

Original Article

Scale-free sex estimation from the bony pelvis using deep learning: A proof of concept**Yasin Etli¹, Erhan Kartal¹, Mahmut Asirdizer²**¹Van Yüzüncü Yıl University, Faculty of Medicine, Department of Forensic Medicine, Van, Türkiye²Bahçeşehir University, Faculty of Medicine, Department of Forensic Medicine, İstanbul, Türkiye

Received February 12, 2026; Accepted March 30, 2026; Available online August 15, 2026

Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

**Abstract**

Aim: Sex estimation from the pelvis is a cornerstone of forensic anthropological identification, but the established methods require physical access to the bone together with either metric measurement or trained morphological scoring. We tested, as a proof of concept, whether sex can be estimated automatically from a single, unscaled image of the whole bony pelvis, as a step toward a scale-free, photograph-based field tool.

Materials and Methods: From computed tomography (CT) examinations of 400 adults (200 female, 200 male; mean age 56.9 years), the pelvis was isolated and rendered as two-dimensional projections from five standard orientations (anterior, posterior, superior, inferior and lateral), yielding 2,000 images. After segmenting the bone and discarding absolute scale, an EfficientNet-B0 convolutional neural network was trained by transfer learning with strict subject-level partitioning, and evaluated by subject-grouped five-fold cross-validation and on an internal holdout split. Calibrated posteriors were converted to likelihood ratios (LRs); saliency mapping and a blur-based confound control identified what drove classification.

Results: Under cross-validation, sex was classified with an accuracy of 0.934 (95% confidence interval [CI] 0.922 – 0.946; area under the receiver-operating-characteristic curve [AUC] 0.982) for a single image and 0.993 (AUC 0.9998) when the five views were combined; the internal holdout split gave 0.865 and 0.975. All five orientations exceeded an AUC of 0.97 (range 0.975 – 0.989), the anterior view being modestly but significantly better than the posterior (Δ AUC 0.014; $p = .02$). Combining any two views already raised specimen accuracy to a mean of 0.931. The image-level log-likelihood-ratio cost was 0.420 (calibration-minimised value 0.341). Discriminability was preserved under a blur-based confound control (specimen AUC ≥ 0.996), and saliency was concentrated on bone at 1.77 times its area share.

Conclusions: Scale-free, image-based pelvic sex estimation is feasible; validation on real photographs and across populations, and a calibrated multi-view fusion rule, are required before forensic use.

Keywords: Forensic anthropology; sex determination by skeleton; pelvis; deep learning; neural networks, computer

INTRODUCTION

Estimation of biological sex is one of the first and most consequential steps in reconstructing a biological profile from unidentified skeletal remains, because subsequent inferences about age, stature and ancestry are themselves sex-dependent [1]. Among all skeletal elements, the adult os coxae is widely regarded as the most sexually dimorphic and the most reliable region for sex estimation, reflecting the functional compromise between bipedal locomotion and parturition [2,3]. Classical assessment rests on a small set of well-characterised features — the ventral arc, the subpubic concavity and the medial aspect of the ischiopubic ramus described by Phenice [2], together with

the morphology of the greater sciatic notch [3], the subpubic angle and the shape of the pelvic inlet — and these traits underpin the visual and scoring methods that remain standard in casework [4,5].

Despite their diagnostic value, morphological methods are inherently observer-dependent. Trait scoring is ordinal and partly subjective, demands considerable training, and is affected by intra- and inter-observer variability as well as by population specificity [4,5]. Metric and geometric-morphometric approaches reduce this subjectivity by quantifying shape, but they typically require precise landmarking, specialised digitisers or imaging, and population-specific discriminant

CITATION

Etli Y, Kartal E, Asirdizer M. Scale-free sex estimation from the bony pelvis using deep learning: A proof of concept. NOFOR. 2026;5(2):49-60. DOI: 10.5455/NOFOR.2026.08.013



Corresponding Author: Yasin Etli, Van Yüzüncü Yıl University, Faculty of Medicine, Department of Forensic Medicine, Van, Türkiye
Email: yasinetli@yyu.edu.tr

functions, which constrains their use in routine, resource-limited or field settings.

The convergence of virtual anthropology and deep learning has opened an alternative route. Three-dimensional models reconstructed from computed tomography (CT) reproduce osseous morphology with a precision adequate for forensic analysis [6], and deep convolutional neural networks (CNNs) [7] trained on images of such models now match or exceed human experts. CNNs applied to virtual hemi-pelvic regions of a Han population classified sex with accuracies above 90% and outperformed two anthropologists, while saliency maps confirmed that the networks attended to recognised sexual traits [8]. Two-dimensional approaches using depth images of homologous pelvic models achieved correct-assignment rates above 90% through transfer learning [9], and analogous pipelines have been reported for the skull [10] and the proximal femur [11]. Transfer learning is central to these results, enabling high accuracy to be obtained from the modest sample sizes typical of forensic collections.

Most of these systems, however, presuppose access to a CT scanner, projectional radiography or a calibrated structured-light scanner, and many require manual cropping of specific sub-regions. Even purpose-built portable solutions retain this dependence: Cao et al. integrated a handheld three-dimensional surface scanner with CNNs and reached 97.1% accuracy in a single-blind trial, yet the workflow still relies on dedicated scanning hardware and region-of-interest extraction [12]. In recovery scenes, mass-disaster contexts and archaeological excavations, an examiner often needs a rapid, low-cost triage estimate before specimens reach a laboratory. A method operating on an ordinary, unscaled photograph — for instance a single image captured with a smartphone, without a scale bar, calibration object or strictly controlled viewpoint — would substantially lower this barrier, but to our knowledge no such scale-free, photograph-based pipeline for pelvic sex estimation has yet been established.

The present study is a proof of concept toward that goal. Using CT-derived volume renderings of the bony pelvis as a realistic, near-photographic surrogate for field images, we trained transfer-learning CNNs to estimate sex from single, unscaled views captured in five standard orientations (anterior, posterior, superior, inferior and lateral). Because absolute scale is deliberately discarded, the models can rely only on shape. We additionally address several issues that are critical for forensic validity but frequently neglected: background and framing cues are removed by segmentation so that the network attends to bone rather than to acquisition artefacts; this is verified quantitatively with gradient-based saliency analysis; performance is evaluated with subject-level cross-validation and cluster-aware confidence intervals (CIs); the output is calibrated into likelihood ratios (LRs) suitable for evaluative reporting; and the robustness of the signal to progressive image degradation is examined as a control against texture- or batch-related confounding. Our objective is to determine whether automated sex estimation from a single, rapidly acquired,

unscaled photograph is feasible, as a preliminary step toward a practical mobile application for field use.

MATERIALS AND METHODS

Study Design and Sample

This was a retrospective, cross-sectional methodological study performed on abdomino-pelvic CT examinations of living adults who had been scanned for unrelated clinical indications and whose sex was recorded in the hospital information system. Ethical approval was obtained from the Van Yüzüncü Yıl University Non-Interventional Clinical Research Ethics Committee (Decision No. 2025/11-03, dated 14 November 2025), and the study was conducted in accordance with the principles of the Declaration of Helsinki. Because the analysis used anonymised, pre-existing images, the requirement for individual informed consent was waived.

Examinations with congenital or acquired pelvic pathology, fracture, surgical hardware, marked degenerative deformity, or incomplete coverage of the bony pelvis were excluded. A total of 435 individuals satisfied the inclusion criteria (223 female and 212 male). The sex recorded in the hospital information system served as the reference label throughout. All individuals were skeletally mature adults; the youngest was 21 years old.

