Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07981-0

ORIGINAL ARTICLE

Employing deep learning for sex estimation of adult individuals using 2D images of the humerus

Javier Venema[1,4] · David Peula[2] · Javier Irurita[2] · Pablo Mesejo[1,3,4]

Received: 17 August 2022 / Accepted: 17 October 2022

Abstract

Biological profile estimation, of which sex estimation is a fundamental first stage, is a really important task in forensic human identification. Although there are a large number of methods that address this problem from different bone structures, mainly using the pelvis and the skull, it has been shown that the humerus presents significant sexual dimorphisms that can be used to estimate sex in their absence. However, these methods are often too subjective or costly, and the development of new methods that avoid these problems is one of the priorities in forensic anthropology research. In this respect, the use of artificial intelligence may allow to automate and reduce the subjectivity of biological profile estimation methods. In fact, artificial intelligence has been successfully applied in sex estimation tasks, but most of the previous work focuses on the analysis of the pelvis and the skull. More importantly, the humerus, which can be useful in some situations due to its resistance, has never been used in the development of an automatic sex estimation method. Therefore, this paper addresses the use of machine learning techniques to the task of image classification, focusing on the use of images of the distal epiphysis of the humerus to classify whether it belongs to a male or female individual. To address this, we have used a set of humerus photographs of 417 adult individuals of Mediterranean origin to validate and compare different approaches, using both deep learning and traditional feature extraction techniques. Our best model obtains an accuracy of 91.03% in test, correctly estimating the sex of 92.68% of the males and 89.19% of the females. These results are superior to the ones obtained by the state of the art and by a human expert, who has achieved an accuracy of 83.33% using a state-of-the-art method on the same data. In addition, the visualization of activation maps allows us to confirm not only that the neural network observes the sexual dimorphisms that have been proposed by the forensic anthropology literature, but also that it has been capable of finding a new region of interest.

Keywords Deep learning · Biomedical image analysis · Forensic anthropology · Biological profile estimation · Sex estimation · Image classification

1 Introduction

In the field of forensic anthropology (FA),[1] sex estimation is a task of great importance as a first step in the identification of individuals from their skeletal remains, conditioning how later phases of the identification process are addressed [2]. It is usually carried out by the analysis of the pelvis [3] or the skull [4, 5], although in their absence there are other methods that also obtain good results, such as odontometry [6] and the metric or morphological analysis of the postcranial skeleton [7], as would be the case with the morphological analysis of the distal epiphysis of the humerus [8–10]. The latter is of special interest due to the resistance of the bone and the preservation of its distal end (see Fig. 1). These properties make the humerus an ideal option for sex estimation in the absence of the pelvis and the skull, as it can resist when those two are not available. In general, these methods focus on the visual observation of the sexual dimorphisms (difference in size and shape between male and female individuals) that are present in the bones, which makes them subjective, error-prone and hardly replicable (see Fig. 2). However, there are also methods that focus on a geometric morphometric analysis [10], being less subjective but also much slower.

Javier Venema and David Peula have been contributed equally to this work.

Extended author information available on the last page of the article

[1] Forensic anthropology deals with the study of the skeleton and its application to medico-legal problems [1].

Published online: 12 November 2022

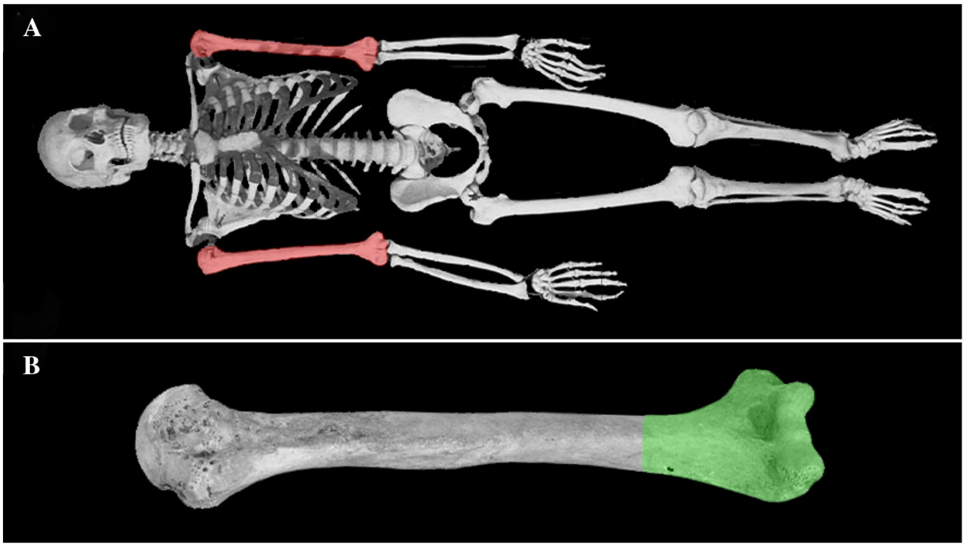


Fig. 1 Location of the humerus in the postcranial skeleton (**A**). Distal epiphysis of the humerus (**B**). Images were obtained and modified from [8]

