

# Sex estimation on panoramic dental radiographs: A methodological approach

Ana Beatriz Hougaz<sup>1</sup>, David Lima<sup>1</sup>, Bernardo Peters<sup>1</sup>, Patricia Cury<sup>2</sup>, Luciano Oliveira<sup>1</sup>

<sup>1</sup>Intelligent Vision Research Lab  
Federal University of Bahia (UFBA)

<sup>2</sup>Faculty of Dentistry  
Federal University of Bahia (UFBA)

{ana.hougaz,davidlima,bpmsilva,patricia.cury,lrebouca}@ufba.br





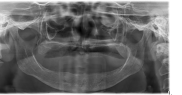




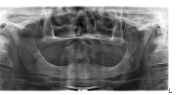
**Abstract.** *Estimating sex using tooth radiographs requires knowledge of a comprehensive spectrum of maxillar anatomy, which ultimately demands specialization on the anatomical structures in the oral cavity. In this paper, we propose a more effective methodological study than others present in the literature for the problem of automatic sex estimation. Our methodology uses the largest publicly available data set in the literature, raises statistical significance in the performance assessment, and explains which part of the images influences the classification. Our findings showed that although EfficientNetV2-Large reached an average F1-score of 91,43% +- 0,67, an EfficientNet-B0 could be more beneficial with a very close F1-score and a much lighter architecture.*

## 1. Introduction

The teeth and mandible are formed by an organic part (mostly collagen) and an inorganic compound (hydroxyapatite), in which calcium phosphate predominates - substances that do not deteriorate quickly. Bones stay long after a person's death, and their decomposition takes thousands of years. Therefore, after accidents, crimes, or natural disasters, person identification based on the analysis of the teeth and jaw is an effective method. However, these inorganic components somewhat show temporal degradation. Mandible is the hardest and most durable bone of the skull, exhibiting a high degree of sexual dimorphism [Saini et al. 2011]. Dental issues, placement of dental implants, cavities, periodontitis, dental plaques, and tooth and bone loss can limit the capacity to identify specific morphological structures that differentiate the sexes. This type of damage is expected, so estimating sex in adults and the elderly is more difficult [Badran et al. 2015] (see Table 1).

Sex determination using skeletal remains represents a great challenge to forensic experts, especially when only body fragments are recovered. Forensic dentists can assist other experts in determining the sex of the remains using teeth and skulls. Many tooth features, such as morphology, crown size, root lengths, and skull patterns, are characteristic of both sexes [Nagare et al. 2018]. However, only specialists in forensic medicine/dentistry can make the evaluations required for sex determination. In cases with numerous images – the result of massive natural disasters and legal services like crime and accidents – the specialist work becomes exhausting and, consequently, human-error prone. Computer-vision models can help accomplish the task quickly and precisely

**Table 1. Examples of radiographs distributed by sex and age. Estimating sex in adults and the elderly is more difficult since dental issues, placement of dental implants, cavities, periodontitis, dental plaques, and tooth and bone loss can limit the capacity to identify specific morphological structures that differentiate the sexes.**

	[0-20]	[21-40]	[41-60]	[61-80]	[81-90]
Female					
Male					

in these situations. These models choose the most helpful image features present in a panoramic dental radiograph (*e.g.*, morphological oral cavity parameters (like the region of the teeth and jaw), outperforming in time the processes of manual measurements.

Sex estimation is fundamental in forensic sciences, becoming the priority in investigating disasters involving many victims. Teeth may be used for differentiating sex by measuring their mesiodistal and buccolingual dimensions. For example, mandibular canines show the greatest dimensional difference with larger teeth in males than in females. Premolars, first and second molars, and maxillary incisors are also known to have significant differences [Sherfudhin et al. 1996]. Using an optical scanner and radiogrammetric measurements on mandibular permanent teeth, sex determination can be performed with 80% of accuracy by measuring root length, and crown diameters [Dayal 1998]. The task is also performed by measuring the canines and the distance between the canine. A study by Anderson and Thompson [Anderson and Thompson 1973] showed that mandibular canine width and intercanine distance were greater in males than in females and permitted a 74% correct classification of sex. In addition, tooth proportions have been used. Sex determination using craniofacial morphology and dimensions is also a common approach. Sex can be predicted correctly in 96% of cases using six traits of skull and mandible: mastoid, supraorbital ridge, size and architecture of skull, zygomatic extensions, nasal aperture, and mandible gonial angle [Williams and Rogers 2006].