CT Image Acquisition and Three-Dimensional Reconstruction

CT image acquisition

Abdominopelvic CT examinations were performed using a 16-slice multidetector CT scanner (Somatom Sensation 16; Siemens Medical Solutions, Erlangen, Germany). Scanning parameters were as follows: tube voltage 120 kV, effective tube current 120 mAs, gantry rotation time 420 ms, physical detector collimation 16×0.6 mm, section thickness 0.75 mm, final section collimation 32×0.63 mm, feed/rotation 6 mm, reconstruction kernel U90u, increment 0.5 mm, and matrix resolution 512×512 pixels. All images were acquired with the patient in the supine position.

Volumetric CT data were exported in Digital Imaging and Communications in Medicine (DICOM) format and imported into RadiAnt DICOM Viewer (Medixant, Poznań, Poland). For each individual a three-dimensional volume rendering of the skeleton was generated in a bone window, and all structures other than the bony pelvis (vertebral column, sacrum-adjacent ribs, proximal femora and soft tissue) were removed with the program's digital-knife tool, leaving the isolated os coxae and sacrum, following a previously described segmentation procedure [13]. No metric scale, scale bar or calibration object was retained, because the intended downstream application operates on unscaled images.

Generation of Two-Dimensional Projection Images

The isolated three-dimensional pelvic model of each individual was rendered in MeshLab (Visual Computing Laboratory, ISTI-CNR, Pisa, Italy) [14]. A consistent surface appearance was obtained by applying the built-in OpenGL Shading Language (GLSL) render shader (Render → Shaders → dimple.gdp),

and two-dimensional screen projections were captured from five standardised orientations — anterior, posterior, superior, inferior and lateral. Each projection therefore reproduced the gross morphology of the pelvis as a single colour image of the bone on a uniform gradient background, providing a realistic, near-photographic surrogate for a field photograph while keeping acquisition fully reproducible. The five-view set per individual constituted the raw image material for the classification pipeline.

Dataset assembly: Completeness filtering and class balancing

Images were inventoried automatically and assigned to one of the five views by filename parsing; auxiliary greyscale renders were excluded, and view keywords were matched after Unicode normalisation to accommodate diacritics. The parent case folder defined the specimen identity, and a sex prefix was added to the identifier to guarantee that no identifier could collide across the two sex groups. Only individuals for whom all five standard views were present were retained (218 of 223 females and 201 of 212 males). Where more than one rendering existed for a given orientation of an individual, a single image was kept, so that each retained individual contributed exactly five images. To prevent the network from exploiting any class-prevalence imbalance, the complete cases were randomly down-sampled to an equal number per sex (200 female and 200 male), yielding a final balanced dataset of 400 individuals and 2,000 projection images (400 per view).

Image Preprocessing: Bone Segmentation and Scale Normalisation

Because the rendered bone appeared as golden-yellow pixels on a blue-violet gradient, the osseous foreground was isolated by colour thresholding, requiring $(R + G) - 2B > 35$ together with $R + G + B > 60$, followed by a morphological opening to remove isolated speckle. The background and all internal openings of the bone (for example the obturator foramina and the pelvic aperture) were set to black so that the network could not use background texture, framing or corner cues. The segmented bone was then cropped to its bounding box, padded symmetrically to a square while preserving the aspect ratio, with an additional margin of 6% of the longer side, and resized to 300×300 pixels. Discarding absolute pixel size in this way removes operator-dependent zoom and framing, which is appropriate for an unscaled application and constrains the model to rely on shape alone. A quality-control montage of one masked image per view was inspected before training. The same preprocessing function is intended to be applied to any new field photograph at deployment.

Network Architecture and Transfer Learning

A single, shared CNN was used for all five views, consistent with the field scenario in which a single photograph from an arbitrary orientation must be classified; under this design each image is an independent training example, which also maximises the number of examples available from a modest set of specimens. The backbone was an EfficientNet-B0 [15] initialised with ImageNet-pretrained weights [16] obtained through the PyTorch Image Models (timm) library [17], with the classification head

replaced by a two-unit (female/male) layer. Transfer learning from natural-image features is essential at this sample size [18]. Inputs were normalised with the standard ImageNet channel statistics. Training-time augmentation comprised resized random cropping, horizontal flipping (admissible given the approximate bilateral symmetry of the pelvis), rotation up to $\pm 12^\circ$, mild colour jitter and random erasing; evaluation used a deterministic resize without augmentation.

Two-Phase Training Procedure

All models were implemented in PyTorch [19] and optimised with AdamW [20] using a weight decay of 1×10^{-4} . To stabilise convergence, training proceeded in two phases. In the first phase the backbone was frozen and only the classification head was trained for two epochs at a learning rate of 1×10^{-3} ; in the second phase the entire network was unfrozen and fine-tuned for up to 35 epochs at a learning rate of 1×10^{-4} with cosine-annealing decay. The objective was a class-weighted cross-entropy loss with label smoothing (0.05), the batch size was 16, and mixed-precision computation was used throughout. The model state yielding the highest validation balanced accuracy was retained, with early stopping after ten epochs without improvement. A single random seed (42) was fixed for all stochastic operations.

Data Partitioning and Specimen-Level Fusion

To preclude information leakage, all partitioning was performed at the level of the individual so that every image of a given specimen remained within a single subset; this avoids the inflation that occurs when different views of the same pelvis are split across training and test sets. For the holdout evaluation the 400 individuals were divided at random, at the individual level, into training, validation and test subsets of 256, 64 and 80 individuals (1,280, 320 and 400 images); the resulting test subset contained 44 females and 36 males (220 and 180 images), and the absence of shared identifiers between subsets was verified programmatically. Because this holdout split was drawn from the same 400 individuals that supplied the cross-validation folds, it is referred to throughout as an internal holdout split rather than as an independent test set. Predictions were reported at two levels: the image level, reflecting the single-photograph field scenario, and the specimen level, obtained by averaging the predicted probability of the male class across the available views of an individual (late fusion). To quantify how much of the specimen-level gain was attributable to each additional view, the fusion was repeated for every subset of one to five views on the holdout test set.

Cross-Validation and Per-View Analysis

Given the sample size, the primary performance estimate was obtained by five-fold cross-validation using a stratified, group-aware scheme (StratifiedGroupKFold from scikit-learn [21]) that stratified folds by sex while keeping all images of an individual together; each fold contained 40 female and 40 male individuals. Out-of-fold predictions were pooled across folds and evaluated at both the image and specimen levels, and the same pooled predictions were stratified by view to compare the discriminative value of the five orientations.

Calibration and Likelihood-Ratio Mapping

For forensic use the model must return calibrated probabilities rather than categorical labels. Calibration was performed by temperature scaling, optimising a single temperature parameter on the validation logits [22], and assessed on the holdout test set with the expected calibration error (ECE, computed over 15 equal-width bins) and the Brier score before and after scaling, together with a reliability diagram. Calibrated posterior probabilities were converted to LR_s under equal prior odds as $LR = p / (1 - p)$, in line with the LR framework for evaluative reporting recommended for forensic science [23]. The quality of the resulting LR_s was quantified with the log-likelihood-ratio cost (*Cl_{lr}*) and its calibration-minimised value (*Cl_{lr}-min*), obtained by a pool-adjacent-violators (isotonic) transformation, and was displayed with a Tippet plot [24,25]. Both quantities were computed at the image level, on the 400 calibrated holdout test images, with a percentile bootstrap over images for the CI, and were repeated at the specimen level on the 80 fused predictions. Because temperature scaling is a monotone transformation applied per image, the mean of the five image posteriors is not itself a calibrated probability; the specimen-level figures were therefore derived from the averaged uncalibrated posteriors and are reported as a decision score whose calibration is assessed rather than assumed. The rates of misleading evidence in favour of each proposition were also recorded [26].