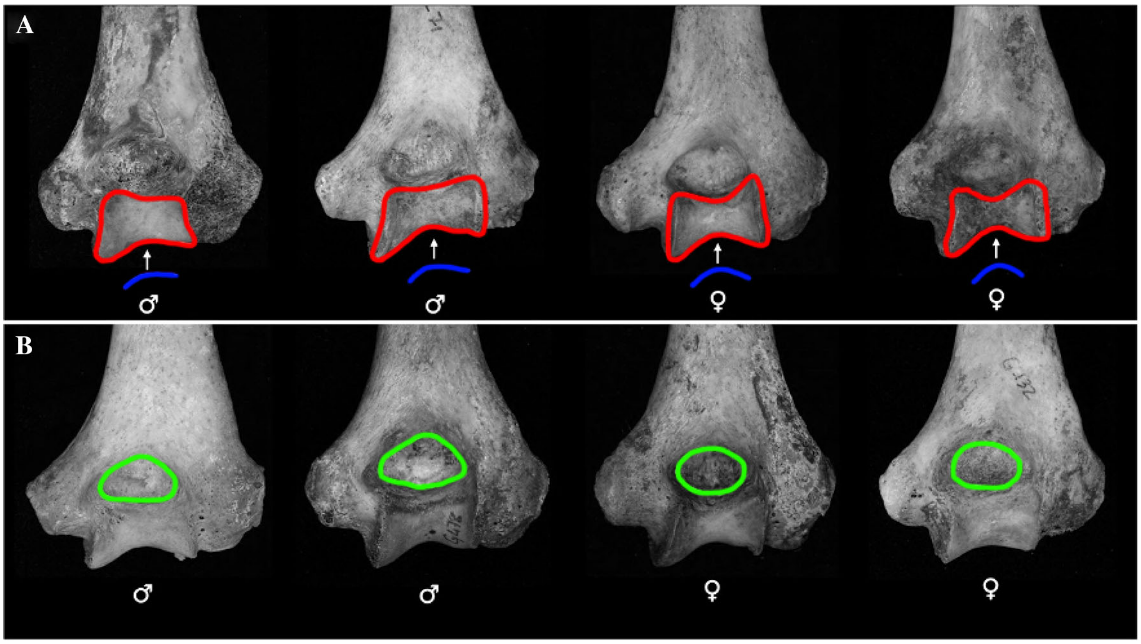


Fig. 2 Sexual dimorphisms proposed by [8] for sex estimation from the humerus. The trochlear constriction (in blue): the angle of the central part of the trochlea respect to the central axis of the humerus tends to curve gradually and to a lesser extent in the masculine, unlike the feminine, which tends to curve sharply and more accentuated; the trochlear symmetry (in red): the central part of the trochlea tends to be more symmetrical in the female humerus than in the male humerus; morphology of the olecranial fossa (in green): it is usually more superficial and triangular in the male humerus, unlike in the female, which tends to be deeper and oval. Images were obtained and modified from [8]

As an alternative, artificial intelligence (AI), concretely *machine learning* (ML), and mainly *deep learning* (DL) as a set of techniques within it could reduce the classical methods subjectivity. In addition, it could automate, accelerate and reduce the costs in time of the techniques used in FA. In this sense, the last advances in deep neural networks, and more specifically in convolutional neural networks (ConvNets), may allow us to address the task of sex estimation in an automated manner.

The aim of this paper is to obtain a model that automates the task of sex estimation in adult individuals from the humerus bone, providing forensic anthropologists with an alternative method for solving human identification problems in a precise, objective, efficient, easy and

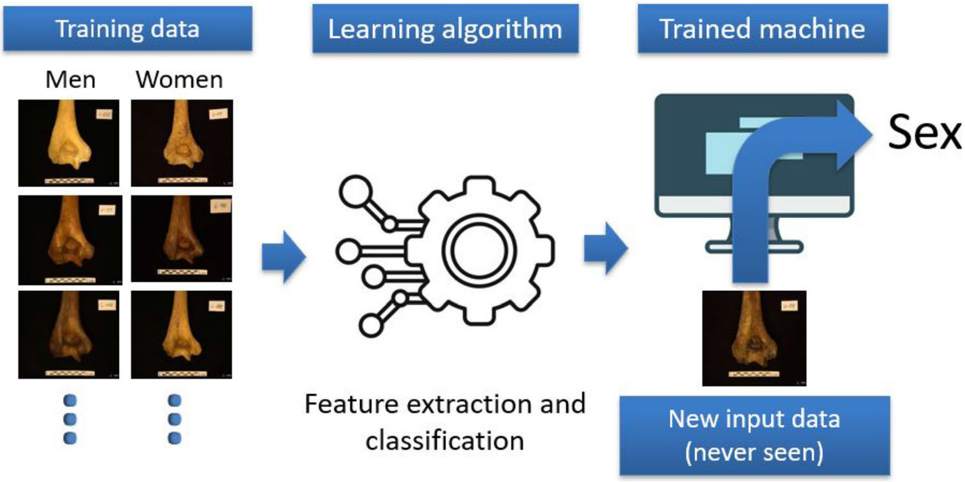

Training data
Men
Women
Learning algorithm
Feature extraction and classification
Trained machine
Sex
New input data (never seen)

Fig. 3 The main objective of this paper is to obtain an automatic, fast and easy to use method that, once trained with labeled images, is capable of estimating sex when receiving new images of the humerus. This figure outlines the methodology that is followed to build that model

reproducible manner (see Fig. 3). Since little data are available, we will perform a comparative analysis between classical computer vision techniques, which use hand-crafted features and a separated classification, and recent DL models trained in an end-to-end manner. Regarding those end-to-end models and also because of the little data available, our first approach will be based on applying transfer learning to well-established networks. After performing the analysis, we will test the best model and compare its results to those obtained by a human expert who used a visual method [8] on the same data and with state-of-the-art results in sex estimation using the humerus [10]. Finally, we will visualize activation maps to have an intuition about which regions of the input images have a larger impact on the output. This will allow us to identify new discriminative regions in anthropological terms. Therefore, the main contributions of this paper are:

- The development of an automatic and precise sex estimation method, which is more objective and reproducible than the visual method proposed in [8] and less time consuming than the morphogeometric method proposed in [10].
- The objective validation of the dimorphic traits that have already been proposed for sex estimation from the humerus, as well as the discovery of a new region of interest.