Forensic odontology is the branch of dentistry that deals with proper handling and examining dental evidence and the appropriate evaluation and preservation of dental findings. However, most undergraduate programs do not include it in the curriculum [Sharma et al. 2015] [Shivani et al. 2017] [Abdul et al. 2019] [Govindaraj et al. 2018]. Therefore, finding an experienced specialist in this field is challenging. An excellent understanding of loss of indicators and natural degradation of anatomical structures in the oral cavity is required.

## 2. Related works

Manual methods for sex estimation in dentistry and forensic science are based on developmental indicators and manual measurements of anatomical parameters. When major disasters happen, forensic specialists might find only fragmented bones of people. Estimating the sex with 100% accuracy is not possible because not all bones have the degree

**Table 2. Literature review in sex automatic estimation on panoramic radiographs.**

Reference	Pre-processing	Classifier	Data set	Evaluation
[Milošević et al. 2019]	Image adjustments + GradCAM	VGG-16	4,000 (images)	96.87% $\pm$ 0.96 (acc)
[Ilić et al. 2019]	Masking	VGG-16	4,155 (images)	94.3% (acc)
[Ke et al. 2020]	Image adjustments + GradCAM	VGG-16	19,976 (images)	94.6% $\pm$ 0.58 (acc)
[Rajee and Mythili 2021]	Image adjustments	ResNet-50	1,000 (images)	98.27% (acc)
[Milošević et al. 2021]	Image adjustments + segmentation	VGG-16	76,293 (teeth)	72.68% (acc)
[Milošević et al. 2022]	Segmentation	VGG-16	86,495 (teeth)	76.41% (acc)
Ours	GradCAM	EfficientNet-V2-Large	16,824 (images)	91.43% $\pm$ 0.67 (F1-score)

of dimorphism (indicators) for the identification task [Saini et al. 2011]. To tackle this problem, some methods have been proposed in the literature to automatically identify sex on panoramic radiographs (see Table 2).

[Milošević et al. 2021] address the problem of sex estimation by analyzing the individual teeth clipped from the full panoramic dental radiograph as indicated by expert annotators. They used a VGG-16 network to assign the sex, achieving an overall accuracy of 72.68% among 76,293 tooth images. Likewise, [Milošević et al. 2022] followed the same strategies but over a larger data set containing 86,495 tooth images and without image adjustments, reaching an accuracy of 76.41%. Both works present very low accuracy over a small data set with a rough split for training and testing.

The literature has shown that the region of the mandible is essential in sex determination because it is the most dimorphic bone in the body when an intact skull is not available. However, indicators of sexual dimorphism in the mandible decrease with age because of wear and tear [Milošević et al. 2022]. So, in this impasse of choosing the best structure that indicates a person's sex, some studies also determine that the two structures (teeth and the mandible) are essential. Ke et al. (2020) indicate that the deep learning (DL) model usually extracts the best features on the area of the mandible and teeth, showing an accuracy of 94.6%  $\pm$ 0.58.

In terms of methodology, other works on automated sex estimation used accuracy, no cross-validation [Rajee and Mythili 2021, Milošević et al. 2021, Milošević et al. 2022], no study of hyperparameters [Ke et al. 2020, Rajee and Mythili 2021, Milošević et al. 2019], and no explanation

using CAM methods [Ilić et al. 2019, Ke et al. 2020]. First, accuracy brings significant biases, while F1-score is more precise than accuracy, especially for imbalanced classification data sets. Cross-validation is an instrumental technique for assessing the model’s effectiveness and raising the statistical significance of the results. Both cross-validation and hyperband techniques are essential to choosing the best neural network architecture. Another problem was the procedure for choosing the best network since most articles were grounded on pre-selection of the testing set, which might bring substantial biases to the results and analyses.

### 3. Contributions

Differently from the literature works, we opted to use EfficientNets [Tan and Le 2019, Tan and Le 2021], which are pre-optimized networks. We propose to explore a methodological approach, having all the versions of EfficientNets to be compared in a benchmark. Although using the optimized internal parameters of these networks, we also used the hyperband technique to determine the best learning rate and optimizer (external parameters). This process controlled how our models learned and generalized from the training data. Although the use of Grad-CAM to determine the region in the image containing the greatest attention for sex estimation was used in other works [Milošević et al. 2019, Ke et al. 2020], our work exploits this method on the largest publicly available data set (gathered by our research group). It means that other researchers can access our data set, allowing fair comparison against other future studies. Finally, the use of F1-score in a k-fold cross-validation manner to evaluate the methods brings more unbiased conclusions from the numerical results. Even though we might compare the EfficientNets with the other old-school networks, we found that this job would be worthless due to the great effort to optimize the great number of parameters present in the former networks; this way, we opted to focus our attention solely in the smallest and the greatest versions of EfficientNets, bringing important conclusions to possibly have it applied in practice.