Explainability and Saliency Validity

To verify that classifications were driven by skeletal morphology rather than by residual rendering artefacts, gradient-weighted

class-activation mapping (Grad-CAM) was applied at the final convolutional block [27]. Beyond qualitative inspection, a quantitative validity measure partitioned each saliency map into three regions — bone, a contour band obtained by dilating the bone mask by six pixels, and the far background — and compared the fraction of saliency energy falling in each region with the fraction of image area it occupied. Concentration of energy on bone and contour above their area share, with depletion in the far background, was taken as evidence that the network attended to morphology rather than to constant background pixels.

Confound Control by Image Degradation

Because the female and male images had been stored and rendered as separate batches, a texture- or batch-specific cue could in principle confound the classifier. This was tested by progressively degrading the images with Gaussian blur of increasing standard deviation (SD; $\sigma = 0, 4, 8$ and 16 pixels), which removes fine texture while preserving coarse shape, and retraining and re-evaluating the model at each level. To keep the four conditions comparable with one another, this experiment used a shorter training budget than the main analysis (up to 18 epochs with early stopping after eight epochs without improvement, versus 35 and ten epochs respectively); the $\sigma = 0$ condition is therefore an independently retrained model rather than the main holdout model, and its specimen-level figures differ marginally from those in Table 1. Preserved discriminability under heavy blurring would indicate that the decision rests on coarse shape, whereas a collapse would indicate dependence on fine texture or batch artefacts.

Table 1. Image- and specimen-level performance from subject-grouped five-fold cross-validation (out-of-fold) and from the internal holdout test set

Evaluation / level	n	Correct	Accuracy (95% CI)	AUC (95% CI)
Cross-validation, image	2,000	1,869	0.934 (0.922 – 0.946)	0.982 (0.976 – 0.987)
Cross-validation, specimen	400	397	0.993 (0.983 – 1.000)	0.9998 (0.9992 – 1.000)
Holdout test, image	400	346	0.865 (0.820 – 0.905)	0.958 (0.938 – 0.974)
Holdout test, specimen	80	78	0.975 (0.938 – 1.000)	0.998 (0.991 – 1.000)

Image-level intervals were obtained by specimen-clustered percentile bootstrap (10,000 replicates), which resamples individuals rather than images; specimen-level intervals used an ordinary bootstrap over individuals. The holdout split was drawn from the same 400 individuals as the cross-validation folds and is therefore an internal, not an independent, validation. Abbreviations: AUC, area under the receiver-operating-characteristic curve; CI, confidence interval.

Test-Time Augmentation

To strengthen the single-image decision relevant to field use, an optional test-time augmentation was evaluated in which the predicted probability for each image was averaged over the original view, its horizontal mirror and $\pm 8^\circ$ rotations.

Model Export and Implementation

The final model was exported as a native PyTorch checkpoint and as a TorchScript module, with an additional Open Neural Network Exchange (ONNX) export to support future on-device

deployment. All computation was performed in Python with PyTorch, timm, scikit-learn and NumPy on a single NVIDIA T4 graphics processing unit (NVIDIA Corporation, Santa Clara, CA, USA).

Statistical Analysis

Performance was summarised by accuracy and the area under the receiver-operating-characteristic (ROC) curve (AUC) at both the image and specimen levels. Because the five views of an individual are not independent observations, two-sided 95% CIs for all

image-level accuracies and AUCs were obtained by specimen-clustered percentile bootstrap resampling, which resamples individuals rather than images; specimen-level intervals used an ordinary bootstrap over individuals. Ten thousand replicates were used for the headline estimates and at least four thousand for the per-view estimates. The anterior and posterior views were compared by bootstrapping the paired difference in AUC over individuals, both views being available for every individual. Confusion matrices and per-class precision, recall and F1-scores were also computed. Ages were compared between the sexes with the Mann-Whitney U test. A random seed was fixed for all stochastic operations to aid reproducibility. Analyses were performed in Python (Python Software Foundation, Wilmington, DE, USA) with the scikit-learn, NumPy and SciPy libraries; a two-sided p value below .05 was considered statistically significant.

RESULTS

Dataset Composition

Of the 435 individuals who met the inclusion criteria, all five standard projection views were available for 218 of 223 females (97.8%) and 201 of 212 males (94.8%). After random down-sampling to equalise the two classes, the final dataset comprised 400 individuals (200 female, 200 male) and 2,000 masked projection images (400 per view). Mean age was 56.9 years (SD 20.9; range 21 – 96) overall, 56.0 years (SD 20.7; range 21 – 93) in females and 57.7 years (SD 21.2; range 21 – 96) in males, with no significant difference between the sexes (Mann-Whitney U test, p = .40). For the holdout evaluation the dataset was partitioned at the individual level into training, validation and test subsets of 256, 64 and 80 individuals (1,280, 320 and 400 images); the test subset comprised 44 females and 36 males.

Classification Performance

The five-fold, subject-grouped cross-validation provided the primary performance estimate. At the image level — the single-photograph field scenario — out-of-fold accuracy was 0.934 (1,869 of 2,000 images; 95% CI 0.922 – 0.946) with an AUC of 0.982 (95% CI 0.976 – 0.987). Per-class values were balanced: precision, recall and F1-score were 0.924, 0.947 and 0.935 for females and 0.946, 0.922 and 0.934 for males. When the five views of an individual were combined by probability averaging, specimen-level accuracy reached 0.993 (397 of 400 individuals; 95% CI 0.983 – 1.000) with an AUC of 0.9998 (95% CI 0.9992 – 1.000).

On the internal holdout test set, specimen-level accuracy was 0.975 (78 of 80 individuals; 95% CI 0.938 – 1.000) with an AUC of 0.998 (95% CI 0.991 – 1.000); both errors were male individuals classified as female, so that every female individual was correctly assigned (recall 1.000) and male recall was 0.944. Image-level accuracy was lower, at 0.865 (346 of 400 images; 95% CI 0.820 – 0.905), with an AUC of 0.958 (95% CI 0.938 – 0.974). At the image level, precision, recall and F1-score were 0.840, 0.932 and 0.884 for females (n = 220 images) and 0.904, 0.783 and 0.839 for males (n = 180 images), reflecting a decision threshold that favoured the female class on this split. Two factors contribute to the gap between the cross-validated and holdout image-level estimates: the holdout model was trained on 256 individuals whereas each cross-validation model was trained on 320, and the pooled out-of-fold predictions average over five separately trained models and therefore behave as an implicit ensemble. The cross-validated estimate is accordingly the more reliable figure (Table 1). Confusion matrices and the ROC curves are shown in Figure 1.

