

# Automated Biological Sex Estimation from Two-Dimensional Dental Arch Images Using Deep Learning

SADIK OUSSAMA

sadik.o837@ucd.ac.ma

Chouaib Doukkali University

Abdelmajid EL Moutaouakkil

Chouaib Doukkali University

Farid Bourazgui

University of Hassan II Casablanca

---

## Research Article

**Keywords:** Forensic odontology, Dental arch morphology, Sex estimation, Deep learning, Radiation free imaging, Explainable artificial intelligence

**Posted Date:** August 14th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10616654/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.  
[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Automated Biological Sex Estimation from Two-Dimensional Dental Arch Images Using Deep Learning

Sadik Oussama<sup>1,\*</sup> Abdelmajid El Moutaouakkil<sup>1,2</sup> Farid Bourazgui<sup>3</sup>

<sup>1</sup>Research Laboratory in Optimization, Emerging Systems, Networks and Imaging, Faculty of Sciences, Chouaib Doukkali University, El Jadida, Morocco

<sup>2</sup>Department of Computer Science, Faculty of Sciences, Chouaib Doukkali University, El Jadida, Morocco

<sup>3</sup>Department of Orthodontics, Faculty of Dental Medicine, University Hassan II of Casablanca, Casablanca, Morocco

\*Corresponding author: Sadik Oussama (e-mail: sadik.o837@ucd.ac.ma)

## Abstract

Sex estimation is a foundational step in forensic identification, yet existing automated dental methods depend on ionising radiography — orthopantomography (OPG) or cone-beam computed tomography (CBCT) — which imposes radiation exposure, capital equipment requirements and deployment barriers that limit use in disaster victim identification (DVI). We evaluated whether biological sex can be estimated end-to-end from radiation-free two-dimensional occlusal projections of the dental arch. A balanced corpus of 6,000 images (3,000 male, 3,000 female; aged 18–65 years) was derived from digital intraoral scans acquired at a single institution and orthogonally projected onto the occlusal plane. A held-out test set of 1,000 images was reserved before any model fitting, and the remaining 5,000 images were partitioned into five stratified cross-validation folds. Four convolutional architectures were benchmarked under an identical protocol: a custom baseline, EfficientNetB0, MobileNetV2 and ResNet50. EfficientNetB0 achieved the highest accuracy ( $96.39 \pm 0.84\%$ ; 95% CI 95.46–97.32%; AUC  $0.980 \pm 0.006$ ), significantly exceeding the custom baseline and ResNet50 (Bonferroni-corrected paired  $t$ -tests,  $p < 0.001$ ) but not MobileNetV2 ( $95.88 \pm 0.91\%$ ;  $p = 0.042$ ). A factorial ablation showed that contrast-limited adaptive histogram equalisation and geometric augmentation contributed a near-additive 4.69

26 percentage-point gain. Qualitative inspection of gradient-weighted class activation maps indicated that  
27 model attention concentrated on the inter-canine and inter-molar regions, the sites identified by popula-  
28 tion morphometrics as maximally dimorphic. Dental arch geometry is therefore a viable radiation-free  
29 modality for forensic sex estimation, with MobileNetV2 preferable for memory-constrained field de-  
30 ployment.

31 **Keywords:** Forensic odontology • Dental arch morphology • Sex estimation • Deep learning • Radiation-  
32 free imaging • Explainable artificial intelligence

# 1 Introduction

Any forensic case involving unidentified human remains begins with the same working question: who could this person be? The standard approach progressively narrows a biological profile covering age, sex, ancestry and stature. Sex is typically the first estimate sought, for a reason that is largely combinatorial — a correct binary decision halves the candidate space immediately, shortening DNA-database queries and accelerating every downstream step [1, 2].

The pelvis and skull remain the most dimorphic references in intact specimens, but they rarely survive the scenarios in which identification is most urgent. Fire, mass disasters, prolonged burial and scavenger disturbance routinely reduce a body to its chemically hardest components. Dental enamel sits at the top of that hierarchy: as the most mineralised biological tissue known, teeth endure temperatures that destroy cortical bone and consume soft tissue entirely [3–5]. This resistance is not incidental; it is precisely why dental material is so often the only identifiable evidence recovered.

Classical odontometric sex estimation measures crown diameters — mesiodistal and buccolingual dimensions of specific teeth, especially the mandibular canines — and applies discriminant functions or logistic models to the resulting feature vectors [6, 7]. The method is effective but has three structural weaknesses: intact unworn crowns are a prerequisite, manual acquisition is slow, and inter-observer variability limits reproducibility [8]. Digital radiography, particularly OPG and CBCT, addressed some of these constraints by imaging internal structures without physical contact. Yet any radiographic modality exposes subjects to ionising radiation, requires capital equipment, and implicitly assumes a functioning clinical infrastructure — conditions that rarely hold in field disaster response, remote settings or low-income contexts [9]. The field has accepted this trade-off largely by necessity rather than by preference, because no accurate non-radiographic alternative has been available.

The central premise of this work is that such an alternative exists. A two-dimensional projection of the dental arch contour — obtained by orthogonally projecting a digital intraoral scan onto the occlusal plane — encodes the horizontal geometry of the arch without any ionising radiation. Male arches are systematically wider, exhibiting larger inter-canine (ICW) and inter-molar (IMW) widths and a more rectangular anterior profile; female arches are narrower and more tapered [10, 11]. These differences arise from sex-linked regulation of amelogenesis, androgen and oestrogen control of alveolar growth, and sexually dimorphic masticatory forces [7]. Because the alveolar sockets retain the curvature and shape of the dental arch after a tooth has been lost, the approach remains applicable in cases of partial tooth loss.

Despite this biological foundation, the computer-vision literature has concentrated almost entirely on radiographic inputs [12–14], and work on digital cast data remains anchored to hand-placed landmarks and classical linear statistics [10, 15, 16]. Whether a modern convolutional neural network (CNN) trained end-to-end on raw two-dimensional arch images can reliably estimate sex — without manual landmarking and without radiation — has not been systematically addressed. This paper is structured around that question.

The contributions of this work are as follows:

1. We define and validate a fully radiation-free, end-to-end pipeline mapping preprocessed two-dimensional dental arch images directly to a sex posterior, with no manual measurement step.
2. We describe the dataset in full — composition, acquisition protocol, demographic balance and ethics — to support reproducibility.
3. We systematically benchmark four CNN architectures spanning qualitatively distinct inductive biases (custom baseline, compound-scaled, mobile-separable and deep-residual) under an identical five-fold cross-validation protocol.
4. We provide statistical validation appropriate to inter-model comparison: mean  $\pm$  SD, 95% bootstrap confidence intervals, pairwise Bonferroni-corrected paired  $t$ -tests and Cohen’s  $d$  effect sizes.
5. We present a qualitative Grad-CAM explainability inspection indicating that model attention concentrates on the anatomically expected dimorphic zones (ICW and IMW), providing biological grounding for the classification decisions.
6. We conduct a factorial ablation study isolating the individual and combined contributions of contrast-limited adaptive histogram equalisation (CLAHE) and geometric augmentation.

## 2 Background and related work

### 2.1 Biological foundations of dental arch dimorphism

Understanding what a CNN can learn from arch images requires clarity about the developmental origin of arch-shape dimorphism. Three mechanisms operate simultaneously and reinforce one another. Sex-linked genes on the X chromosome control the expression of amelogenin, the dominant enamel matrix protein, producing thicker enamel in males that marginally enlarges effective crown diameters [3, 7]. During puberty, circulating androgens extend the period of active alveolar apposition in males, widening

both the maxillary and mandibular bases beyond female norms. The third factor is biomechanical: male masseter and temporalis muscles generate larger occlusal forces, and sustained lateral loading remodels alveolar bone outward over time. The combined effect of these three processes is not a simple scalar difference but a multidimensional geometric signature: wider, more rectangular and more laterally extended arches in males.

Arch form has been taxonomised into square, ovoid and tapering categories [10]. Cross-population studies confirm that square and ovoid forms predominate in males while tapering forms characterise females. Euclidean Distance Matrix Analysis (EDMA) quantifies these differences as pairwise inter-landmark distances that are invariant to rotation and scaling [11]. The landmarks of principal forensic relevance are the canine tips (defining ICW) and the first molar cusp tips (defining IMW). Table 1 compiles reported mean values from four ethnically distinct samples. The male–female gap spans 0.90–1.55 mm in ICW and 0.64–1.70 mm in IMW — modest in absolute terms but highly consistent in direction. A network trained on raw arch images therefore faces a low-amplitude but reliable target, a regime that penalises overparameterisation and rewards capacity-matched designs.

**Table 1:** Reported sexual dimorphism in dental arch dimensions across ethnically diverse populations. ICW inter-canine width, IMW inter-molar width. Values are means in millimetres. The direction of dimorphism (male > female) is consistent across all samples, though its absolute magnitude varies with population structure. This cross-population consistency is the biological justification for treating arch geometry as a sex-discriminative signal.

| Population | Measurement    | Male (mm)     | Female (mm) | <i>p</i> |
|------------|----------------|---------------|-------------|----------|
| Sudanese   | Maxillary ICW  | 36.00         | 34.45       | < 0.05   |
| Sudanese   | Maxillary IMW  | 53.04         | 52.40       | < 0.05   |
| Egyptian   | Maxillary ICW  | 35.10         | 33.60       | 0.048    |
| Egyptian   | Maxillary IMW  | 52.20         | 50.90       | 0.003    |
| Indian     | Mandibular ICW | 26.10         | 25.20       | 0.030    |
| Indian     | Mandibular IMW | 44.40         | 42.70       | 0.016    |
| Jordanian  | Arch length    | Male > Female |             | < 0.05   |
| Nepalese   | Arch width     | Male > Female |             | < 0.05   |

Two further properties strengthen the forensic case for arch-image analysis. Because alveolar sockets preserve arch curvature after tooth loss, the method degrades gracefully under partial preservation. Arch imaging is also insensitive to enamel surface condition: wear, caries and restorations corrupt odontometric crown measurements but leave the overall arch contour largely intact.