This paper is organized as follows. Section 2 describes previous work, both in FA methods to estimate sex using the humerus and in sex estimation using AI techniques. Section 3 discusses the experimental protocols (dataset and AI techniques) used to find an accurate sex estimation model. The results of these experiments are shown and discussed in Sect. 4, which also compares the best method with a human expert. Conclusions and future works are outlined in Section 5.

2 Related works

In this case, related works should be divided into two different groups. On the one hand, since this work is related to FA, and more precisely with sex estimation using the humerus bone, we must describe which are the main methods used in the field to perform this task. Moreover, since we want to compare our results with those that can be obtained with classical FA methods, it is important to show their performance. On the other hand, it is also fundamental to describe the main works that use AI for the task of forensic sex estimation.

ConvNets have been extensively applied in the field of medicine, but there is little work in FA and even less in sex estimation. There are proposals that have applied DL techniques for classifying images of different bones to estimate sex. Bewes et al. [11] achieved a 95% accuracy using skull images of adult individuals by applying transfer learning to a GoogleNet [12] model that had been pre-trained on ImageNet [13]. Rajee and Maythili applied fine-tuning to the ResNet50 [14] architecture using 1000 noise-filtered dental X-ray images, achieving an accuracy of 98.27%. They also visualized activation maps to observe what the model was learning. Vila et al. proposed a method for sex estimation using panoramic dental radiographs. They tested three different approaches to perform this task. The first one is DASNet, which was proposed in [15] as a ConvNet for age estimation from panoramic radiographs. DASNet is composed by two ConvNets, one for age estimation and another one for sex estimation. The latter is used to extract sex features that are then concatenated with the age features obtained by the other network just before performing the age prediction (as sex is important to estimate age). Although DASNet is not conceived for sex estimation, it obtains state-of-the-art results, so the authors proposed DSANet, which uses the same structure as DASNet but inversely, that is, it uses age features to estimate sex instead of sex features to estimate age. They also tested using a VGG16-based [16] architecture, but DSANet gave the best results, getting an accuracy over 80% in every age group, including children, and an accuracy between 90% and 96% in every adult group (over 16 years old). Cao et al. [17] used pelvis CT scans to estimate sex. They obtained 3D shapes from those CT scans and then extracted three views of interest from them. As each of those views is a 2D image, they can use a 2D ConvNet, in this case GoogleNet [12] pre-trained on ImageNet [13], for sex estimation from each of them. When combining the three models, which is done using an average weighted by the test accuracy of each of them, they obtain an accuracy of 97.1% in a single-blind trial, being more precise than the anthropologist with whom they compare. In [18], the authors followed a similar approach, but in this case, they obtained an accuracy of 100% in test using only the images corresponding to the view of the ventral pubis.

Some authors, like Ortega et al. [19] and Kaloi et al. [20], also used ConvNets for estimating sex, but they focused on children individuals. The former used pelvis images to perform a comparative study between different methods, of which VGG16 [16] was the best option, getting an accuracy of 59% that was very close to the 61% obtained by a human expert. The latter used hand radiographs and obtained a 98% accuracy by training their own architecture, which contained four convolutional layers and two dense layers, ReLU as activation function and a dropout [21] rate of 0.8, *from scratch* using the Adam [22] algorithm.

Other research has focused on using techniques that do not involve the use of ConvNets. On the one hand, there are some papers that develop semi-automatic methods. In these works, a forensic anthropologist manually extracts some geometric features from the bone. Those features are then used to train a classifier. [23] and [24] are relevant examples in this regard. In the former, the authors extracted 38 features from the pelvis and used them to train a partial least squares model. In the latter, the authors positioned 16 semilandmarks over images of the posterior view of the humerus and used them to train an LDA model. As it can be seen, since this kind of methods requires a manual feature extraction phase, they are slow and difficult to apply.

On the other hand, there are also some fully automatic methods that just do not use ConvNets. For instance, in [25], the authors employ the wavelet transform to create a method for the objective quantification of sexually dimorphic features and use it to successfully estimate sex in a pilot sample of three-dimensional meshes of the skull. In [26], the authors present a hybrid approach of artificial neural networks and metaheuristics to estimate sex from hand radiographs of children. To do that, they divide the images into six age groups and measure the length of 19 bones of the hand in an automated manner. Using those lengths as features, they get an accuracy of $\sim 70\%$. Finally, Imaizumi et al. [27] developed a method for sex estimation of adult individuals using 3-dimensional shapes of the skull. They obtained 100 skull shapes from CT scans and created three different models from them: use the whole skull, the cranium only and the mandible only. They obtained a point cloud from the meshes and then used partial least squares regression for dimensionality reduction. Using SVM for classification, they obtained a 100% accuracy both using the skull and its separate parts. Those results were obtained in a double-looped cross-validation process, where the first loop is used for hyperparameter tuning and the second one for accuracy estimation.

It can be seen that every work that has been mentioned uses images from the skull, pelvis, hand or teeth. As a result, neither the humerus nor any other long bone has been used in the development of any automated sex estimation method. However, sex estimation from a robust bone, such as the humerus, can be important in multiple situations in which the aforementioned methods would not be applicable because of the deterioration, breakage or disappearance of the used bones, as well as in situations that require the study of isolated bone remains without anatomical connection, such as the study of ossuaries. Moreover, these situations often require the identification and therefore sex estimation of a great quantity of individuals, which makes the automation of the estimation even more important. Some examples include natural disasters, genocides, terrorism, accidents involving several people or mass graves.

3 Materials and methods

3.1 Dataset

We have worked with a dataset of humerus photographs obtained from two identified collections (from the Cemetery of San Jose and the Cemetery of Lucena) located in the Physical and Forensic Anthropology Laboratory of the University of Granada (Spain). Individuals in the collections used are of current chronology (20th century) and population of Mediterranean origin. These are identified collections, of which reliable information is available thanks to the existence of death and/or burial data. Around 90% of the skeletons that make up these collections are in a good state of conservation. Exclusion criteria for the study have been a poor state of preservation, pathological alterations, subadult individuals and the lack of antemortem information. Once they have been applied, we are left with 401 individuals, with 213 males and 188 females whose ages range between 22 and 102 years old (see Table 1 for more information). We used images of the right humeri.