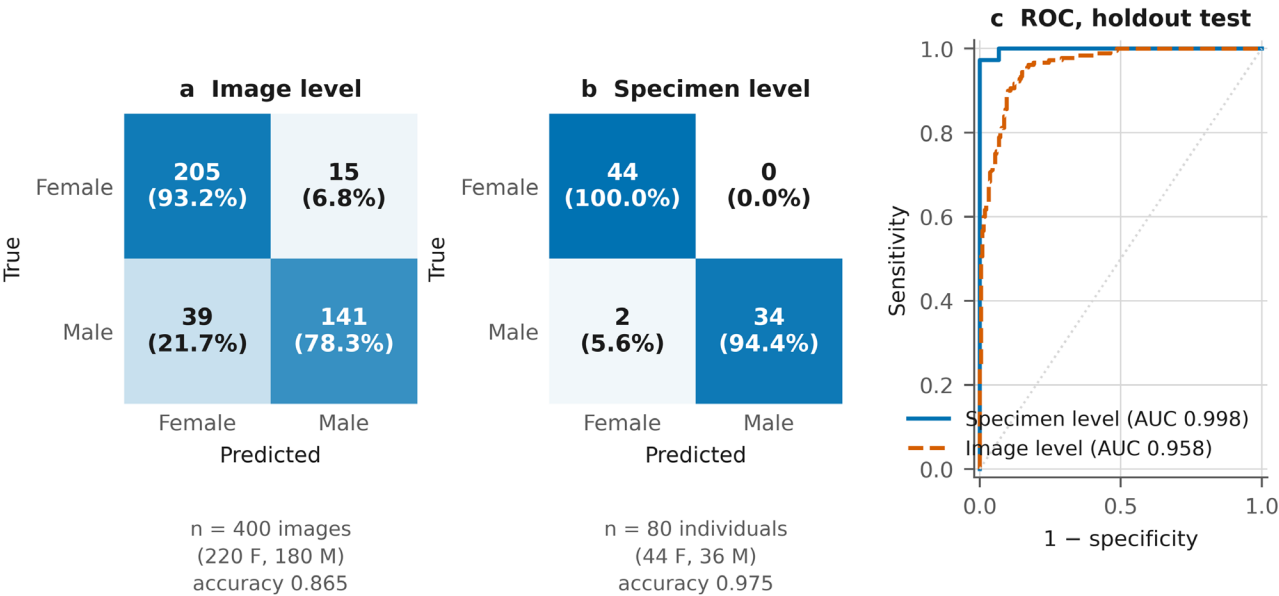


Figure 1. Classification performance on the internal holdout test set. (a) Confusion matrix at the image level (n = 400 images from 80 individuals); (b) confusion matrix at the specimen level after averaging the five views of each individual; (c) receiver-operating-characteristic curves at both levels. Cell labels give counts with the row percentage in parentheses.

Per-View Performance

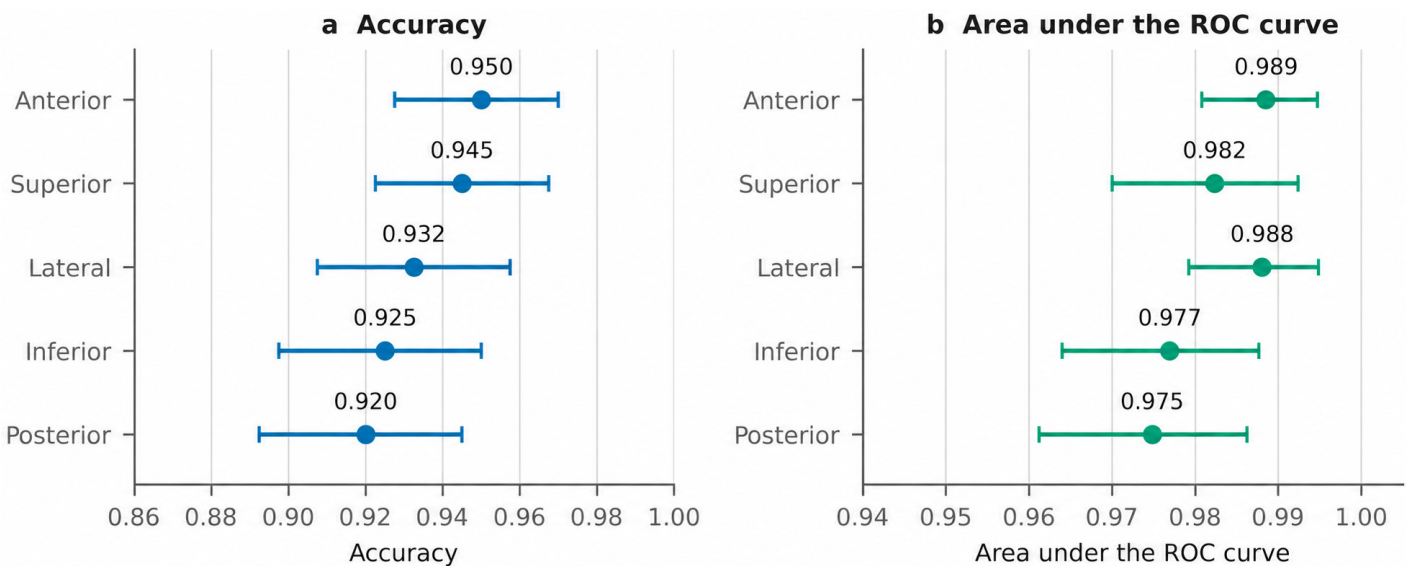
Performance was analysed separately for each of the five orientations using the pooled out-of-fold predictions (Table 2; Figure 2). Single-view accuracy ranged from 0.920 (posterior) to 0.950 (anterior) and the AUC from 0.975 (posterior) to 0.989 (anterior), so that every orientation was highly informative and the CIs overlapped substantially. A paired comparison of the two

extremes nevertheless showed a small but statistically significant advantage for the anterior view over the posterior (ΔAUC 0.014, 95% CI 0.002 – 0.027; $p = .02$). The practical implication is not that the views are indistinguishable but that the spread between the best and the worst is narrow — 1.4 percentage points of AUC — so that no single orientation is indispensable and an examiner is not obliged to reproduce one particular view.

Table 2. Per-view image-level performance from pooled out-of-fold predictions, ordered by accuracy

View	n	Accuracy (95% CI)	AUC (95% CI)
Anterior	400	0.950 (0.928 – 0.970)	0.989 (0.981 – 0.995)
Superior	400	0.945 (0.923 – 0.968)	0.982 (0.970 – 0.992)
Lateral	400	0.933 (0.908 – 0.958)	0.988 (0.979 – 0.995)
Inferior	400	0.925 (0.898 – 0.950)	0.977 (0.964 – 0.988)
Posterior	400	0.920 (0.893 – 0.945)	0.975 (0.961 – 0.986)

Intervals are percentile bootstrap (4,000 replicates) over individuals. A paired comparison of the best and worst views gave ΔAUC (anterior – posterior) 0.014 (95% CI 0.002–0.027; $p = .02$). Abbreviations: AUC, area under the receiver-operating-characteristic curve; CI, confidence interval.



Pooled out-of-fold predictions, n = 400 images per view; bars are 95% bootstrap confidence intervals.

Figure 2. Per-view image-level performance from pooled out-of-fold predictions for the five orientations: (a) accuracy and (b) area under the receiver-operating-characteristic curve. Points are the estimates and bars the 95% bootstrap confidence intervals (n = 400 images per view).

Effect of the Number of Views Combined

Because the specimen-level figures require more than one photograph, the fusion was repeated for every subset of one to five views on the holdout test set (Figure 3). Mean specimen accuracy rose from 0.865 with a single view (range across the five views 0.825 – 0.900) to 0.931 with any two views (range

0.888 – 0.975), 0.958 with three, 0.965 with four and 0.975 with all five; the corresponding mean AUCs were 0.957, 0.987, 0.993, 0.996 and 0.998. Most of the benefit of fusion was therefore obtained from the second photograph, which is relevant to a field workflow in which capturing two views of a bone costs little more than capturing one.