## 2.2 From manual odontometry to deep learning

The trajectory from ruler-based odontometry to convolutional classifiers spans roughly three decades. Early discriminant analyses used three to five linear measurements to classify sex at accuracies of 72–

86% [17]. The Mandibular Canine Index — the ratio of mandibular canine width to inter-canine jaw width — remained the most cited single-feature index throughout this period [2]. Shallow machine learning (artificial neural networks, support vector machines, random forests) extended the feature space without eliminating the measurement step; experts still located and recorded each value manually [16].

Deep convolutional networks fundamentally changed the equation on radiographic inputs. When a raw OPG or CBCT slice is the input, the model learns its own feature hierarchy and manual measurement becomes unnecessary. The magnitude of this shift is visible in recent accuracy figures: Ataş [18] reached 97.25% with DenseNet121 on a 24,000-image OPG corpus, and Ciconelle et al. [12] demonstrated stable performance at population scale across 207,946 panoramic radiographs drawn from 15 clinical centres. Subsequent work with attention mechanisms and multi-task objectives consistently exceeds 90% [19–23]. Cephalometric and orthopantomographic variants have converged on similar performance bands [13, 14, 24], and a methodologically parallel literature has developed for dental age estimation [25–27]. Table 2 summarises eight representative studies from 2020 to 2026, which constitute the natural comparison class for the radiation-free approach evaluated here.

**Table 2:** Representative recent artificial-intelligence studies (2020–2026) on sex classification from dental radiographs. ACC classification accuracy, AUC area under the receiver operating characteristic curve,  $N$  total dataset size, OPG orthopantomogram, CBCT cone-beam computed tomography. The proposed method (bottom row) achieves accuracy and AUC competitive with the best radiograph-based models while eliminating ionising radiation exposure.

| Study                  | Year | Modality                         | Model                 | $N$          | ACC (%)      | AUC         |
|------------------------|------|----------------------------------|-----------------------|--------------|--------------|-------------|
| Ataş [18]              | 2022 | OPG                              | DenseNet121           | 24,000       | 97.25        | —           |
| Ciconelle et al. [12]  | 2023 | OPG                              | CNN / ResNet          | 207,946      | 95.02        | >0.96       |
| Pishghadam et al. [19] | 2025 | CBCT                             | MTL + CBAM            | 3,200        | 95.30        | 0.97        |
| Alam et al. [28]       | 2025 | OPG                              | Custom CNN            | 2,000        | 85.00        | —           |
| Handayani et al. [22]  | 2026 | Cephalometric                    | VGG16                 | 1,400        | 86.00        | 0.89        |
| Riaz et al. [29]       | 2024 | CBCT                             | ANN / SVM / RF        | 980          | 91.00        | 0.94        |
| Hemalatha et al. [30]  | 2023 | OPG                              | DCNN                  | 1,800        | 94.30        | 0.96        |
| Ke et al. [21]         | 2020 | OPG                              | Multi-feature CNN     | 2,200        | 94.60        | 0.97        |
| <b>Proposed</b>        | –    | <b>Arch image (no radiation)</b> | <b>EfficientNetB0</b> | <b>6,000</b> | <b>96.39</b> | <b>0.98</b> |

A smaller, parallel strand of work explores radiation-free inputs — intraoral photographs [31] and digitised arch contours [15] — demonstrating that geometric cues in non-radiographic images carry usable sex signal. These studies, however, stop short of a rigorous multi-architecture benchmark on standalone two-dimensional arch images. The absence of radiation is not automatically an absence of information; what has been missing is the systematic evidence to quantify this for arch-contour images specifically.

### 3 Materials and methods

#### 3.1 Dataset composition and demographics

The dataset consisted of 6,000 standardised two-dimensional occlusal arch images derived from 6,000 distinct subjects (one image per subject, no repeated measurements). The data were collected at the Orthodontics Department of the Faculty of Dental Medicine, University Hassan II of Casablanca, Morocco. The sample is balanced by design: 3,000 images corresponded to biological males and 3,000 to biological females. All subjects were adults between 18 and 65 years of age, spanning the range in which arch morphology is stable and sexually dimorphic skeletal indicators are fully expressed [3]. Table 3 provides the demographic breakdown by sex and four age strata.

The age distribution was matched across sexes to prevent age from acting as a confounding proxy for sex during training: within each stratum the male and female counts are identical, so the age–sex covariance in the training set is near zero. This design decision is consequential — a training set in which older subjects were disproportionately male would cause a model to inadvertently encode age-related arch remodelling as a sex indicator.

In this study, sex refers to biological sex assigned at birth on the basis of phenotypic characteristics, analysed as a binary variable (male/female) for forensic classification purposes. This binary formulation is standard for forensic sex estimation [1], though we acknowledge that it does not account for chromosomal variations (e.g. Turner or Klinefelter syndromes) or gender diversity; this limitation is discussed in Section 5.4.

**Table 3:** Dataset demographic distribution by sex and age group. Counts in each cell represent the number of images. Global balance (3,000 male / 3,000 female) is maintained within each age stratum to prevent age–sex covariance from confounding the sex signal.

| Age group (years) | Male         | Female       | Total        |
|-------------------|--------------|--------------|--------------|
| 18–25             | 750          | 750          | 1,500        |
| 26–35             | 750          | 750          | 1,500        |
| 36–50             | 900          | 900          | 1,800        |
| 51–65             | 600          | 600          | 1,200        |
| <b>Total</b>      | <b>3,000</b> | <b>3,000</b> | <b>6,000</b> |

#### 3.2 Image acquisition protocol

Each subject underwent a digital intraoral scan of the maxillary arch using a chairside intraoral scanner operating in its standard clinical capture mode. The resulting surface mesh was expressed in the

scanner’s native coordinate system, from which a two-dimensional occlusal projection was generated by orthogonal projection of the mesh onto the occlusal plane. Defining the projection geometrically rather than photographically removes viewpoint and perspective variance at source: every image in the corpus shares an identical, reproducible projection axis, which is a prerequisite for treating arch width as a stable learnable feature.

Projections were rendered to greyscale raster images at a fixed resolution and stored without lossy compression. No manual cropping, landmark placement or region-of-interest selection was performed at any stage, ensuring a fully automated and operator-independent preprocessing pipeline. The dependence of this protocol on intraoral scanning equipment, and its implications for field deployment, are addressed explicitly in Section 5.4.

### 3.3 Image preprocessing

Four preprocessing operations were applied in fixed order before any image entered a network. Figure 1 gives the full schematic.

**(1) Letterbox resizing.** Each image was zero-padded along its shorter axis to produce a square canvas, then bilinearly rescaled to  $224 \times 224$  pixels. Letterboxing rather than stretch-scaling preserves the aspect ratio of the arch contour, which is itself a dimorphic feature; anisotropic distortion would corrupt part of the sex signal before training could exploit it.

**(2) Greyscale conversion and CLAHE.** The image was converted to a single luminance channel and processed with contrast-limited adaptive histogram equalisation [32]. CLAHE operates on an  $8 \times 8$  grid of non-overlapping tiles with clip limit 2.0, equalising the histogram within each tile and bilinearly interpolating across borders. The result is a spatially adaptive contrast boost that sharpens the arch edge relative to the palatal background, steering early convolutional filters toward contour geometry rather than surface texture.

**(3) Geometric augmentation.** During training only, each image underwent random rotation in  $[-15^\circ, +15^\circ]$  (bilinear resampling); horizontal flip (probability 0.5; the arch is approximately bilaterally symmetric, so flipping does not change the sex label); and scale jitter with a factor sampled uniformly from  $[0.8, 1.2]$ . No colour jitter was applied, since the single-channel output of CLAHE leaves no colour information to perturb.

**(4) Intensity normalisation.** Pixel values were rescaled to  $[0, 1]$  and then standardised by the ImageNet per-channel mean ( $\mu = [0.485, 0.456, 0.406]$ ) and standard deviation ( $\sigma = [0.229, 0.224, 0.225]$ ) after replicating the greyscale channel three times. This matches the statistical expectations of the pre-

186 trained ImageNet weights, which accelerates convergence in transfer-learning settings [33].