The photographs were made with a Lumix DC-GH5 camera (Panasonic Corp, Japan) with Lumix G Vario 14-42 mm lens (Panasonic Corp, Japan). In order for the epiphysis to remain focused and undistorted, the distal end of the humerus was placed in the center of the frame with a scale and label of the corresponding individual. The photographs were taken with a diaphragm value f/8 to 0.4 meters with a focal length of 42 mm. The resulting images are shown in Fig. 4.

These images have been divided in a random and stratified manner into a training set, with 80% of the images, and a test set, which will be used to evaluate the best model once it has been selected by the application of 5-fold cross-validation. This test set initially contained all of the images that were not in the training one, but we dropped three of them since the expert with whom we want to compare our results dropped them because of their bad state of preservation. By doing this, we can perform this comparison with the exact same data.

Table 1 Information of age and number of individuals used in the study

	Men	Women	Total
Cementery of San Jose	125	104	229
Average age (years)	65	76	
Cementery of Lucena	87	85	172
Average age (years)	73	82	

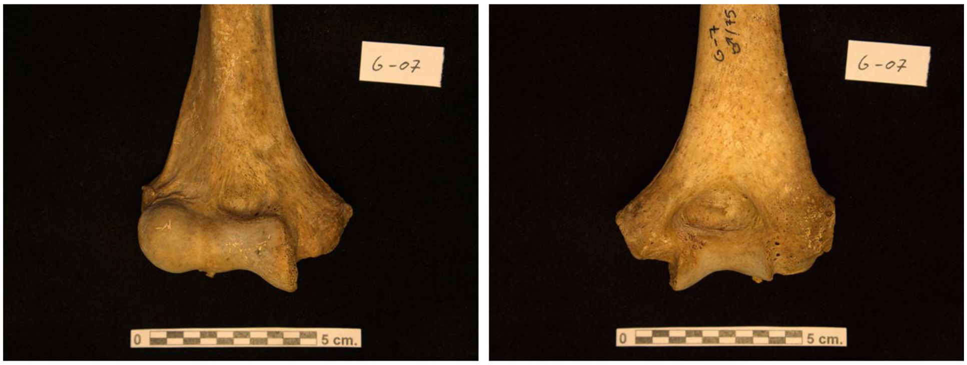
3.1.1 Methods

As previously said, because of the little data available, we have carried out a comparative analysis between handcrafted feature extraction methods followed by a classifier and neural models trained end-to-end. In the first case, HOG features [28] have been extracted and provided as the input of three different models: SVM [29], random forest [30] and logistic regression. HOG features are usually good enough for image classification tasks, as they capture shape information well. Other well-known feature extraction algorithms, such as SIFT [31], SURF [32] or ORB [33], which detect points of interest before the feature extraction, are usually more suitable for different tasks, such as object matching. With respect to the tested classification algorithms, SVM is usually used after HOG feature extraction, while logistic regression and random forest have been introduced as a simple and a more complex but powerful model.

As for the DL-based approach, we have used transfer learning techniques over two different architectures: VGG16 [16] and ResNet50 [14], both of them pre-trained on ImageNet, as well as early stopping to halt training when it starts worsening and the Grad-CAM algorithm [34] to obtain activation maps that allow us to visualize what the network is learning. Adam [22] was employed as optimization algorithm. The use of transfer learning over well-established architectures has two main reasons. Firstly, it has been shown that using a properly tuned well-established architecture works just as well as using an ad hoc network [35]. Secondly, we have little data. In this sense, we do not have enough data to train a model from scratch, as it is difficult to obtain big amounts of annotated data in the field of FA. Transfer learning is the general approach to address this problem [36].

For model selection, we follow a simple procedure. In the case of the more traditional ML techniques, we first obtain the HOG features of the images to use them as the input to the models. Once this is done, we use the grid search technique with 5-fold cross-validation for hyperparameter tuning. In the case of DL techniques, it is impossible to use grid search cross-validation due to the computational complexity of the methods. In this case, starting from a common initial configuration for all the architectures that are tested, we follow an iterative approach in which we change the value of a certain hyperparameter or introduce some technique at each step. Then, we observe how that change affects the values of the error metrics using 5-fold cross-validation, and what effect does it have in the evolution of the loss function (both in training and validation for every fold) along the training procedure. More details on the aforementioned initial setup will be given in the experiments section.

Fig. 4 Images that are taken for each individual of the collection. They include both the anterior and posterior view of the bone



(A) Anterior view. **(B)** Posterior view.

The usage of 5-fold cross-validation allows us to obtain an optimal version of every model by tuning their hyperparameters. Once we have these optimal models, we select the best among them and evaluate it on a never seen test set. By doing this, we are able to obtain non-biased results that can be used to compare the proposed method with a human expert and the state of the art in FA. After performing this comparison and to increase the interpretability of the selected model, which ended up being ResNet50, we used the Grad-CAM algorithm [34], which obtains activation maps that allow us to visualize what the network is learning. More precisely, these activation maps highlight the characteristics of the input that strongly influence the output of the network, allowing us to explain why that output is given and to compare if the important regions are the same for the human expert and the model. This visualization technique is related to explainable AI, which is a field of great and increasing importance [37, 38] in AI research, and that could be even of greater importance in FA because of the need to justify decisions when applied to medico-legal problems.