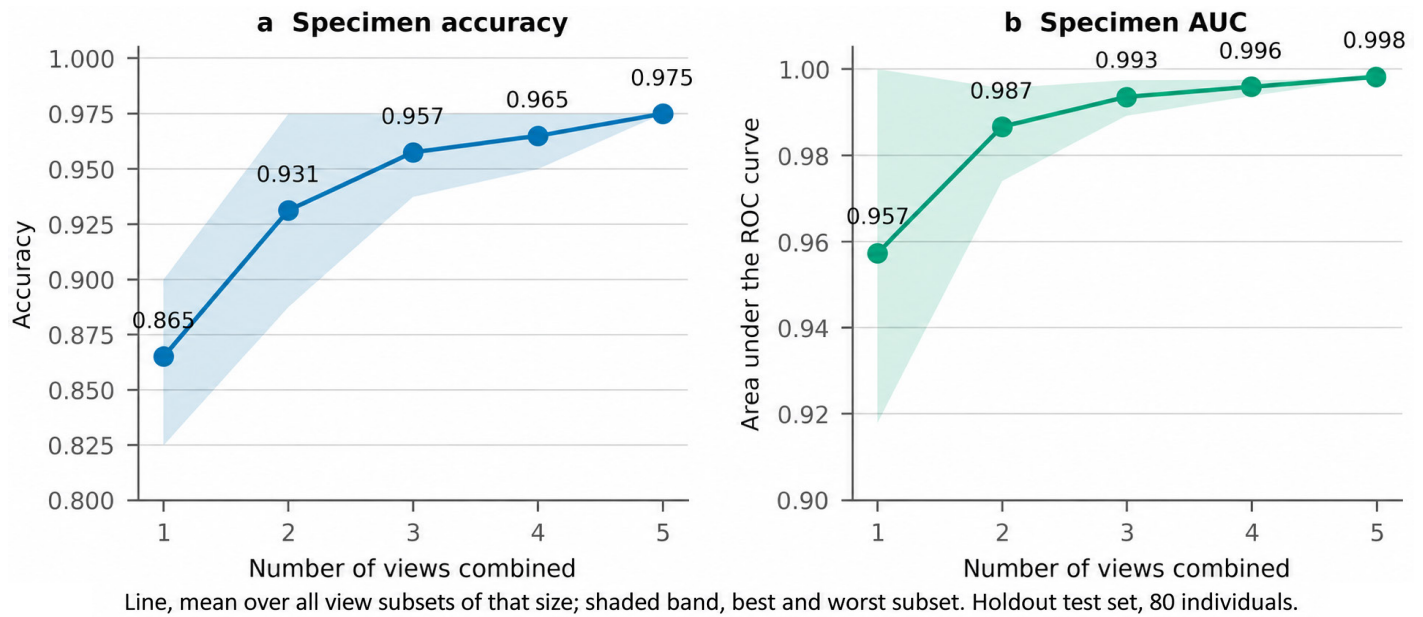


Figure 3. Effect of the number of views combined by probability averaging, holdout test set. (a) Specimen-level accuracy and (b) specimen-level area under the curve as a function of how many of the five orientations were fused. The line is the mean over all subsets of that size and the shaded band spans the best and the worst subset.

Probability Calibration and Likelihood Ratios

Temperature scaling returned an optimal temperature of 1.075, indicating that the raw outputs were already close to calibrated, and it left the test-set calibration essentially unchanged: the ECE moved from 0.0556 to 0.0565 and the Brier score from 0.0892 to 0.0886 (Figure 4). Converting the calibrated posteriors to LRs under equal prior odds yielded, at the image level, a $Cllr$ of 0.420 (95% CI 0.343 – 0.502) with a $Cllr$ -min of 0.341, so that the calibration loss component was

0.078, or 18.7% of the total. Observed LRs spanned $\log_{10}$ LR -4.92 to $+3.76$, and 11.5% of images received an LR greater than 39, confirming that the label smoothing applied during training did not impose a practical ceiling on the strength of support that could be expressed. The rates of misleading evidence were 0.217 for male images receiving an LR below one and 0.068 for female images receiving an LR above one. The Tippett plot (Figure 4) showed clear separation of the two propositions.

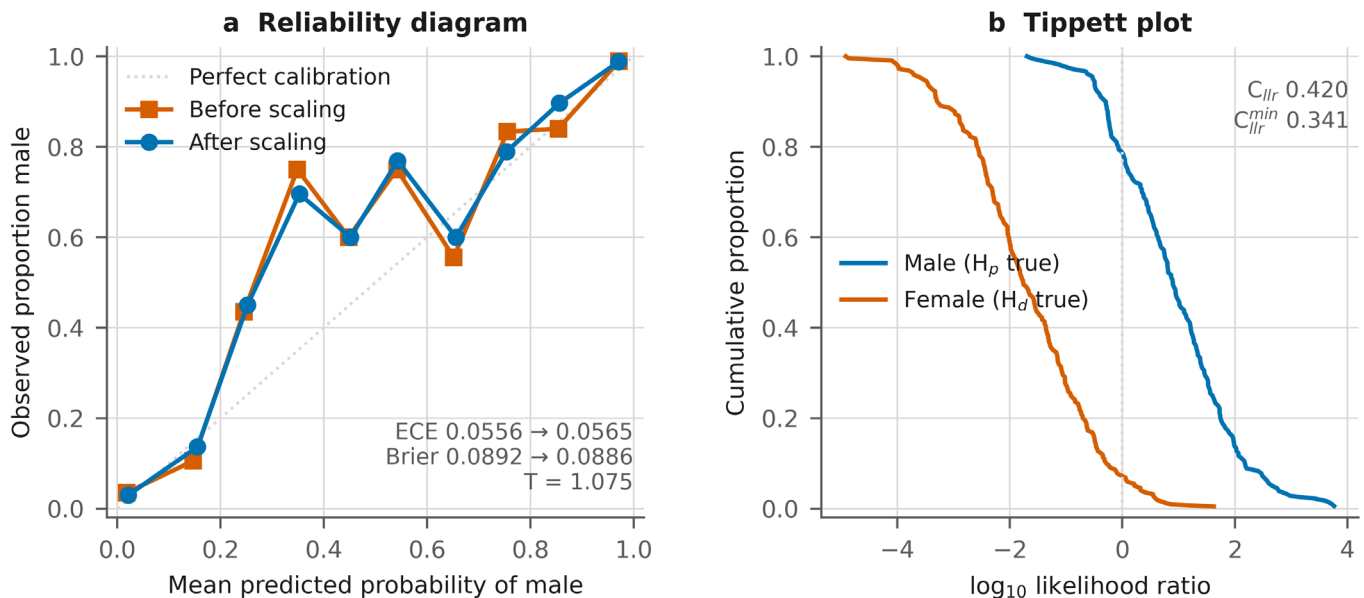


Figure 4. Probability calibration and evidential value on the holdout test set. (a) Reliability diagram before and after temperature scaling, with the expected calibration error and Brier score; (b) Tippett plot of the calibrated image-level likelihood ratios, showing for each value on the abscissa the proportion of male and of female images receiving a greater likelihood ratio.

At the specimen level the picture differed instructively. Discrimination was far better ($Cllr$ -min 0.042, corresponding to near-perfect rank separation) and the rates of misleading evidence fell to 0.056 and 0.000, but the $Cllr$ of the fused output was 0.302, so that 86% of the total cost was calibration loss. Averaging five highly correlated posteriors and mapping the mean through $p / (1 - p)$ is therefore an adequate decision rule but not a calibrated LR: the fused output compresses toward the centre, spanning only $\log_{10} LR -2.46$ to $+1.52$. A fusion rule trained and calibrated at the specimen level would be required before specimen-level LRs could be reported as evidential weight.

Saliency Validity

Grad-CAM maps concentrated on anatomically meaningful regions, including the pubic body and subpubic region on

anterior views and the greater sciatic notch and iliac regions on posterior and superior views (Figure 5). The quantitative three-region analysis gave the same ordering: bone attracted 52.9% of the saliency energy while occupying 29.8% of the image area (an energy-to-area ratio of 1.77) and the adjacent contour band 8.4% of the energy for 7.1% of the area (ratio 1.18), whereas the far background received 38.7% of the energy despite covering 63.0% of the area (ratio 0.61). Attention was therefore concentrated on the bone and depleted over the constant background, consistent with classification being driven by skeletal morphology; the residual energy over the uniformly black background reflects the limited spatial resolution of a class-activation map upsampled from the final convolutional block rather than an informative signal, since those pixels are identical in every image.