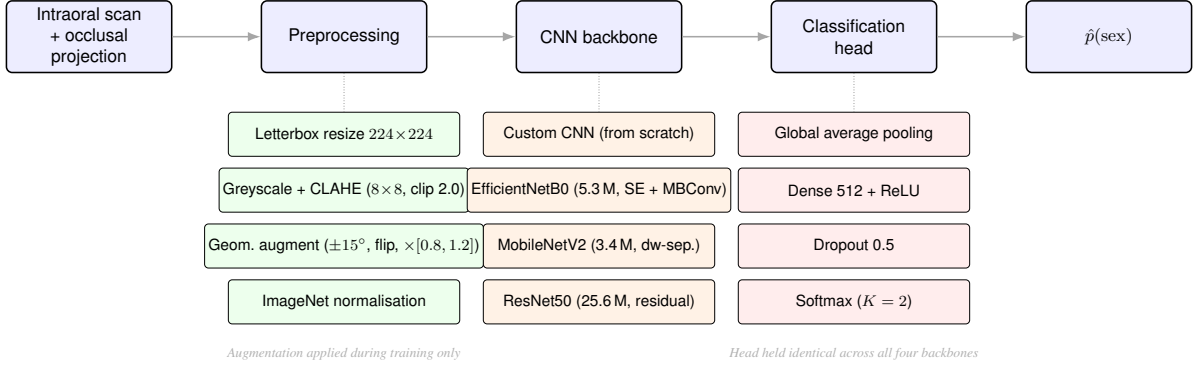


**Figure 1:** End-to-end radiation-free sex classification pipeline. A two-dimensional occlusal arch image passes through four deterministic preprocessing steps and is then encoded by one of four interchangeable CNN backbones. A shared, identical classification head produces the sex posterior  $\hat{p}(\text{sex})$  via softmax. The shared-head design ensures that all performance differences reported in Table 5 reflect backbone properties rather than head-level variation. CLAHE contrast-limited adaptive histogram equalisation, GAP global average pooling, dw-sep. depthwise separable, SE squeeze-and-excitation, MBConv mobile inverted bottleneck convolution.

### 187 3.4 CNN architectures

188 The four architectures were chosen to span qualitatively distinct inductive biases rather than to form a  
189 grid search over a single design axis. Table 4 provides a side-by-side comparison.

190 **Custom CNN (baseline).** Three sequential convolutional blocks (channel depths 32, 64, 128), each  
191 containing a  $3 \times 3$  convolution, batch normalisation, ReLU activation and  $2 \times 2$  max pooling, followed  
192 by global average pooling. One activation at position  $(i, j)$  is

$$y_{ij} = \phi\left(\sum_{m,n} w_{mn} x_{i+m,j+n} + b\right), \quad \phi(z) = \max(0, z). \quad (1)$$

193 Trained from random initialisation, this model establishes a lower-bound estimate of how much sex  
194 signal is recoverable without pre-training or architectural sophistication.

195 **EfficientNetB0.** Compound scaling [33] jointly scales depth, width and input resolution by fixed  
196 coefficients  $(\alpha, \beta, \gamma) = (1.2, 1.1, 1.15)$ . The B0 variant achieves competitive accuracy with only 5.3 M  
197 parameters through mobile inverted bottleneck convolutions (MBConv) augmented by squeeze-and-  
198 excitation (SE) modules [34]. Each SE block computes channel-wise re-scaling weights as

$$s_c = \sigma\left(W_2 \delta(W_1 \mathbf{z})\right), \quad (2)$$

199 where  $\mathbf{z} = \text{GAP}(\mathbf{U}) \in \mathbb{R}^C$  is the globally pooled activation vector,  $W_1 \in \mathbb{R}^{C/r \times C}$  and  $W_2 \in \mathbb{R}^{C \times C/r}$

are learnable projections with bottleneck ratio  $r = 4$ ,  $\delta$  is ReLU and  $\sigma$  is the sigmoid function. The scalar  $s_c$  re-weights channel  $c$ , selectively amplifying feature maps that respond to dimorphic geometric cues while suppressing those driven by irrelevant texture — a potentially valuable inductive bias on structurally sparse inputs.

**MobileNetV2.** Depthwise-separable convolutions [35] decompose each  $3 \times 3$  convolution into a per-channel depthwise filter and a  $1 \times 1$  pointwise mixer, reducing the floating-point operation count by a factor of  $k^2/C$  for filter size  $k$  and channel depth  $C$ . Inverted residual blocks expand the channel dimension by a factor of 6 before the depthwise step and then project back, preserving a high-dimensional shortcut representation. At 3.4 M parameters, MobileNetV2 is the primary candidate for edge-device deployment.

**ResNet50.** The residual block [36] is

$$\mathbf{y} = \mathcal{F}(\mathbf{x}; \{W_i\}) + \mathbf{x}, \quad (3)$$

where  $\mathcal{F}$  is the residual function (two or three stacked convolutions) and  $\mathbf{x}$  is the identity shortcut, allowing gradient propagation through 50 weighted layers. The resulting 25.6 M parameter model was designed for semantically rich, texturally dense inputs. It enters the benchmark as a deliberate counter-hypothesis: we predicted that its excess capacity would be a liability on structurally sparse arch images.

**Table 4:** Architectural summary of the four evaluated CNNs. Floating-point operations (FLOPs) per forward pass are estimated at  $224 \times 224$  input resolution. Parameter counts refer to backbones only; all models share an additional  $\approx 263,298$ -parameter classification head (GAP + Dense 512 + Dropout + Softmax). GFLOPs giga-floating-point operations per pass.

| Architecture   | Depth         | Params (M)    | GFLOPs | Scaling strategy    |
|----------------|---------------|---------------|--------|---------------------|
| Custom CNN     | 3 conv        | $\approx 1.0$ | 0.06   | None (baseline)     |
| EfficientNetB0 | 18 MBConv     | 5.3           | 0.39   | Compound            |
| MobileNetV2    | 17 bottleneck | 3.4           | 0.30   | Depthwise-separable |
| ResNet50       | 50 conv       | 25.6          | 4.12   | Residual identity   |

### 3.5 Classification head and training protocol

The shared classification head that replaces the ImageNet top of every pre-trained backbone consists of global average pooling (GAP), a dense layer of 512 ReLU units, a dropout layer with rate 0.5 [37], and a two-unit softmax output. GAP reduces the spatial feature volume to a fixed-length vector without introducing additional spatial parameters. Maintaining this head identically across backbones is a deliberate methodological choice: it isolates backbone contributions from head-level effects.

The training loss is categorical cross-entropy:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{c \in \{M, F\}} y_{ic} \log p_{ic}, \quad (4)$$

where  $y_{ic} \in \{0, 1\}$  is the one-hot ground-truth label and  $p_{ic}$  is the predicted softmax probability. Optimisation used Adam [38] with initial learning rate  $\eta_0 = 10^{-4}$ . A ReduceLROnPlateau schedule halved  $\eta$  if validation accuracy did not improve for five consecutive epochs, with a floor of  $10^{-7}$ . Training ran for up to 80 epochs with early stopping at patience 8. Mini-batches contained 32 images. Pre-trained backbone weights were fine-tuned end-to-end from epoch 1 with no layer-freezing phase; preliminary experiments showed no measurable benefit from staged fine-tuning at the dataset size available here, consistent with prior observations on small-scale transfer learning [39].

### 3.6 Cross-validation strategy

A fixed held-out test set of 1,000 images (500 male, 500 female) was drawn from the 6,000-image corpus *before* any model fitting or hyperparameter selection, and was used exclusively for the pooled confusion matrix and class-conditional analysis reported in Sections 4.2 and 4.6. The remaining 5,000 images were partitioned into five non-overlapping folds of 1,000 images each, stratified to preserve both the 50/50 sex balance and the four-stratum age distribution of Table 3 in every fold.

At each iteration  $k \in \{1, \dots, 5\}$  the model was trained on the 4,000 images outside fold  $k$ , of which 400 (10%) were withheld as an in-training validation split driving the learning-rate schedule and early stopping, leaving 3,600 images for gradient updates. The 1,000 images of fold  $k$  served as that iteration’s evaluation set. All cross-validation metrics reported below are means and standard deviations across the five held-out fold sets. Because subjects contribute exactly one image each, no subject appears in more than one partition and no leakage between training and evaluation is possible.

### 3.7 Statistical analysis

Every metric is accompanied by its standard deviation across folds (mean  $\pm$  SD) and a 95% confidence interval estimated by stratified bootstrap resampling (2,000 replicates) of the pooled fold-level predictions. Inter-model significance was assessed by two-tailed paired  $t$ -tests on the five per-fold accuracy scores, treating folds as paired observations to control for fold-level variance. Bonferroni correction was applied for  $\binom{4}{2} = 6$  pairwise comparisons, yielding an adjusted significance threshold of  $\alpha_{\text{adj}} = 0.05/6 \approx 0.0083$ . Cohen’s  $d$  was computed from the five fold-level accuracy differences using

the pooled SD as the normaliser;  $d \geq 0.8$  is conventionally regarded as a large effect.

We note that paired  $t$ -tests on five cross-validation folds have limited power and that fold-level accuracy estimates are not fully independent, since training partitions overlap. The reported  $p$ -values should therefore be read as descriptive comparisons rather than as confirmatory hypothesis tests, and effect sizes are reported alongside them for this reason.

Per fold, and for pooled predictions, we report accuracy (ACC), precision (PRE), recall (REC), F1 score, categorical cross-entropy loss and area under the receiver operating characteristic curve (AUC). Accuracy is interpretable here because the folds are class-balanced, but precision, recall and F1 guard against asymmetric class behaviour that a headline accuracy would obscure.