4 Experiments

4.1 Preliminary experiments

Although we have two photographs per individual, with one image of the posterior view and another one of the anterior view, we cannot use them all to train the same model. To select which of them would be used, we performed an initial test in which we took VGG16 [16] pretrained on ImageNet [13], substituted the last layer by another one with just one neuron (for binary classification) and applied transfer learning freezing the whole network but the added layer. Then, we separately trained the model with both the anterior and the posterior views. The results showed that the posterior view was slightly better for estimating sex, as the model reached an *accuracy* of 86% when used, being higher than the *accuracy* of 85.5% obtained when using the anterior view.

Table 2 Comparison of the best results obtained by each model in validation (best in bold)

Model	Accuracy (%)	Precision (%)	Recall (%)
VGG16	**88.8**	89.1	**89.9**
ResNet50	87.8	**91.4**	85.9
SVM	86.6	88.0	86.4
Random forest	86.9	88.7	86.4
Logistic regression	86.3	88.8	84.6
Random	50.0	52.8	49.9

We include a random model (last row of the table) to show that every classifier is clearly superior to it

The *accuracy* (percentage of correctly classified examples) has been the main metric that we have observed to evaluate model performance, although due to the slight imbalance in the data, *precision* and *recall* have also been calculated. In this case, *precision* refers to the percentage of examples classified as males that are truly males (so it decreases when the number of incorrectly classified females increases), while *recall* refers to the percentage of examples that are truly males and that were classified as such (so it decreases when the number of incorrectly classified males increases). In addition, to improve the interpretability of the results in the final comparison, the confusion matrix has also been obtained, which allows us to observe the specific results without summarizing them in a single value.

Table 2 shows the values of the metrics in cross-validation for the best version of each of the tested models. While it can be seen that all models perform considerably well, it is important to note that the DL techniques are slightly superior. Regarding the selection of the best model, although it can be observed from the values of the metrics that VGG16 seems better than ResNet50, it is necessary to highlight two important factors. The first one is that ResNet50 is lighter than VGG16, so it can be more easily aggregated into an application. The second and most important one is that in the graphs that show the evolution of the validation loss along the training process we observed a much higher variability for VGG16 than for ResNet50. By this, we mean that while ResNet50 loss gradually declined along consecutive epochs, VGG16 loss was varying a lot, with high increases and decreases (what we called variability) during the training process. This is due to the higher learning rate used for the optimizer in the case of VGG16 (as it improved cross-validation results). This second factor is very important due to the use of early stopping, since the stopping and evaluation of the model are done on the same validation set for every fold. As it can be noted, this introduces a slight bias in the validation process, because we stop when the model works best and then evaluate it on the same data. This bias will be higher if there is a lot of variability, since the model worsens more after stopping training, which makes us think that ResNet50 will extrapolate better to the test set and to a real scenario. The best version of VGG16 that does not have so much variability between epochs is slightly worse than ResNet50, and that conditioned us to take ResNet50 as the best model.

The tested and best hyperparameters for each model were the following:

- VGG16: Fine-tuning of the last 0, 2, 4 and 6 layers. We started with 0 (no fine-tuning) and increased from there until the results started worsening because of an increase in variance. We retrain only a few layers because of the danger of overfitting due to the little data available. Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 0.001, 0.0001, 0.01 and 0.1 for the first tuning phase (before unfreezing the layers at the end of the model, which means training only the classification layer); and 1e−5 and 1e−6 for the second tuning phase (after unfreezing the layers). Batch size of 32, 64 and 128. Early stopping with patience of 5 epochs and a maximum of 30 epochs, as we observed that we needed no more than that. The best version retrained the last four layers, used the Adam optimizer with a learning rate of 0.01 in the first phase and of 1e−5 in the second phase and took a batch size of 32.
- ResNet50: Fine-tuning of the last 0, 6, 10 and 14 layers. As in VGG16, we started with 0 (no fine-tuning) and increased from there. In this case retraining 10 and 14 layers gave similar results, obtaining a bias close to 0 that made it unnecessary to increase the number of retrained layers to reduce it. Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 0.001, 0.0001 and 0.01 for the first tuning phase (before unfreezing the layers at the end of the model, which means training only the classification layer); and 1e−5 and 1e−6 for the second tuning phase (after unfreezing the layers). Batch size of 32 and 64. Early stopping with patience of 5 epochs and a maximum of 30 epochs. The best version retrained the last ten layers, used Adam optimizer with a learning rate of 0.001 in the first phase and of 1e−5 in the second phase and took a batch size of 32.
- SVM: For the regularization hyperparameter C, where the higher the C the weaker the regularization, we start trying with consecutive values in a logarithmic scale from 0.01 to 10. After observing that 0.1 was the best option we refined the hyperparameter with close values to it (from 0.04 to 0.08 by 0.02 and from 0.2 to 0.8 by 0.2) only for the linear kernel once we saw it was the best one. For the kernel, we tried linear (no kernel), Gaussian, sigmoid and polynomial with degrees 2, 3, 4 and 5. For the γ parameter of the Gaussian and sigmoid kernel, which is used in scikit-learn to obtain σ, we tried $1/m$ and $1/(m \times Var(X_{train}))$ where m is the number of features and $Var(X_{train})$ is the variance of the training set, as those are common values. The best option was using a linear kernel with $C = 0.2$.
- Random forest: 50 and 100 trees, as we saw these were enough estimators. For m the number of features used to split each node, we tried with $m = p$, $m = \sqrt{p}$ and $m = log_2 p$, where p is the total number of features, as these are the common values. For the splitting criteria, we used Gini and the entropy; and for the minimum number of samples required to split a node (this is to reduce variance) we tried 2 (splitting at every impure node), $0.1n$ and $0.2n$, being n the number of samples. The best version uses 50 estimators, Gini as splitting criteria, $m = p$ (so we end up using bagging) and $0.1n$ samples required to split a node.
- Logistic regression: We tried both L1 and L2 regularization with a C of 0.01 to 10 once more in consecutive values of a logarithmic scale. After seeing that the best option was L2 regularization with $C = 1$, we tried to refine C by trying with 0.5, 2, 3, 4 and 5 first, and with every value from 1.25 to 2.75 by 0.25 after that. The best option was using L2 regularization with $C = 1.5$.