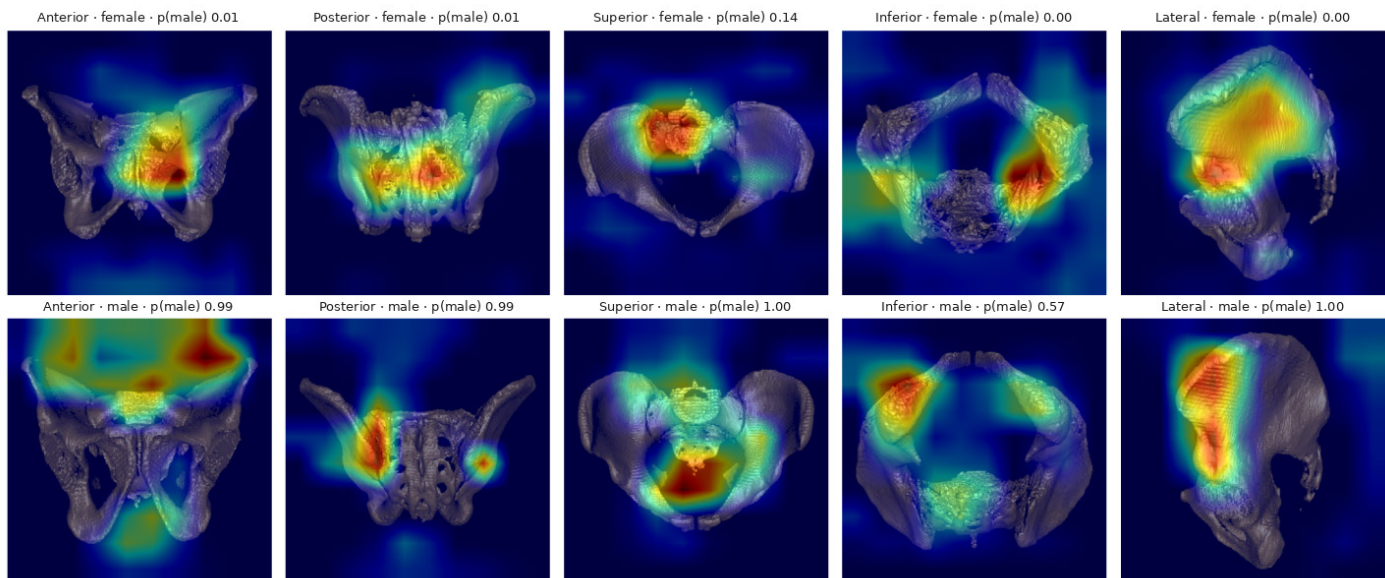


Figure 5. Representative gradient-weighted class-activation mapping (Grad-CAM) saliency maps overlaid on masked projection images: the five standard views of one female individual (top row) and of one male individual (bottom row), with the predicted probability of the male class above each panel. Warm colours indicate the regions most influential for the sex classification, which concentrate on the pubic and subpubic region, the greater sciatic notch and the iliac areas. The quantitative three-region (bone, contour, far background) energy partition reported in the text was computed across all holdout-test images.

Confound Control

Discriminability was preserved under progressive image degradation. As the Gaussian-blur SD increased from 0 to 16 pixels — removing fine texture while preserving coarse shape — the specimen-level AUC never fell below 0.9968 (lowest value, at $\sigma = 8$) and specimen-level accuracy stayed between 0.875 and 0.988 (Table 3; Figure 6). Image-level accuracy did not decline monotonically: it rose from 0.865 at $\sigma = 0$ to 0.913 at $\sigma = 4$ before falling to 0.833 at $\sigma = 8$ and 0.778 at $\sigma = 16$. Because the AUC was essentially unchanged throughout, these movements reflect shifts in the operating threshold rather than a loss of separability. Had the classifier depended on a fine-texture or batch-specific cue arising from the separate rendering of the female and male sets, heavy blurring would have degraded discriminability; the

preservation of the AUC indicates instead that the decision rests on coarse shape. It should be noted that each condition was trained from scratch, so this experiment establishes that coarse shape suffices rather than that the main model ignored texture, and that with a single seed per condition the fluctuation in image-level accuracy cannot be separated from run-to-run variability.

Test-Time Augmentation

Averaging each image's prediction over its original, horizontally mirrored and $\pm 8^\circ$ -rotated versions raised single-image (image-level) accuracy from 0.865 to 0.898 (AUC 0.967) on the test set, with specimen-level accuracy unchanged at 0.975, indicating that a simple inference-time augmentation can further stabilise the single-photograph decision relevant to field use.

Table 3. Confound control: performance after retraining and evaluating on progressively blurred images.

Blur σ (px)	Image accuracy	Specimen accuracy	Specimen AUC
0	0.865	0.988	0.999
4	0.913	0.988	1.000
8	0.833	0.950	0.997
16	0.778	0.875	1.000

Each condition was trained from scratch with a shorter budget than the main analysis (up to 18 epochs, early stopping after eight), so the $\sigma = 0$ row is an independently retrained model and its specimen-level values differ marginally from the holdout row of Table 1 (0.988 versus 0.975 accuracy; 79 versus 78 of 80 individuals correct). Discriminability was preserved as fine texture was removed, while image-level accuracy varied non-monotonically — a shift in the operating threshold rather than a loss of separability. A single seed was used per condition. Abbreviations: AUC, area under the receiver-operating-characteristic curve; px, pixels; σ , standard deviation of the Gaussian blur kernel.

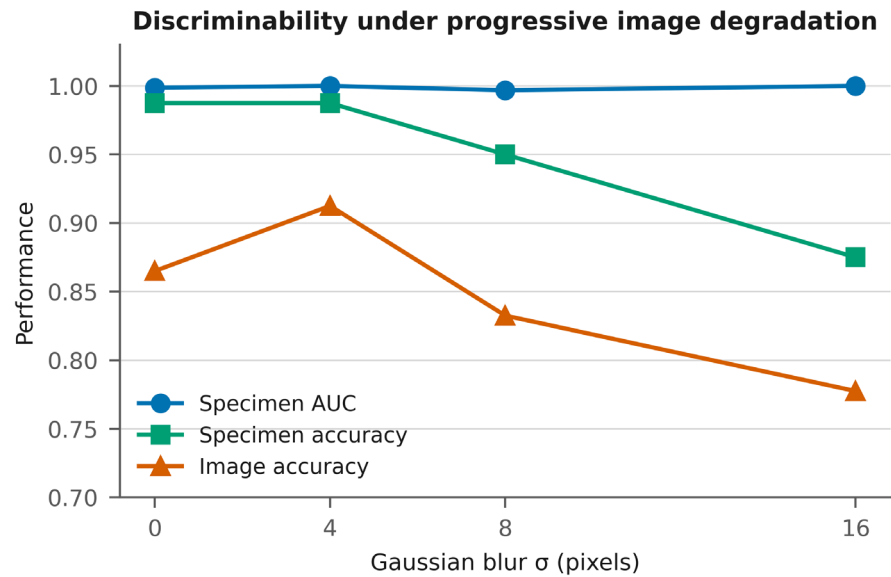


Figure 6. Classification performance (image accuracy, specimen accuracy and specimen area under the curve) as a function of Gaussian-blur strength, used as a confound control. Each blur condition was trained from scratch with an identical shortened budget.