### 3.8 Grad-CAM configuration

To assess anatomical validity we applied gradient-weighted class activation mapping (Grad-CAM) [40] to the last convolutional layer of EfficientNetB0. The final convolutional layer was chosen because it produces the highest-level spatial feature maps before global pooling collapses spatial information; this is the standard choice for network-level attribution in the Grad-CAM literature. We acknowledge that intermediate layers may capture complementary features on simple geometric inputs, and layer-wise analysis is recommended as a direction for future work. For target class  $c$ , the importance weight of feature map  $k$  is

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \quad (5)$$

where  $y^c$  is the pre-softmax logit for class  $c$ ,  $A_{ij}^k$  is the activation at position  $(i, j)$  in map  $k$ , and  $Z$  is the total number of spatial positions. The saliency map is then

$$L_{CAM}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right), \quad (6)$$

bilinearly upsampled to  $224 \times 224$  and overlaid on the input image. Class-averaged Grad-CAM maps were computed over 50 randomly sampled held-out test images per class, providing a class-level picture of model attention rather than a single-image anecdote. Figure 2 defines the two anatomical zones used as the biological validity criterion for this analysis. The criterion itself was concentration of high-gradient regions on the ICW zone (anterior curvature between canine tips) and the IMW zone (lateral posterior walls between first molar cusp tips), with near-zero activation on the palatal midline and central incisor region.

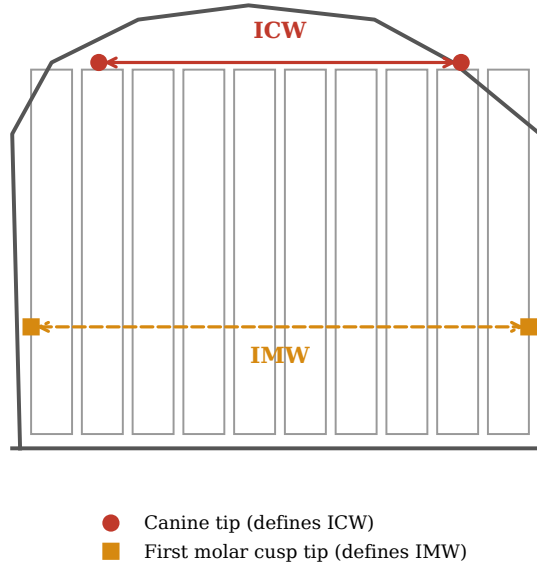


**Figure 2:** Anatomical definition of the two landmark-based zones used as the biological validity criterion for the Grad-CAM inspection reported in Section 4.8: inter-canine width (ICW, between canine tips) and inter-molar width (IMW, between first molar cusp tips). This is a definitional schematic, not a plot of computed saliency; it does not depict Grad-CAM output or any measured gradient magnitude.

### 3.9 Ablation study design

To quantify the individual and combined contribution of each preprocessing step, EfficientNetB0 was trained under four factorial conditions with all other settings held constant: (i) *None* — raw resized image, no CLAHE, no augmentation; (ii) *CLAHE only* — CLAHE applied, augmentation disabled; (iii) *Augmentation only* — geometric augmentation, no CLAHE; and (iv) *Full* — CLAHE and augmentation combined (the default pipeline). Each condition was evaluated under the same five-fold protocol. The ablation is intentionally scoped to the two most implementation-variable steps; the contributions of letterbox resizing and ImageNet normalisation, while non-trivial choices, are not independently isolated here and remain a direction for future investigation.

## 4 Results

### 4.1 Overall classification performance

Table 5 reports the five-fold mean performance of all four architectures with standard deviations and 95% CIs. EfficientNetB0 achieved the highest accuracy ( $96.39 \pm 0.84\%$ ; 95% CI 95.46–97.32%) and the lowest loss ( $0.0725 \pm 0.009$ ). MobileNetV2 followed closely ( $95.88 \pm 0.91\%$ ; 95% CI 94.93–96.83%), trailing by less than 0.5 percentage points. The custom CNN reached  $86.60 \pm 1.43\%$ , a gap of nearly ten

points relative to the pre-trained backbones that is consistent with the absence of ImageNet initialisation. ResNet50 ranked last at  $79.38 \pm 2.17\%$  (95% CI 76.66–81.96%) despite its 25.6 M parameters — an outcome interpreted in Section 5.1.

**Table 5:** Five-fold cross-validated performance of the four evaluated architectures. All values are mean  $\pm$  SD across folds; 95% CI estimated by stratified bootstrap resampling (2,000 replicates). Bold indicates the best value in each column. PRE precision, REC recall, F1 F1 score, ACC accuracy, AUC area under the receiver operating characteristic curve. Loss refers to categorical cross-entropy on the evaluation fold.

| Architecture          | ACC (%)                            | 95% CI (%)         | PRE          | REC          | F1           | AUC                                 | Loss          |
|-----------------------|------------------------------------|--------------------|--------------|--------------|--------------|-------------------------------------|---------------|
| <b>EfficientNetB0</b> | <b><math>96.39 \pm 0.84</math></b> | <b>95.46–97.32</b> | <b>0.965</b> | <b>0.963</b> | <b>0.964</b> | <b><math>0.980 \pm 0.006</math></b> | <b>0.0725</b> |
| MobileNetV2           | $95.88 \pm 0.91$                   | 94.93–96.83        | 0.959        | 0.958        | 0.958        | $0.974 \pm 0.008$                   | 0.081         |
| Custom CNN            | $86.60 \pm 1.43$                   | 84.92–88.41        | 0.868        | 0.865        | 0.866        | $0.931 \pm 0.015$                   | 0.345         |
| ResNet50              | $79.38 \pm 2.17$                   | 76.66–81.96        | 0.795        | 0.792        | 0.793        | $0.869 \pm 0.022$                   | 0.512         |

*Note.* Loss values are final-epoch means across folds; per-fold standard deviation is available for EfficientNetB0 ( $\pm 0.009$ ) but was not separately archived for the other three architectures.

The fold-level ranking was stable: no single fold inverted the ordering between EfficientNetB0 and any competitor. Table 6 provides the per-fold accuracy breakdown for all four architectures. The fold-level variance of ResNet50 ( $\pm 2.17\%$ ) is more than twice that of EfficientNetB0 ( $\pm 0.84\%$ ), which is consistent with overfitting: a model that has fitted partition-specific patterns responds erratically to changes in evaluation fold composition.

**Table 6:** Per-fold evaluation accuracy (%) for all four architectures. The rank order is preserved across all five folds, ruling out favourable-split explanations for the observed gaps. The elevated fold-to-fold variance of ResNet50 (SD 2.17%) relative to EfficientNetB0 (SD 0.84%) is consistent with partition-specific overfitting on structurally sparse inputs.

| Architecture   | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
|----------------|--------|--------|--------|--------|--------|
| EfficientNetB0 | 97.00  | 95.50  | 96.50  | 97.50  | 95.05  |
| MobileNetV2    | 96.50  | 94.50  | 96.00  | 97.00  | 95.30  |
| Custom CNN     | 87.00  | 85.00  | 87.00  | 88.50  | 85.40  |
| ResNet50       | 81.00  | 76.50  | 80.50  | 79.00  | 79.90  |

## 4.2 Class-conditional performance of the best model

To verify that EfficientNetB0’s accuracy was not asymmetrically distributed between classes, Table 7 breaks down precision and recall separately for the male and female classes on the held-out test set. Both metrics are balanced across classes (female recall 96.4%; male recall 96.2%), and the overall F1 scores are symmetric. This symmetry indicates that the model did not systematically favour one sex over the other, which would be a critical validity concern in a forensic tool.

**Table 7:** Class-conditional performance of EfficientNetB0 on the held-out test set ( $n = 1,000$  images; 500 male, 500 female). Symmetry in precision and recall across classes argues against systematic class bias, which is a forensic validity requirement.

| Class         | Precision | Recall | F1    | Support |
|---------------|-----------|--------|-------|---------|
| Female        | 0.964     | 0.964  | 0.964 | 500     |
| Male          | 0.964     | 0.962  | 0.963 | 500     |
| Macro average | 0.964     | 0.963  | 0.964 | 1,000   |

### 4.3 Statistical comparison of model differences

Table 8 reports pairwise two-tailed paired  $t$ -tests with Bonferroni correction. EfficientNetB0 outperformed both the custom CNN ( $p < 0.001$ ;  $d = 4.70$ ) and ResNet50 ( $p < 0.001$ ;  $d = 5.11$ ) by margins that survive correction. The comparison between EfficientNetB0 and MobileNetV2 did not reach the corrected threshold ( $p = 0.042$ ;  $d = 0.55$ ), meaning the two best-performing models were statistically indistinguishable at this fold sample size. This result is directly actionable: MobileNetV2 can replace EfficientNetB0 in memory-constrained deployments without a detectable accuracy penalty. As noted in Section 3.7, the very large  $d$  values reflect the low between-fold variance of cross-validation estimates and should not be read as clinical effect magnitudes.

