from the long bones of modern South Africans, Int. J. Leg. Med. 131 (2017) 275–285, https://doi.org/10.1007/s00414-016-1488-z.
[33] J. Bruzek, A method for visual determination of sex, using the human hip bone, Am. J. Phys. Anthr. 117 (2002) 157–168, https://doi.org/10.1002/ajpa.10012.
[34] T.W. Phenice, A newly developed visual method of sexing the os pubis, Am. J. Phys. Anthr. 30 (1969) 297–301, https://doi.org/10.1002/ajpa.1330300214.
[35] R. Martin, K. Saller, Lehrbuch der, Anthropologie 1 (1957) 433–476.
[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. (2011) 2825–2830.
[37] P.R. Avent, C.E. Hughes, H.M. Garvin, Applying posterior probability informed thresholds to traditional cranial trait sex estimation methods, J. Forensic Sci. 67 (2022) 440–449, https://doi.org/10.1111/1556-4029.14947.
[38] P. Galeta, J. Brůžek, Sex estimation using continuous variables: Problems and principles of sex classification in the zone of uncertainty. Proceeding of Statistics and Probability in Forensic Anthropology, Elsevier, 2020, pp. 155–182, https://doi.org/10.1016/B978-0-12-815764-0.00016-2.
[39] I. Jerković, Ž. Bašić, Š. Anđelinović, I. Kružić, Adjusting posterior probabilities to meet predefined accuracy criteria: a proposal for a novel approach to osteometric sex estimation, Forensic Sci. Int. 311 (2020) 110273, https://doi.org/10.1016/j.forsciint.2020.110273.
[40] F. Santos, P. Guyomarc'h, J. Bruzek, Statistical sex determination from craniometrics: comparison of linear discriminant analysis, logistic regression, and support vector machines, Forensic Sci. Int. 245 (2014) 204.e1-8, https://doi.org/10.1016/j.forsciint.2014.10.010.
[41] E. Nikita, P. Nikitas, On the use of machine learning algorithms in forensic anthropology, Leg. Med. 47 (2020) 101771, https://doi.org/10.1016/j.legalmed.2020.101771.
[42] Ph Du Jardin, J. Ponsaillé, V. Alunni-Perret, G. Quatrehomme, A comparison between neural network and other metric methods to determine sex from the upper femur in a modern French population, Forensic Sci. Int. 192 (2009) 127.e1–127.e6, https://doi.org/10.1016/j.forsciint.2009.07.014.
[43] S. Knecht, P. Morandini, L. Biehler-Gomez, Y. Ardagna, M. Perrin, C. Cattaneo, C. Roman, P. Adalian, Interpretable machine learning for individualized sex estimation from long bones, Int. J. Leg. Med. (2025), https://doi.org/10.1007/s00414-025-03635-7.
[44] S. Toy, Y. Secgin, Z. Oner, M.K. Turan, S. Oner, D. Senol, A study on sex estimation by using machine learning algorithms with parameters obtained from computerized tomography images of the cranium, Sci. Rep. 12 (2022) 4278, https://doi.org/10.1038/s41598-022-07415-w.
[45] E.K. Kranioti, J.G. García-Donas, P.S. Almeida Prado, X.P. Kyriakou, H.C. Langstaff, Sexual dimorphism of the tibia in contemporary Greek-Cypriots and Cretans: forensic applications, Forensic Sci. Int. 271 (2017) 129.e1–129.e7, https://doi.org/10.1016/j.forsciint.2016.11.018.
[46] S. Zeng, E. Cunha, F. Curate, Sex estimation using adult femora and humeri trained on a Portuguese reference sample, and tested on Portuguese and South African samples, Aust. J. Forensic Sci. (2024) 1–17, https://doi.org/10.1080/00450618.2024.2429720.
[47] P. Oura, J.-A. Junno, D. Hunt, P. Lehenkari, J. Tuukkanen, H. Maijanen, Deep learning in sex estimation from knee radiographs – A proof-of-concept study utilizing the Terry Anatomical Collection, Leg. Med. 61 (2023) 102211, https://doi.org/10.1016/j.legalmed.2023.102211.