Table 3 Comparison between the best model (in test), a human expert (with the same data) and the morphogeometric method without using the centroid size (NCS) and using it (WCS)

	Men (%)	Women (%)	Total (%)
Human expert	80.49	86.49	83.33
ResNet50	**92.68**	89.19	91.03
Morphogeometric NCS [10]	77.92	71.78	75.19
Morphogeometric WCS [10]	90.77	**94.88**	**92.60**

The table presents the percentage of correct classifications (best in bold)

- For the HOG features extraction, we used the default parameters of the algorithm, which are detailed in [28], as they are not usually modified.

All the experiments have been performed using *Google Colaboratory*. We have used Keras (2.8.0) over tensorflow (2.8.2) for the DL experiments, and OpenCV (4.6.0) and scikit-learn (1.0.2) for traditional ML techniques. The code and the best model (ResNet50) weights are available as supplementary material. A web application for sex estimation using humerus images will be available at the Panacea Cooperative Research website (https://www.panacea-coop.com/).

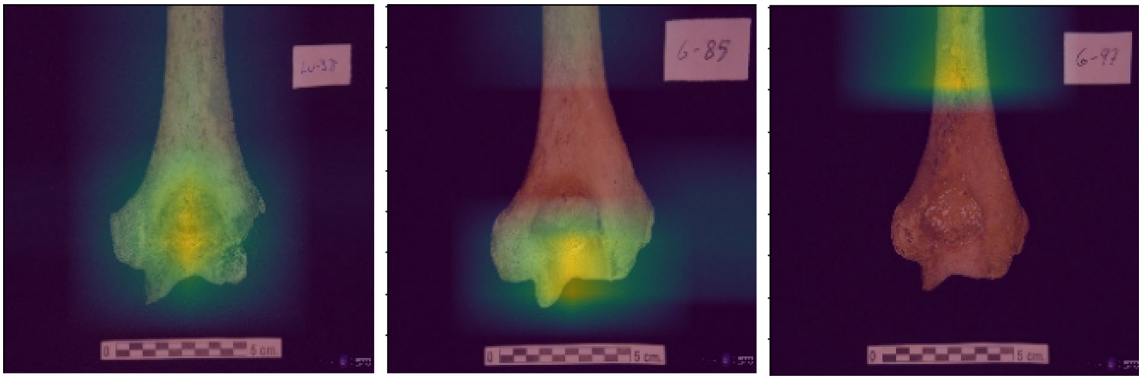
4.2 Comparison with state-of-the-art and discussion

Once ResNet50 has been selected as the best model, its results in test can be compared with the human expert using the state-of-the-art visual method [8], as well as with the morphogeometric method proposed in [10] without using the centroid size (since ResNet50 does not have that information) and using it (since it is part of the method). The comparison with the expert is performed using exactly the same data, while for the morphogeometric method we used the best results (obtained with the same posterior view that we used) that are given in [10] with a set of 32 adult females and 40 adult males. The overall comparison is shown in Table 3, while Table 4 displays the confusion matrices obtained by ResNet50 and by the human expert.

Results show that the DL model that we have developed is able to obtain better results than a human expert using the same data. It is also more accurate than the method proposed in [10] when the centroid size is not introduced, but slightly worse than this same method when the centroid size is added. In respect with our objectives, we have not only succeeded in the development of a competitive, efficient and objective automatic model, but we have also improved the results obtained by the methods that are currently used and that add no extra information such as the centroid size. When the centroid size is added, the

Table 4 Confusion matrices for ResNet50 (in test) and the human expert

	Predicted woman	Predicted man		Predicted woman	Predicted man
Woman	32	5	Woman	33	4
Man	8	33	Man	3	38
Human expert			ResNet50		



(A) Olecranon fossa. **(B)** Trochlea. **(C)** New region.

Fig. 5 Results of applying Grad-CAM on correctly classified images (the more yellow the region, the higher its importance in the classification). We show the three regions that are mainly observed independently, but for most images two or all of them are observed at the same time

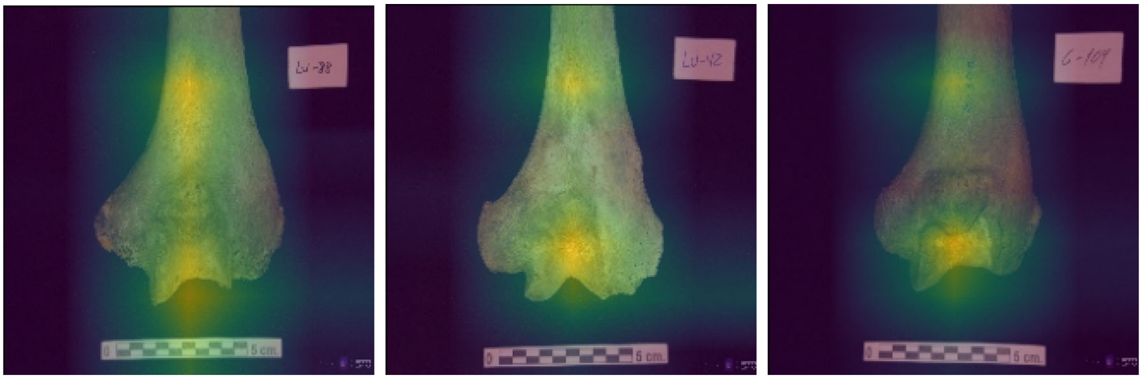


Fig. 6 Examples that where misclassified by the human expert but correctly classified by the model. While the expert focuses on the olecranon fossa, the network gives more attention to the other two dimorphic regions in these specific examples

morphogeometric method is slightly superior but comparable to our model. Given that, it could be concluded that the bone shape information (without considering the size of the centroid) is sufficient to perform an adequate estimation, although it cannot be ruled out that the introduction of information about the size of the centroid could contribute to improving the results of DL and ML methods.