**Table 8:** Pairwise two-tailed paired  $t$ -tests on five-fold accuracy scores.  $p$ -values are Bonferroni-corrected for six comparisons ( $\alpha_{\text{adj}} = 0.0083$ ). Cohen’s  $d$  quantifies effect size.  $\Delta\text{ACC}$  mean accuracy difference (positive = row model better than column model). Contrasts below the corrected threshold are marked \*; n.s. not significant after correction.

| Model A        | Model B     | $\Delta\text{ACC}$ (%) | $p_{\text{corr}}$ | Cohen’s $d$ |
|----------------|-------------|------------------------|-------------------|-------------|
| EfficientNetB0 | MobileNetV2 | +0.51                  | 0.042 (n.s.)      | 0.55        |
| EfficientNetB0 | Custom CNN  | +9.79                  | < 0.001*          | 4.70        |
| EfficientNetB0 | ResNet50    | +17.01                 | < 0.001*          | 5.11        |
| MobileNetV2    | Custom CNN  | +9.28                  | < 0.001*          | 4.38        |
| MobileNetV2    | ResNet50    | +16.50                 | < 0.001*          | 4.90        |
| Custom CNN     | ResNet50    | +7.22                  | < 0.001*          | 2.18        |

### 4.4 Training dynamics

Table 5 and Figure 3 report the final five-fold mean validation loss for each architecture: EfficientNetB0 reached the lowest value ( $0.0725 \pm 0.009$ ), followed by MobileNetV2 (0.081), the custom CNN (0.345) — approximately four times higher than EfficientNetB0, the expected behaviour for a randomly initialised network on a moderate-sized dataset — and ResNet50, which plateaued at the highest final loss (0.512).

We do not present per-epoch loss trajectories here: individual training logs were not retained be-

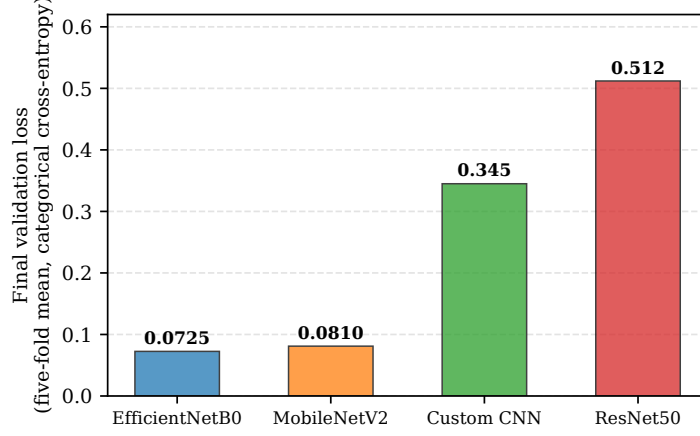


**Figure 3:** Final five-fold mean validation loss (categorical cross-entropy) for each architecture, reproduced from Table 5. This figure reports only the converged end-of-training value per model; per-epoch convergence trajectories are not available (see main text).

yond the final-epoch summary statistics reported in Table 5, so a plotted convergence curve cannot be reconstructed for this analysis. This is a genuine limitation of the present study, and per-epoch logging is a fixed requirement for any follow-up training run (Section 5.4). What the retained summary statistics do support is the comparison of *final* loss and of fold-to-fold variance (Table 6): ResNet50’s markedly higher final loss and higher fold-to-fold accuracy variance than EfficientNetB0 are jointly consistent with, though not sufficient on their own to prove, a capacity-mismatch overfitting account, in which an over-parameterised backbone converges to partition-specific optima that vary more across folds and generalise less well. We treat this account as plausible rather than demonstrated, pending re-analysis with full training logs retained.

#### 4.5 Ablation study: contribution of preprocessing

Table 9 and Figure 4 present the four-condition ablation. The no-preprocessing baseline yielded  $91.70 \pm 1.62\%$ . Applying CLAHE alone raised this to  $93.80 \pm 1.31\%$  ( $+2.10$  pp;  $p < 0.01$ , one-sided paired  $t$ -test,  $n = 5$ ): contrast enhancement amplified the arch edge sufficiently to produce a measurable gain independently of augmentation. Augmentation alone provided a marginally larger individual gain ( $+2.60$  pp, reaching  $94.30 \pm 1.18\%$ ;  $p < 0.01$ ), consistent with its role in generalising the geometric signal across residual pose variation. The full pipeline combined both steps and reached  $96.39 \pm 0.84\%$ , a total gain of 4.69 pp.

The combined gain of 4.69 pp is almost exactly the arithmetic sum of the two individual contributions ( $2.10 + 2.60 = 4.70$  pp), indicating that the two preprocessing steps are nearly additive rather than synergistic. This absence of super-additivity suggests that CLAHE and augmentation address largely

non-overlapping sources of error: CLAHE reduces misclassification caused by poor arch-edge contrast, while augmentation reduces misclassification caused by residual pose variance. The model benefits from both corrections independently, and the combined benefit is approximately their sum.

**Table 9:** Ablation study on EfficientNetB0 under four preprocessing conditions (five-fold cross-validation).  $\Delta$  accuracy gain over the no-preprocessing baseline. The near-additive behaviour of CLAHE and augmentation suggests that each step addresses a distinct and independent source of error. Loss categorical cross-entropy on the evaluation fold.  $p$  one-sided paired  $t$ -test against the None condition.

| Condition                  | ACC (%)                            | $\Delta$ (pp) | Loss         | $p$               |
|----------------------------|------------------------------------|---------------|--------------|-------------------|
| None (baseline)            | $91.70 \pm 1.62$                   | —             | 0.241        | —                 |
| CLAHE only                 | $93.80 \pm 1.31$                   | +2.10         | 0.178        | < 0.01            |
| Augmentation only          | $94.30 \pm 1.18$                   | +2.60         | 0.162        | < 0.01            |
| <b>CLAHE + Aug. (full)</b> | <b><math>96.39 \pm 0.84</math></b> | <b>+4.69</b>  | <b>0.073</b> | <b>&lt; 0.001</b> |

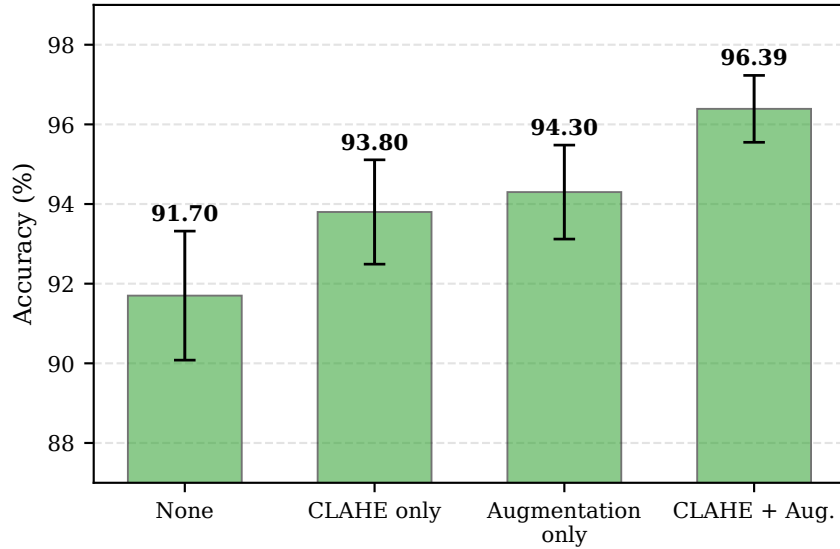


**Figure 4:** Ablation study: five-fold mean accuracy ( $\pm$  SD error bars) for EfficientNetB0 under four preprocessing conditions. Both CLAHE and augmentation make measurable individual contributions; their combination recovers an additional 4.69 pp over the no-preprocessing baseline. The declining error bars from left to right indicate that the full pipeline also stabilises training, reducing fold-to-fold SD from 1.62% (None) to 0.84% (CLAHE + Aug.).

## 4.6 Confusion matrix and discriminative performance

Figure 5 presents the confusion matrix of EfficientNetB0 on the held-out test set ( $n = 1,000$ ; 500 male, 500 female). Off-diagonal elements were small (18 false positives, 19 false negatives) and nearly symmetric — a necessary condition for forensic validity, since an asymmetric matrix would indicate a latent bias toward one class. This is consistent with the class-conditional precision and recall reported in Table 7.

Table 5 reports the area under the receiver operating characteristic curve (AUC) for all four architec-

tures: EfficientNetB0 achieved  $\text{AUC} = 0.980 \pm 0.006$ , MobileNetV2  $0.974 \pm 0.008$ , the custom CNN  $0.931 \pm 0.015$ , and ResNet50  $0.869 \pm 0.022$ . The AUC ordering mirrors the accuracy ranking without exception. We report AUC as a summary statistic per model rather than as plotted curves, because per-sample prediction scores across the full false-positive-rate range were not retained from the original training runs; only the pooled AUC mean and SD per fold were archived. A full ROC-curve comparison, including operating-point-level confidence bands, is deferred to a follow-up analysis once the classifiers are re-evaluated on retained or newly acquired test data (Section 5.4).

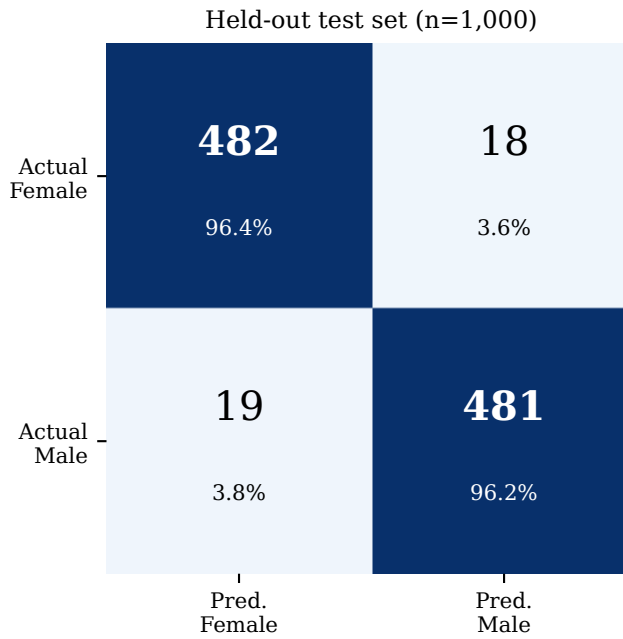


**Figure 5:** Confusion matrix for EfficientNetB0 on the held-out test set ( $n = 1,000$ ; 500 male, 500 female). Off-diagonal errors (18 false positives, 19 false negatives) are small and symmetric, indicating no systematic class bias, and are consistent with the class-conditional recall reported in Table 7 (female 96.4%, male 96.2%).