DISCUSSION

This study set out to determine whether biological sex can be estimated automatically from a single, unscaled image of the bony pelvis, as a proof of concept toward a practical field tool. Using CNNs trained by transfer learning on scale-free projections of the whole pelvis, sex was classified with a cross-validated accuracy of 0.934 at the level of a single image and 0.993 when the five views of an individual were combined, with AUCs of 0.982 and 0.9998, respectively. The model was reasonably calibrated at the image level and, crucially, was shown to base its decisions on skeletal shape rather than on background, framing or rendering artefacts. Taken together, these findings indicate that automated, scale-free, photograph-based sex estimation from the pelvis is feasible in principle.

The accuracy obtained is comparable with that of the classical morphological methods that remain the reference standard in

casework. The pelvis is the most sexually dimorphic and the most reliable element of the skeleton for sex estimation [2,3], and the visual hip-bone methods built on the ventral arc, subpubic concavity, ischiopubic ramus and greater sciatic notch achieve accuracies of the order of 90 – 95% in experienced hands [2,4,5]. The principal weakness of these methods is their dependence on observer training and the resulting intra- and inter-observer variability [4,5]. An automated classifier reaching similar accuracy removes this subjective element. Our approach also shares the statistical philosophy of metric, probabilistic tools such as the Diagnose Sexuelle Probabiliste (DSP), a population-independent os coxae method that exceeds 95% accuracy and expresses its result as a posterior probability rather than a categorical call [28]; like that method, our network yields a calibrated probability that can be reported as a graded statement of support.

Several groups have already shown that deep learning can estimate sex from pelvic imaging. CNNs trained on virtual hemi-pelvic

regions reconstructed from CT classified sex with accuracies above 90% and outperformed experienced anthropologists, with saliency maps confirming attention to recognised sexual traits [8]; comparable results were obtained from two-dimensional depth images of homologous pelvic models, where correct-assignment rates exceeded 90% [9]; and a purpose-built portable system combining a handheld three-dimensional surface scanner with CNNs reached 97.1% accuracy in a single-blind trial [12]. Our accuracy lies within this established range. The contribution of the present work is not a higher accuracy but a lower barrier to acquisition: the cited systems all require a CT scanner, projectional radiography or a calibrated structured-light scanner, and most depend on manual cropping of specific sub-regions, whereas our pipeline operates on a single, unscaled image of the whole bone and discards absolute scale entirely. The finding that all five orientations reached an AUC above 0.97, with only 1.4 percentage points separating the best from the worst, supports a scenario in which an examiner photographs the bone from whichever of these standard orientations is convenient; this contrasts with the earlier depth-image study, in which one particular oblique frontal view was clearly superior [9]. The view-count analysis adds a practical refinement: a second

photograph recovers about 60% of the gap in accuracy between the single-image and the five-view result, and 72% of the gap in AUC, so a workable field protocol need not demand a complete standard series.

A distinctive feature of this work is its emphasis on forensic validity rather than on a categorical label alone. By temperature-scaling the output and converting it to an LR under equal prior odds, the model produces a quantity that can serve directly as a measure of evidential weight, in keeping with the LR framework for evaluative reporting that is increasingly advocated in forensic science [23]. At the image level the quality of these LRs was moderate ($Cllr$ 0.420) and not far from its $Cllr$ -min (0.341), indicating that the probabilities carry an interpretable weight of evidence rather than serving merely as classification scores [24,25]. Reporting the two levels separately also exposes a limitation that a single headline accuracy would have hidden: although fusing the five views produces the best decisions, simple probability averaging is not a calibrated combination rule, and 86.1% of the specimen-level $Cllr$ was calibration loss (Table 4). Framing automated skeletal sex estimation in LR terms is what makes this shortcoming visible at all, and it defines a concrete next step rather than undermining the approach.

Table 4. Quality of the likelihood ratios at the image and specimen levels, holdout test set

Metric	Image level (n = 400)	Specimen level (n = 80)
$Cllr$	0.420 (95% CI 0.343 – 0.502)	0.302
$Cllr$ -min	0.341	0.042
Calibration loss	0.078 (18.7% of $Cllr$)	0.260 (86.1% of $Cllr$)
$\log_{10}$ LR range	-4.92 to +3.76	-2.46 to +1.52
Rate of misleading evidence, H_p	0.217 (39/180)	0.056 (2/36)
Rate of misleading evidence, H_a	0.068 (15/220)	0.000 (0/44)

$Cllr$, log-likelihood-ratio cost; CI, confidence interval; H_p , the proposition that the individual is male; H_a , the proposition that the individual is female; LR, likelihood ratio. Rates of misleading evidence are the proportions of cases receiving an LR that supports the false proposition. Specimen-level LRs were obtained by averaging the five image-level posteriors and are not separately calibrated, which accounts for the large calibration loss at that level

The interpretability and confound analyses are a further strength. Grad-CAM localised the network’s attention to anatomically meaningful regions — the pubic body and subpubic region on anterior views and the greater sciatic notch and iliac regions on posterior and superior views — mirroring both the saliency findings of earlier pelvic networks [8] and the traits used in classical assessment [2,3]. We went beyond qualitative inspection by quantifying the distribution of saliency energy, showing that bone attracted attention at 1.77 times its area share while the constant background was attended to at 0.61 times its own, and by demonstrating that discriminability was preserved under heavy blurring that removes fine texture while sparing coarse shape. Because the female and male images had been rendered as separate batches, this blur control addresses the concern

that a network might exploit a batch-specific texture cue rather than genuine morphology. Such checks are important for the credibility of any forensic application of deep learning, where uninterpreted “black-box” performance is difficult to defend.

These strengths must be weighed against the principal constraint on generalisability. Skeletal sex estimation is notoriously population-dependent: a CNN trained on proximal-femur radiographs fell from 94.6% accuracy in the population on which it was developed to 82.9% in a different population [11], and metric methods are similarly affected except for population-independent tools such as the DSP [28]. Our model was trained on a single regional sample, and, more fundamentally, on volume renderings derived from CT rather than on real photographs of dry or skeletonised bone. The images therefore share a uniform

appearance and lighting that a field photograph would not, and the present results should be read as a feasibility demonstration rather than as evidence of field-ready performance.

Study Limitations

The most fundamental limitation concerns the nature of the images. The model was developed entirely on volume renderings derived from CT, each showing the bone in a consistent colour on a uniform synthetic background under standardised lighting. Real photographs of dry, skeletonised or in-situ remains would introduce variable illumination, shadow, soil and staining, background clutter, perspective and lens distortion, and damaged or incomplete bone — none of which is represented in the present material. The colour-based segmentation used here to isolate the bone exploits the synthetic gold-on-violet contrast and would not transfer to ordinary photographs, where a learned segmentation step would be required. The intended single-photograph field application therefore remains a hypothesis, and the present work should be regarded as a feasibility study rather than as evidence of field-ready performance.

The sample was drawn from a single institution in one region and consisted of living adults imaged for unrelated clinical reasons, with sex taken from hospital records rather than independently verified. It does not represent the dry, weathered, fragmentary or commingled remains typical of casework, and because skeletal sex estimation is strongly population-dependent [11], the accuracy reported here cannot be assumed to hold in other populations. The sample was also predominantly middle-aged and older (mean age 56.9 years), so performance in young adults, in whom some pelvic traits are less pronounced, is not well characterised, and age was not modelled as a covariate. Parity, a known modifier of pubic morphology, could not be ascertained from the data source. Balancing the classes to 200 individuals each discarded 19 complete cases, and because the completeness filter removed a larger proportion of male than of female examinations (5.2% versus 2.2%) the retained sample is not a simple random subset of those eligible.

Although the primary estimate was obtained by subject-grouped cross-validation with cluster-aware CIs, the holdout test rested on only 80 individuals drawn from the same 400-individual pool, so it is an internal rather than an independent validation, and its specimen-level accuracy carries a wide lower confidence bound. That split was also not exactly balanced by sex (44 females, 36 males), which contributes to the asymmetry between the per-class recalls observed at the image level. No independent external dataset was available for confirmation. Before any forensic use, the method would need to be validated on independent data — ideally real photographs from multiple populations — following the established guidance for the validation of LR methods in forensic evidence evaluation [26].