That said, accuracy is not the only thing that is important in sex estimation. On the one hand, our model, which gives precise, replicable and observer-independent metrics, is more objective than the visual method proposed in [8] and used by the human expert. On the other hand, the developed method is less time consuming than the morphogeometric method proposed in [10] that requires the location of landmarks over the bone. This improvement in time efficiency is greater when the centroid size is obtained, since it has to be measured from those landmarks. In our case, once the network has been trained, it takes less than a second to estimate sex.

Once the comparison is done, we use the Grad-CAM [34] algorithm to obtain heat maps as the ones shown in Fig. 5. These visualizations allow us to verify that some of the regions that have more weight in the estimation are the ones proposed by the FA literature [8]. More precisely, the model observes the olecranon fossa and the trochlea in its estimations. The fact that the neural network has been able to identify sexual dimorphisms in these regions without previous information is not only a guarantee that the model learns what it should, but also an objective validation of the dimorphic traits proposed in FA and a demonstration of the ability of the model to replicate human knowledge. In addition, the visualizations show that the neural network has detected other possible dimorphic traits of the humerus not yet considered, so it does not only replicate knowledge, but also generates it. In this case, we have detected a new region of interest in the humerus shaft that could be relevant in order to achieve better results than the expert. More precisely, we think it is the width of the yellow region in Fig. 5c what could be an important trait for sex estimation. That being said, further anthropological studies are needed to corroborate these new hypotheses.

The superiority with respect to the human expert could be due to the ability of the model to perform the estimation with the combination of various sexual dimorphisms, whereas the human expert decided to focus only on the area of the olecranon fossa (see Fig. 6). This is because, as has happened to other authors [9], observing it exclusively is what gave him the best results. The ability of the model to use the new detected region could also be having an impact in its good results.

5 Conclusions and future works

In this paper, we have addressed sex estimation from humerus bone images using DL techniques. This is a highly complex and useful task in FA, since it contributes to forensic human identification from skeletal remains. For this purpose, we have compared DL and classical computer vision techniques. Our best model obtains better results than a human expert applying the visual method proposed in [8]. It also outperforms the morphogeometric method proposed in [10] when the centroid size is not considered and has comparable results to it when considering it. The visualization of activation maps allows us to confirm that the model observes the regions proposed by [8], as well as a new region that has not been considered before.

Two main conclusions arise from the results of this work. Firstly, it has been shown, considering the criterion that a sex estimation method must exceed 80% of correctly classified cases to be considered usable [39], that the humerus bone, and more specifically images of its posterior view, allows us to obtain an effective automatic method for sex estimation. Secondly, it has been proven that the application of AI techniques allows to replicate and even improve the results that human experts are able to obtain by visual methods of sex estimation. In this sense, while the human expert is able to correctly classify 83.33% of the individuals, the best developed model achieves an accuracy of 91.03%. The morphogeometric method proposed in [10] reaches only 75.19% accuracy when not considering the centroid size, which is much lower than ours. However, this accuracy increases to a 92.60% when considering the centroid size.

It should be noted that the proposed model obtains objective error metrics that do not depend on the observer analyzing the bones, which is the case with visual FA methods, and that it is fast and easy to apply, which are the downsides of the morphogeometric method [10]. Thus, we have developed an automatic method of sex estimation that is not only useful and precise, but also cheap, fast and objective at the same time. This objectivity, as well as the ability to provide error metrics, is especially relevant in the medico-legal field.

The study, in the field of FA, of the proposed dimorphic region; the search or validation of sexual dimorphisms in other bones by using algorithms such as Grad-CAM; and the validation of the model in populations that are not of Mediterranean origin, constitute our main lines of future work. Other research may focus on improving results, either by adding the centroid size as input to the developed models or through experimentation with other ConvNet architectures or feature extractors. This improvement in results does not only include increasing the accuracy of the model, but also making it usable for its application to images that have not been obtained using the acquisition protocol described in Sect. 3. In this regard, our model is the first existing prototype for automatic sex estimation from the humerus bone, but we cannot assure that it would work with images obtained under different conditions. Because of that, we ought to keep training the model incrementally with new images that are not acquired using the protocol that we have described.

For its good results, the method proposed in this paper will be included in the biological profile estimation toolbox of Skeleton-ID,[2] the only commercial solution for AI-driven forensic identification when using DNA or fingerprint analysis is not feasible.

Acknowledgements This work was supported by European Regional Development Funds (ERDF) under grant FORAGE (B-TIC-456-UGR20).

Funding Funding for open access charge: Universidad de Granada/CBUA.

Data availability The datasets generated during and/or analyzed during the current study are not publicly available due to privacy issues, but can be available from the corresponding author on reasonable request.

Declarations

Conflicts of interest The authors have no conflicts of interest that are relevant to the work reported in this paper.

References

1. Ubelaker DH (2008) Forensic anthropology: methodology and diversity of applications. The bio-logical anthropology of the human skeleton, pp 41–71
2. Mesejo P, Martos R, Ibáñez O, Novo J, Ortega M (2020) A survey on artificial intelligence techniques for biomedical image analysis in skeleton-based forensic human identification. Appl Sci 10:4703
3. Hayashizaki Y, Usui A, Hosokai Y, Sakai J, Funayama M (2015) Sex determination of the pelvis using Fourier analysis of postmortem CT images. Forensic Sci Int 246:122-e1
4. Raghavendra Babu YP, Kanchan T, Attiku Y, Dixit PN, Kotian MS (2012) Sex estimation from foramen magnum dimensions in an Indian population. J Forensic Leg Med 19(3):162–167
5. Bruzek J, Murail P (2006) Methodology and reliability of sex determination from the skeleton. In: Forensic Anthropology and Medicine. Springer, pp 225–242
6. Viciano J, López-Lázaro S, Alemán I (2013) Sex estimation based on deciduous and permanent dentition in a contemporary Spanish population. Am J Phys Anthropol 152(1):31–43

[2] Skeleton-ID: https://skeleton-id.com/.