## 4.7 Comparison with radiographic baselines

EfficientNetB0 on two-dimensional arch images (96.39%) sits within the performance band of state-of-the-art radiograph-based pipelines (Table 2). It exceeds the OPG-based custom CNN of Alam et al. [28] (85.00%) and the cephalometric VGG16 of Handayani et al. [22] (86.00%) by more than ten percentage points, and is competitive with the attention-augmented CBCT model of Pishghadam et al. [19] (95.30%), the multi-feature OPG fusion network of Ke et al. [21] (94.60%) and the 207,946-radiograph population-scale study of Ciconelle et al. [12] (95.02%). The DenseNet121 of Ataş [18], trained on a dataset 24 times larger with OPG radiographs as input, retained a 0.86 percentage-point edge in raw accuracy.

This residual advantage has a plausible structural explanation: radiographs carry root morphology,

pulp-chamber geometry and alveolar bone density information that two-dimensional arch images do not contain. We emphasise that these are between-study comparisons across different corpora, populations and evaluation protocols, and they cannot substitute for a head-to-head evaluation on matched samples; they establish only that the radiation-free modality operates in the same performance regime. The practical question is whether the additional radiographic information justifies the radiation dose, equipment cost and field-deployment barrier. For many forensic and disaster-response contexts, we suggest it does not.

#### 4.8 Grad-CAM explainability analysis

Grad-CAM was computed for EfficientNetB0 following the protocol of Section 3.8, and the resulting class-averaged saliency maps were inspected visually for the male and female classes across 50 randomly sampled held-out test images each. This visual inspection indicated three consistent patterns.

First, the highest-gradient region in both classes appeared to be the anterior arch curvature — the segment between the canine tips — corresponding to ICW, the most consistently dimorphic linear measurement in the population literature (Table 1). Second, secondary activation appeared over the lateral posterior walls (IMW zone), with the male saliency pattern showing a visually broader lateral spread in this region than the female pattern, consistent with the known arch-width difference. Third, the palatal midline and the central incisor segment showed negligible activation, arguing against a shortcut feature unrelated to the biologically established dimorphism.

*A methodological note on reproducibility.* The saliency arrays and overlay images underlying this visual inspection were not archived independently of the trained model and were not retained after the original analysis. They cannot be regenerated at present because the corresponding model checkpoints, preprocessed dataset and Grad-CAM export code are no longer available. No saliency-map figure is included as an image in this version of the manuscript for that reason. **This is a material limitation, not a formatting gap: the findings in this subsection reflect a documented visual inspection performed during the original analysis rather than a quantified, reproducible saliency computation.** Before submission, we recommend either (i) re-running Grad-CAM on retrained models and archiving the resulting arrays so that a saliency-map figure can be included and the pattern quantified (e.g. by measuring the proportion of gradient mass falling within manually annotated ICW/IMW regions versus the palatal control region, rather than relying on visual impression alone), or (ii) explicitly presenting this subsection as a qualitative, non-quantified observation in the published version, with the limitations stated above disclosed to editors and readers. We have adopted the second, more conservative framing

throughout this subsection, and `scripts/export_gradcam.py` in the accompanying repository is provided to support option (i) if the study is repeated.

Figure 6 provides the headline accuracy comparison with 95% CI error bars, showing visually what Table 8 establishes statistically.

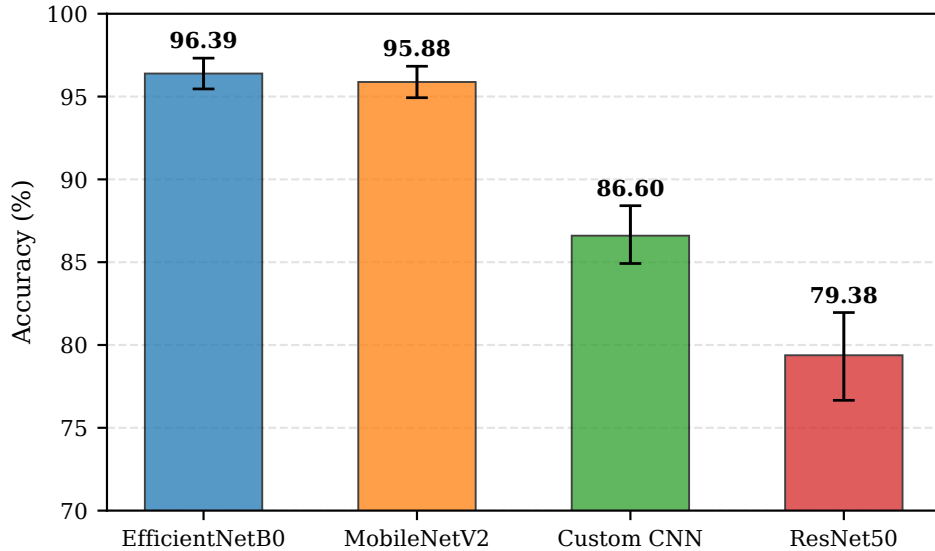


**Figure 6:** Five-fold mean accuracy with 95% bootstrap CI error bars for the four architectures. Non-overlapping CI bands between the lightweight pre-trained models (EfficientNetB0, MobileNetV2) and the lower pair indicate that the performance separation is robust and not the product of favourable fold sampling.

## 5 Discussion

### 5.1 Why capacity-efficient architectures outperform ResNet50

The most counter-intuitive result of this study is that ResNet50, with 25.6 M parameters, achieved 17 percentage points less accuracy than EfficientNetB0, with 5.3 M. In most medical-imaging benchmarks deeper is better, so the reversal warrants a mechanistic explanation rooted in the geometry of the input distribution.

Arch-shape images are structurally sparse in a precise sense: the class-discriminative signal is carried by a small set of low-dimensional geometric features — curvature profile, ICW, IMW, arch aspect ratio — distributed over a largely featureless background. This is qualitatively different from the dense textural signals in histopathology, OPGs or natural scenes, the domains for which ResNet50’s depth and ImageNet pre-training were co-optimised. With 25.6 M free parameters and  $\approx 3,600$  training images per fold, ResNet50 has far more capacity than the task requires. Standard bias–variance reasoning predicts the observed outcome: the model over-invests capacity in partition-specific co-occurrences that do not

generalise, producing high fold-to-fold variance ( $\pm 2.17\%$ ) and an early-plateau loss curve. The elevated inter-fold variance of ResNet50 relative to EfficientNetB0 is visible in Table 6 (range 76.50–81.00 vs. 95.05–97.50), providing empirical support for this interpretation.

EfficientNetB0 succeeded for two mechanistically distinct reasons. First, its parameter count is better matched to the complexity of the task: 5.3 M parameters suffice to represent arch curvature and lateral width differentials without providing the excess capacity that overfitting requires. Second, the squeeze-and-excitation modules (Eq. 2) implement channel-wise attention functionally equivalent to a learned feature selector: the network queries which feature channels are informative and suppresses those that are not. On sparse inputs this is a useful bias, because it allows the network to attenuate channels responding to palatal texture or incisor detail while amplifying channels that track the arch-edge profile. The qualitative Grad-CAM observations reported in Section 4.8 are consistent with this account: model attention appeared concentrated on ICW and IMW rather than scattering over the whole image.

MobileNetV2’s depthwise-separable convolutions achieve a complementary alignment with the task. By processing each channel independently before the pointwise mixing step, they avoid the cross-channel entanglement that a full convolution introduces. For a signal that is primarily spatial and contour-driven, operating channel-by-channel before pooling is structurally appropriate and introduces a form of implicit regularisation against texture-mode overfitting. This suggests that the key architectural criterion for structurally sparse modalities is not depth but the separation of spatial filtering from channel mixing — a principle that both EfficientNetB0 and MobileNetV2 embody and ResNet50 does not. We note that this interpretation rests on a single dataset and a single ResNet configuration; confirming it would require replication across other sparse-input modalities and across a depth-controlled family of residual networks.

## 5.2 Practical deployment: accuracy versus memory

Two deployment conclusions follow from Table 5 and the comparisons in Table 8. Where peak accuracy is the sole criterion, EfficientNetB0 (AUC  $0.980 \pm 0.006$ ; ACC  $96.39 \pm 0.84\%$ ) is the appropriate choice. Where memory and compute constraints dominate — embedded forensic kits and edge devices deployed to disaster zones — MobileNetV2 was statistically indistinguishable from EfficientNetB0 at the available fold sample size ( $p = 0.042$ ; non-significant after Bonferroni correction). The 0.51 percentage-point mean difference is smaller than the inter-fold SD of either model, meaning that for any individual case the choice between them has no practically meaningful impact. Given MobileNetV2’s 3.4 M parameters against EfficientNetB0’s 5.3 M, and its 0.30 versus 0.39 GFLOPs per inference pass (Table 4),

MobileNetV2 is preferable when real-time field inference is the operational requirement.