Several methodological choices bound the strength of the conclusions. The blur experiment retrained the network at each level and therefore shows that coarse shape is sufficient rather

than that the deployed model ignored texture; evaluating the frozen model on degraded inputs, and repeating each condition over several random seeds, would be the more direct control. A single seed was used throughout, so no estimate of run-to-run variability is available. Specimen-level LRs were obtained by averaging posteriors and are not calibrated, as the decomposition of the specimen-level *Cllr* makes explicit. Only the five canonical orientations were rendered, so genuinely oblique or off-axis viewpoints remain untested; the in-plane rotation used during augmentation is not a substitute. A single network backbone and one late-fusion rule were used, alternative architectures and uncertainty-quantification strategies were not compared, and no scale-retaining ablation or classical metric comparator such as the DSP was evaluated on the same individuals, so the cost of discarding absolute size cannot be quantified from these data. Finally, the LRs were computed under equal prior odds; in actual reporting the relevant propositions and any case-specific priors would be defined by the instructing authority in accordance with the principles of evaluative reporting [23].

CONCLUSION

This proof-of-concept study shows that a CNN trained by transfer learning can estimate biological sex from a single, unscaled projection image of the bony pelvis with high accuracy — 0.934 at the level of an individual image and 0.993 when the five available views are combined under cross-validation — without any scale, calibration object or manual region selection. All five standard orientations were informative, and a second photograph recovered most of the benefit of a full five-view series. The model returns reasonably calibrated image-level probabilities that can be converted into LRs for evaluative reporting, and saliency and confound analyses indicate that its decisions rest on skeletal morphology rather than on background or rendering artefacts.

These findings establish the feasibility of scale-free, image-based pelvic sex estimation and of embedding it within an LR reporting framework, an approach intended ultimately to support a quick, smartphone-based field tool. Before any application to casework, however, the method must be validated on real photographs and across multiple populations, with domain-adaptation techniques to bridge the gap between CT renderings and ordinary photographs, and the multi-view fusion must be trained and calibrated explicitly rather than obtained by averaging. Addressing these requirements defines the path from the present demonstration toward a deployable and forensically defensible tool.

Conflict of Interests

The authors declare that there is no conflict of interest in the study.

Financial Disclosure

The authors declare that they have received no financial support for the study.

Ethical Approval

The study was approved by the Van Yüzüncü Yıl University Non-Interventional Clinical Research Ethics Committee (Decision No. 2025/11-03, dated 14 November 2025) and was conducted in accordance with the principles of the Declaration of Helsinki.

Patient Consent

Not applicable. The study used anonymised, pre-existing imaging and the requirement for individual informed consent was waived by the ethics committee.

REFERENCES

1. Buikstra JE, Ubelaker DH, eds. Standards for Data Collection from Human Skeletal Remains. Arkansas Archeological Survey Research Series No. 44. Fayetteville: Arkansas Archeological Survey, 1994.
2. Phenice TW. A newly developed visual method of sexing the os pubis. *Am J Phys Anthropol.* 1969;30:297-301.
3. Walker PL. Greater sciatic notch morphology: sex, age, and population differences. *Am J Phys Anthropol.* 2005;127:385-91.
4. Bruzek J. A method for visual determination of sex, using the human hip bone. *Am J Phys Anthropol.* 2002;117:157-68.
5. Klaes AR, Ousley SD, Vollner JM. A revised method of sexing the human innominate using Phenice's nonmetric traits and statistical methods. *Am J Phys Anthropol.* 2012;149:104-14.
6. Colman KLA, Dobbe JGG, Stull KE, Ruijter JM, Oostra RJ, van Rijn RR, et al. The geometrical precision of virtual bone models derived from clinical computed tomography data for forensic anthropology. *Int J Legal Med.* 2017;131:1155-63.
7. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521:436-44.
8. Cao Y, Ma Y, Yang X, Xiong J, Wang Y, Zhang J, et al. Use of deep learning in forensic sex estimation of virtual pelvic models from the Han population. *Forensic Sci Res.* 2022;7:540-9.
9. Fukuta M, Kato C, Biwasaka H, Usui A, Horita T, Kanno S, et al. Sex estimation of the pelvis by deep learning of two-dimensional depth images generated from homologous models of three-dimensional computed tomography images. *Forensic Sci Int Rep.* 2020;2:100129.
10. Imaizumi K, Bermejo E, Taniguchi K, Ogawa Y, Nagata T, Kaga K, et al. Development of a sex estimation method for skulls using machine learning on three-dimensional shapes of skulls and skull parts. *Forensic Imaging.* 2020;22:200393.
11. Li Y, Niu C, Wang J, Xu Y, Dai H, Xiong T, et al. A fully automated sex estimation for proximal femur X-ray images through deep learning detection and classification. *Leg Med (Tokyo).* 2022;57:102056.
12. Cao Y, Ma Y, Vieira DN, Guo Y, Wang Y, Deng K, et al. A potential method for sex estimation of human skeletons using deep learning and three-dimensional surface scanning. *Int J Legal Med.* 2021;135:2409-21.
13. Etli Y, Demir U, Kartal E. Testing the usefulness of the generally accepted pelvic inlet classification in forensic anthropological sex determination and possible alternative methods. 3rd International and 19th National Forensic Sciences Congress, 3-6 November 2022. Didim, Türkiye, 521-33.
14. Cignoni P, Callieri M, Corsini M, Dellepiane M, Ganovelli F, Ranzuglia G. MeshLab: an open-source mesh processing tool. 6th Eurographics Italian Chapter Conference, 2-4 July 2008. Salerno, Italy, 129-36.
15. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. 36th International Conference on Machine Learning, 9-15 June 2019. Long Beach, CA, USA, 6105-14.
16. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition, 20-25 June 2009. Miami, FL, USA, 248-55.
17. Wightman R. PyTorch image models (timm). <https://github.com/huggingface/pytorch-image-models> access date [12.01.2026]
18. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? 28th Conference on Neural Information Processing Systems, 8-13 December 2014. Montreal, Canada, 3320-8.
19. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. 33rd Conference on Neural Information Processing Systems, 8-14 December 2019. Vancouver, Canada, 8024-35.
20. Loshchilov I, Hutter F. Decoupled weight decay regularization. 7th International Conference on Learning Representations, 6-9 May 2019. New Orleans, LA, USA.
21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825-30.
22. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. 34th International Conference on Machine Learning, 6-11 August 2017. Sydney, Australia, 1321-30.
23. European Network of Forensic Science Institutes. ENFSI Guideline for Evaluative Reporting in Forensic Science. Strengthening the Evaluation of Forensic Results across Europe (STEOFRAE). Dublin: ENFSI, 2015.
24. Ramos D, Gonzalez-Rodriguez J. Reliable support: measuring calibration of likelihood ratios. *Forensic Sci Int.* 2013;230:156-69.
25. Brümmer N, du Preez J. Application-independent evaluation of speaker detection. *Comput Speech Lang.* 2006;20:230-75.
26. Meuwly D, Ramos D, Haraksim R. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Sci Int.* 2017;276:142-53.
27. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. IEEE International Conference on Computer Vision, 22-29 October. 2017 Venice, Italy, 618-26.
28. Brůžek J, Santos F, Dutailly B, Murail P, Cunha E. Validation and reliability of the sex estimation of the human os coxae using freely available DSP2 software for bioarchaeology and forensic anthropology. *Am J Phys Anthropol.* 2017;164:440-9.