7. Alemán Aguilera I, Botella López MC, du Souich Henrici P (1997) Aplicación de las funciones discriminantes en la determinación del sexo. Estudios de Antropología Biológica 9
8. Rogers TL (1999) A visual method of determining the sex of skeletal remains using the distal humerus. J Forensic Sci 44(1):57–60
9. Falys C, Schutkowski H, Weston D (2005) The distal humerus-a blind test of Rogers' sexing technique using a documented skeletal collection. J Forensic Sci 50:1289–1293
10. López-Lázaro S, Pérez-Fernández A, Alemán I, Viciano J (2020) Sex estimation of the humerus: a geometric morphometric analysis in an adult sample. Leg Med 47:101773
11. Bewes J, Low A, Morphett A, Pate FD, Henneberg M (2019) Artificial intelligence for sex determination of skeletal remains: application of a deep learning artificial neural network to human skulls. J Forensic Leg Med 62:40–43
12. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D et al (2015) Going deeper with convolutions. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1–9
13. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 248–255
14. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778
15. Vila-Blanco N, Carreira MJ, Varas-Quintana P, Balsa-Castro C, Tomás I (2020) Deep neural networks for chronological age estimation from OPG images. IEEE Trans Med Imaging 39(7):2374–2384
16. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
17. Cao Y, Ma Y, Vieira DN, Guo Y, Wang Y, Deng K et al (2021) A potential method for sex estimation of human skeletons using deep learning and three-dimensional surface scanning. Int J Legal Med 135(6):2409–2421
18. Cao Y, Ma Y, Yang X, Xiong J, Wang Y, Zhang J et al (2022) Use of deep learning in forensic sex estimation of virtual pelvic models from the Han population. Forensic Sci Res 02:1–10
19. Ortega R, Irurita J, Estévez Campo E, Mesejo P (2021) Analysis of the performance of machine learning and deep learning methods for sex estimation of infant individuals from the analysis of 2D images of the ilium. Int J Legal Med 07:135
20. Kaloi MA, He K (2018) Child Gender Determination with Convolutional Neural Networks on Hand Radio-Graphs. arXiv preprint arXiv:1811.05180
21. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958
22. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
23. d'Oliveira Coelho J, Curate F (2019) CADOES: An interactive machine-learning approach for sex estimation with the pelvis. Forensic Sci Int 07(302):109873
24. Ammer S, d'Oliveira Coelho J, Cunha EM (2019) Outline shape analysis on the trochlear constriction and Olecranon fossa of the humerus: insights for sex estimation and a new computational tool. J Forensic Sci 64(6):1788–1795
25. Pinto SCD, Urbanová P, Cesar-Jr RM (2016) Two-dimensional wavelet analysis of supraorbital margins of the human skull for characterizing sexual dimorphism. IEEE Trans Inf Forensics Secur 11(7):1542–1548
26. Darmawan MF, Yusuf SM, Rozi MA, Haron H (2015) Hybrid PSO-ANN for sex estimation based on length of left hand bone. IEEE student conference on research and development (SCOReD), pp 478–483
27. Imaizumi K, Bermejo E, Taniguchi K, Ogawa Y, Nagata T, Kaga K et al (2020) Development of a sex estimation method for skulls using machine learning on three-dimensional shapes of skulls and skull parts. Forensic Imaging 22:200393
28. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. IEEE Conf Comput Vis Pattern Recognit (CVPR) 1:886–893
29. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
30. Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition (ICDAR). IEEE 1:278–282
31. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vision 60(2):91–110
32. Bay H, Tuytelaars T, Gool LV (2006) Surf: Speeded up robust features. In: European Conference on Computer Vision (ECCV). Springer, pp 404–417
33. Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: an efficient alternative to SIFT or SURF. In: International conference on computer vision (ICCV). IEEE, pp 2564–2571
34. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: IEEE international conference on computer vision (ICCV), pp 618–626
35. Lathuilière S, Mesejo P, Alameda-Pineda X, Horaud R (2019) A comprehensive analysis of deep regression. IEEE Trans Pattern Anal Mach Intell 42(9):2065–2081
36. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C (2018) A survey on deep transfer learning. In: International conference on artificial neural networks (ICANN). Springer, pp 270–279
37. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A et al (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fus 58:82–115
38. Ras G, Xie N, van Gerven M, Doran D (2022) Explainable deep learning: a field guide for the uninitiated. J Artif Intell Res 73:329–397
39. Basic Z, Kružić I, Jerković I, Andelinović D, Andelinović S (2017) Sex estimation standards for medieval and contemporary Croats. Croat Med J 58:222–230

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

Authors and Affiliations

Javier Venema[1,4] · David Peula[2] · Javier Irurita[2] · Pablo Mesejo[1,3,4]

✉ Javier Venema
javiervenema@correo.ugr.es

David Peula
davidpeula@correo.ugr.es

Javier Irurita
javieri@ugr.es

Pablo Mesejo
pmesejo@decsai.ugr.es

[1] Department of Computer Science and Artificial Intelligence, University of Granada, C. Periodista Daniel Saucedo Aranda, 18071 Granada, Spain

[2] Department of Legal Medicine, Toxicology and Physical Anthropology, University of Granada, Parque Tecnológico de la Salud, Av. de la Investigación 11, 18006 Granada, Spain

[3] Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI Institute), Granada, Spain

[4] Panacea Cooperative Research, S.Coop, Ponferrada, Spain