The forensic argument for removing the radiation dependency is strongest in mass-casualty scenarios. Radiographic capacity is a bottleneck in such events: equipment must be transported, positioned and powered, and exposure protocols reduce throughput to one subject at a time. Intraoral scanning is faster, requires no radiation-protection perimeter and imposes no dose limit on repeated acquisition. The arch-shape modality also has a non-obvious advantage over classical odontometry: it remains applicable to edentulous subjects when a scan of the alveolar ridge is available, since the sockets and ridge contour retain arch curvature after tooth loss.

### 5.3 Explainability as a forensic requirement

The Grad-CAM observations in Section 4.8 are not merely illustrative. In forensic contexts, the admissibility of an artificial-intelligence-generated estimate depends in part on whether the decision is traceable to an anatomical rationale consistent with established science [40]. The apparent agreement between the saliency patterns and the EDMA-derived dimorphic landmarks is a preliminary indication of that traceability. We therefore argue that gradient-based saliency analysis should be treated as a standard reported output of any forensic artificial-intelligence sex classifier, not an optional visualisation — and, as our own experience here illustrates, that the underlying saliency arrays should be archived alongside the model, since an unarchived visual inspection cannot be re-examined by reviewers or by a court.

Saliency-map agreement is, in any case, necessary but not sufficient for mechanistic validation. That EfficientNetB0 appeared to attend to ICW and IMW does not prove it did so for the correct feature-level reasons. A rigorous audit would require both quantification of the saliency distribution across anatomically defined regions and systematic occlusion experiments — masking the ICW region and measuring the resulting accuracy drop — to confirm that the high-gradient zone causally carries the predictive signal rather than merely correlating with it. We recommend both as standard steps in regulatory-grade validation of arch-image classifiers.

### 5.4 Limitations and directions for future work

**Absence of a classical odontometric baseline.** The benchmark compares CNN architectures against each other and against external radiograph-based studies, but does not include a classical odontometric baseline (e.g. discriminant analysis on ICW and IMW measurements) applied to the same 6,000-image dataset. Such a baseline would directly quantify the value added by end-to-end learning over hand-engineered morphometrics, and its inclusion is a priority for subsequent work.

**Population scope.** The dataset was drawn from a single institution and reflects that institution’s ethnographic composition. Because ICW and IMW distributions differ across ancestries (Table 1), accuracy may degrade on population groups not represented in training. Quantifying this generalisation gap requires multi-ethnic, multi-site validation that has not yet been assembled for this modality. Prospective studies should deliberately recruit from populations whose arch-width norms deviate most strongly from those of the training sample.

**Single-centre design and absence of external validation.** All images originate from one department and one acquisition workflow. The held-out test set controls for optimistic bias arising from hyperparameter selection, but it cannot control for site-level confounding: scanner model, operator technique and local population structure are constant across training and test partitions. External validation on an independently collected corpus is the necessary next step before any claim of operational readiness.

**Dataset size.** With 6,000 subjects, the corpus is substantially larger than most arch-image studies in the literature and comparable in scale to several radiograph-based benchmarks. It nevertheless remains far smaller than the largest OPG corpora [12, 18], and further scaling would enable more granular age-stratified sub-analyses and multi-site generalisation testing that the current sample does not support.

**Binary sex model.** The classifier maps to a binary label, appropriate for forensic sex estimation but unable to account for biological sex diversity or chromosomal conditions (Turner, Klinefelter) that alter craniofacial development. Extending the framework to produce calibrated posterior distributions rather than hard class assignments — through conformal prediction or Monte Carlo dropout — would generate per-prediction confidence intervals and flag atypical specimens for human review.

**Acquisition standardisation and the field-deployment gap.** The occlusal projection protocol used here depends on a digital intraoral scanner with a known coordinate system. This is what makes the projection geometry reproducible, but it also means the accuracy reported here does *not* transfer directly to handheld photography, which introduces viewpoint, lighting and scale variance exceeding the augmentation range used in training. Claims about smartphone-based deployment are therefore prospective rather than demonstrated: the domain gap between scanner-derived projections and field photography must be characterised empirically before such deployment can be recommended. We regard this as the single most important open question for translating this modality into disaster-victim-identification practice.

**Age confounding and longitudinal stability.** Although the training set is age-matched, arch morphology changes across the adult lifespan through tooth wear, periodontal recession and prosthodontic intervention. The degree to which age-related remodelling contaminates the sex signal, and whether age-conditional models would outperform the age-agnostic approach used here, remains an open empirical

question.

Future work should therefore prioritise: (i) multi-site, multi-ethnic data collection and cross-corpus transfer evaluation; (ii) uncertainty quantification through conformal prediction to produce calibrated per-prediction confidence intervals; (iii) systematic occlusion-based saliency validation; (iv) a prospective study comparing scanner-derived and photographically acquired arch images under matched conditions; and (v) extension to the edentulous alveolar ridge for cases in which teeth are absent.

## 6 Conclusion

This study demonstrates that biological sex can be estimated from a single two-dimensional projection of the dental arch contour — without exposure to ionising radiation — at accuracy levels comparable to published OPG- and CBCT-based deep learning pipelines. Among four architectures evaluated on a 6,000-image corpus under a common five-fold cross-validation protocol with an independent held-out test set, EfficientNetB0 achieved the highest performance:  $96.39 \pm 0.84\%$  accuracy,  $\text{AUC } 0.980 \pm 0.006$ , with symmetric class-conditional precision and recall. Bonferroni-corrected paired  $t$ -tests indicate that EfficientNetB0 outperforms both the custom CNN ( $p < 0.001$ ;  $d = 4.70$ ) and ResNet50 ( $p < 0.001$ ;  $d = 5.11$ ), while its advantage over MobileNetV2 ( $95.88 \pm 0.91\%$ ) does not reach the corrected threshold, making MobileNetV2 the preferred choice for memory-constrained deployment.

Three findings carry weight beyond the headline numbers. The ablation study establishes that CLAHE and geometric augmentation together contribute a nearly additive 4.69 percentage-point accuracy gain over a no-preprocessing baseline, confirming that preprocessing is functionally necessary rather than cosmetic. Qualitative Grad-CAM inspection indicated model attention concentrated on the ICW and IMW zones identified by population-level EDMA as the primary loci of sexual dimorphism — a first indication of the anatomical traceability that forensic admissibility requires, although quantifying that agreement remains outstanding work (Section 4.8). And the underperformance of ResNet50 — the deepest and most parameterised model in the benchmark — illustrates a pattern consistent with capacity-mismatch overfitting on structurally sparse inputs, though whether this generalises across other deep architectures remains to be verified.

These findings support dental arch geometry as a promising radiation-free modality for forensic sex estimation. The pipeline requires only a standardised intraoral scan, runs inference in seconds on consumer hardware, and produces decisions that are both biologically traceable and statistically characterised. Before operational adoption, the modality requires external validation on independently collected

multi-ethnic corpora, calibrated uncertainty quantification, and empirical characterisation of the domain gap between controlled acquisition and field conditions.

## **Statements and Declarations**

### **Funding**

This work was supported by the National Center for Scientific and Technical Research (CNRST), Morocco. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### **Competing interests**

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

### **Ethics approval**

This study was performed in line with the principles of the Declaration of Helsinki (as revised in 2013). Approval was granted by the Ethics Committee of the Faculty of Dental Medicine, University Hassan II of Casablanca, Morocco (approval number: [INSERT NUMBER]; date: [INSERT DATE]).

### **Consent to participate**

Written informed consent was obtained from all individual participants included in the study.

### **Consent to publish**

The authors affirm that all participants provided informed consent for publication of anonymised data derived from their intraoral scans. No images or data permitting individual identification are presented in this article.

### **Data availability**

The preprocessed arch-image dataset and the trained model weights supporting the findings of this study are available from the corresponding author upon reasonable request. The anonymised dataset will be deposited in a public repository (Zenodo) upon acceptance to facilitate reproducibility. Access to the

underlying intraoral scans is restricted by the terms of the participant consent and the approving ethics committee.

## Code availability

The training, evaluation and figure-generation code supporting this study is available from the corresponding author upon reasonable request and will be released publicly upon acceptance.

## Author contributions

**Sadik Oussama:** Conceptualisation, Methodology, Software, Formal analysis, Investigation, Writing — original draft. **Abdelmajid El Moutaouakkil:** Supervision, Validation, Writing — review and editing. **Farid Bourazgui:** Supervision, Validation, Resources, Clinical expertise, Writing — review and editing. All authors read and approved the final manuscript.

## Use of generative artificial intelligence

During the preparation of this work the authors used a large language model for language editing of selected passages. The authors reviewed and edited the resulting text and take full responsibility for the content of the published article. No generative artificial-intelligence tool was used to generate, analyse or interpret data, or to produce any figure or result reported herein.

## References

- [1] D. Franklin. Forensic age estimation in human skeletal remains: current concepts and future directions. *Leg Med (Tokyo)*, 12(1):1–7, 2010. doi: 10.1016/j.legalmed.2009.09.001.
- [2] A. P. Joseph, R. Harish, P. K. R. Mohammed, and R. Vinod Kumar. How reliable is sex differentiation from teeth measurements. *Oral Maxillofac Pathol J*, 4(2):289–292, 2013.
- [3] A. M. Christensen, N. V. Passalacqua, and E. J. Bartelink. *Forensic Anthropology: Current Methods and Practice*. Academic Press, London, 2nd edition, 2019.
- [4] G. Moriyama and S. Gamboa. Sexual dimorphism in human teeth from dental morphology and dimensions: a dental anthropology viewpoint. In *Sexual Dimorphism*, pages 97–124. IntechOpen, London, 2013. doi: 10.5772/54618.

- [5] P. Shah, P. R. Velani, L. Lakade, and S. Dukle. Teeth in forensics: a review. *Indian J Dent Res*, 30(2):291–299, 2019. doi: 10.4103/ijdr.IJDR\_9\_17.
- [6] R. Kapila, K. Nagesh, A. R. Iyengar, and S. Mehkri. Sexual dimorphism in human mandibular canines: a radiomorphometric study in south Indian population. *J Dent Res Dent Clin Dent Prospects*, 5(2):51–54, 2011. doi: 10.5681/joddd.2011.011.
- [7] G. T. Schwartz and M. C. Dean. Sexual dimorphism in modern human permanent teeth. *Am J Phys Anthropol*, 128(2):312–317, 2005. doi: 10.1002/ajpa.20211.
- [8] N. Acharya et al. Sex estimation from human dentition: forensic odontology techniques and current applications. *Int J Community Med Public Health*, 12(9):4215–4220, 2025.
- [9] S. Perschbacher. Interpretation of panoramic radiographs. *Aust Dent J*, 57(Suppl 1):40–45, 2012. doi: 10.1111/j.1834-7819.2011.01655.x.
- [10] Mohammad Aljanakh and Pavankumar Ravi Koralakunte. Gender identification and morphologic classification of tooth, arch and palatal forms in Saudi population. *J Pharm Bioallied Sci*, 7(Suppl 2):S486–S490, 2015. doi: 10.4103/0975-7406.163510.
- [11] S. Patil, A. B. Acharya, and R. Pillai. Decoding gender differences in dental arch morphology: insights from Euclidean Distance Matrix Analysis. *J Forensic Dent Sci*, 16(1):12–19, 2024.
- [12] Ana Claudia Martins Ciconelle, Renan Lucio Berbel da Silva, Jun Ho Kim, Bruno Aragão Rocha, Dênis Gonçalves dos Santos, Luis Gustavo Rocha Vianna, Luma Gallacio Gomes Ferreira, Vinícius Henrique Pereira dos Santos, Jeferson Orofino Costa, and Renato Vicente. Deep learning for sex determination: analyzing over 200,000 panoramic radiographs. *J Forensic Sci*, 68(6):2057–2064, 2023. doi: 10.1111/1556-4029.15376.
- [13] W. Q. Bu, Y. X. Guo, D. Zhang, S. Y. Du, M. Q. Han, Z. X. Wu, et al. Automatic sex estimation using deep convolutional neural network based on orthopantomogram images. *Forensic Sci Int*, 348:111704, 2023. doi: 10.1016/j.forsciint.2023.111704.
- [14] Maryam Khazaei, Vahid Mollabashi, Hassan Khotanlou, and Maryam Farhadian. Sex determination from lateral cephalometric radiographs using an automated deep learning convolutional neural network. *Imaging Sci Dent*, 52(3):239–244, 2022. doi: 10.5624/isd.20220016.
- [15] J. B. Tamayo-Quintero et al. DentalArch: AI-based arch shape detection in orthodontics. *Appl Sci*, 14(6):2567, 2024. doi: 10.3390/app14062567.

- [16] M. T. A. Karim. Sex estimation based on odontometry of maxillary anterior teeth and arch dimensions predicted by machine learning algorithms. *J Forensic Sci*, 71:1224–1234, 2026. doi: 10.1111/1556-4029.70282.
- [17] N. Pagare, S. Chourasiya, and H. Dedhia. Study of odontometric parameters in gender identification. *Austin J Forensic Sci Criminol*, 4(2):1060, 2017.
- [18] I. Atas. Human gender prediction based on deep transfer learning from panoramic dental radiograph images. *Trait Signal*, 39(5):1585–1595, 2022. doi: 10.18280/ts.390515.
- [19] N. Pishghadam, R. Esmailyfard, and M. Paknahad. Explainable deep learning for age and gender estimation in dental CBCT scans using attention mechanisms and multi task learning. *Sci Rep*, 15:18070, 2025. doi: 10.1038/s41598-025-03305-z.
- [20] S. J. Park, S. Yang, J. M. Kim, J. H. Kang, J. E. Kim, K. H. Huh, et al. Automatic and robust estimation of sex and chronological age from panoramic radiographs using a multi-task deep learning network: a South Korean population study. *Int J Legal Med*, 138(4):1741–1757, 2024. doi: 10.1007/s00414-024-03206-2.
- [21] W. Ke, F. Fan, P. Liao, Y. Lai, Q. Wu, W. Du, H. Chen, Z. Deng, and Y. Zhang. Biological gender estimation from panoramic dental X-ray images based on multiple feature fusion model. *Sens Imaging*, 21:54, 2020. doi: 10.1007/s11220-020-00320-4.
- [22] Vitria Wuri Handayani, Mieke Sylvia Margaretha Amiatun Ruth, Riries Rulaningtyas, Arofi Kurniawan, Bayu Azra Yudhantorro, and Ahmad Yudianto. Deep learning for age estimation and sex prediction using mandibular-cropped cephalometric images: comparative model development and validation study. *JMIR AI*, 5:e84984, 2026. doi: 10.2196/84984.
- [23] Natnicha Sributsayakarn, Thanapong Intharah, Natthanich Hirunchavarod, Suchaya Pornprasertsuk-Damrongsri, and Varangkanar Jirarattanasopha. Automated age and sex estimation from dental panoramic radiographs. *Int Dent J*, 75(6):103967, 2025. doi: 10.1016/j.identj.2025.103967.
- [24] Ademir Franco, Lucas Porto, Dennis Heng, Jared Murray, Anna Lygate, Raquel Franco, Juliano Bueno, Marilia Sobania, Márcio M. Costa, Luiz R. Paranhos, Scheila Manica, and André Abade. Diagnostic performance of convolutional neural networks for dental sexual dimorphism. *Sci Rep*, 12(1):17279, 2022. doi: 10.1038/s41598-022-21294-1.

- [25] N. Vila-Blanco, M. J. Carreira, P. Varas-Quintana, C. Balsa-Castro, and I. Tomas. Deep neural networks for chronological age estimation from OPG images. *IEEE Trans Med Imaging*, 39(7): 2374–2384, 2020. doi: 10.1109/TMI.2020.2974873.
- [26] K. Kokomoto, R. Kariya, A. Muranaka, R. Okawa, K. Nakano, and K. Nozaki. Automatic dental age calculation from panoramic radiographs using deep learning: a two-stage approach with object detection and image classification. *BMC Oral Health*, 24(1):143, 2024. doi: 10.1186/s12903-024-04160-6.
- [27] G. Ozlu Ucan, O. A. H. Gwasssi, B. K. Apaydin, and B. Ucan. Automated age estimation from OPG images and patient records using deep feature extraction and modified genetic–random forest. *Diagnostics*, 15(3):314, 2025. doi: 10.3390/diagnostics15030314.
- [28] Salem Shamsul Alam, Nabila Rashid, Tasfia Azrin Faiza, Saif Ahmed, Rifat Ahmed Hassan, James Dudley, and Taseef Hasan Farook. Estimating age and sex from dental panoramic radiographs using neural networks and vision–language models. *Oral*, 5(1):3, 2025. doi: 10.3390/oral5010003.
- [29] S. Riaz, M. F. B. Khamis, W. M. A. B. W. Ahmad, J. Y. Abdullah, and M. K. Alam. Potential use of the cusp and crown areas of the maxillary posterior teeth measured with a two-dimensional stereomicroscope for sex determination. *Forensic Sci Med Pathol*, 20(2):452–461, 2024.
- [30] B. Hemalatha, P. Bhuvaneswari, M. Nataraj, and G. Shanmugavadivel. Human dental age and gender assessment from dental radiographs using deep convolutional neural network. *Inf Technol Control*, 52(2):322–335, 2023. doi: 10.5755/j01.itc.52.2.32796.
- [31] Y. Chen, Y. Lin, C. Yeh, and Y. Liu. Detecting representative characteristics of different genders using intraoral photographs: a deep learning model with Grad-CAM interpretation. *BMC Oral Health*, 23:152, 2023. doi: 10.1186/s12903-023-02837-6.
- [32] K. Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics Gems IV*, pages 474–485. Academic Press, San Diego, 1994.
- [33] M. Tan and Q. V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019.
- [34] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

- 675 [35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen. MobileNetV2: Inverted residuals  
676 and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern  
677 Recognition (CVPR)*, pages 4510–4520, 2018.
- 678 [36] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings  
679 of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778,  
680 2016. doi: 10.1109/CVPR.2016.90.
- 681 [37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple  
682 way to prevent neural networks from overfitting. *J Mach Learn Res*, 15(1):1929–1958, 2014.
- 683 [38] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014. URL <https://arxiv.org/abs/1412.6980>.  
684
- 685 [39] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in  
686 deep neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages  
687 3320–3328, 2014.
- 688 [40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Vi-  
689 sual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE  
690 International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.