## 4. Materials and methods

### 4.1. Data set

Our public data set contains 16,824 panoramic dental radiographs, separated into males and females from Brazilian patients. The data were collected to provide a solid baseline of what we usually find in the “real world” with a support of a specialist in the field. So, no process was done to include synthetically high-quality images. The images show some types of dental problems, such as placement of dental implants, caries, periodontitis, dental plaque, the natural loss of teeth, and damage to the jaw’s skeletal structure. Data distribution is found as follows: 6,341 are from males - corresponding to 37,7% - and 10,483 are from females, corresponding to 61,3%. For assessing the performance of sex estimation, we used the complete data set of 16,824 images that we split in 6 folds of 2,804 images each. Fold division was done to aid the k-fold cross-validation: 11,216 images or four folds for training, 2,804 images or one fold for validation, and 2,804 images or one fold for testing. Figure 1 depicts some samples of our data set. We create some augmentations provided by the *albumations* package, such as resize and horizontal flip, to guarantee a good balance of the two classes to be classified.



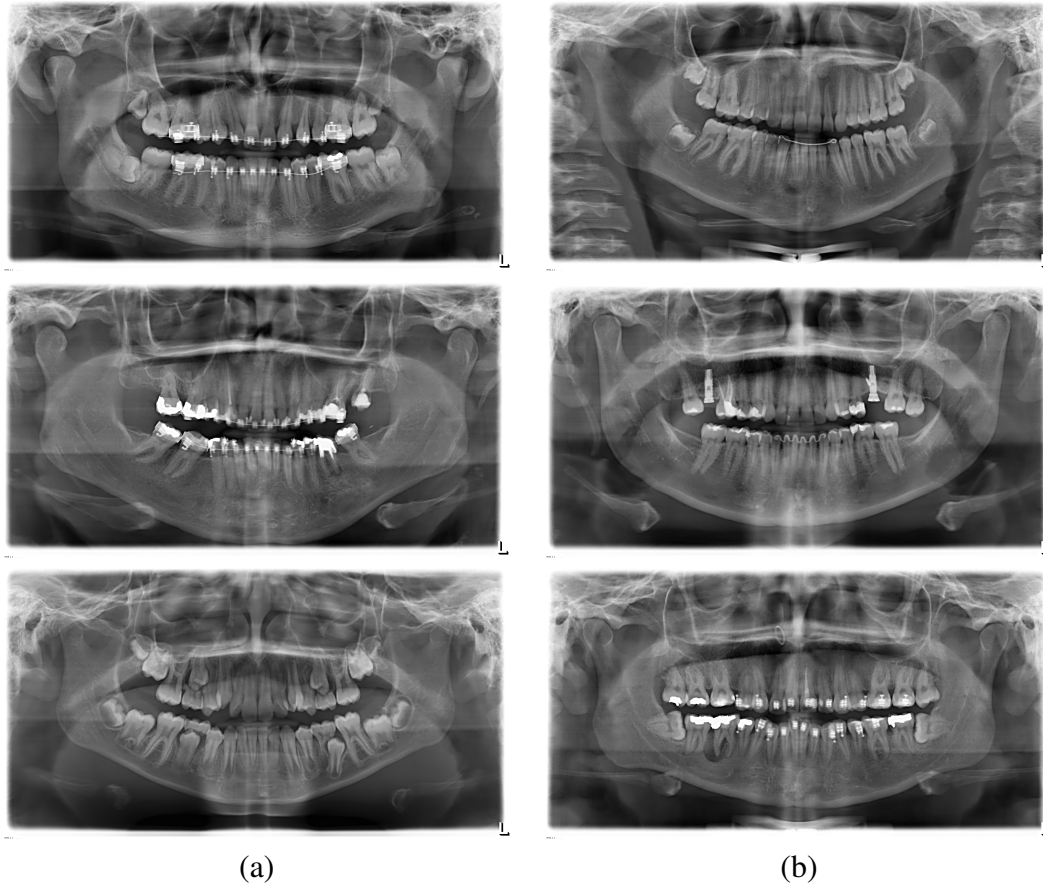


Figure 1. Image samples of our data set: (a) males and (b) females.

Table 3. List of hyperparameters found by the hyperband procedure.

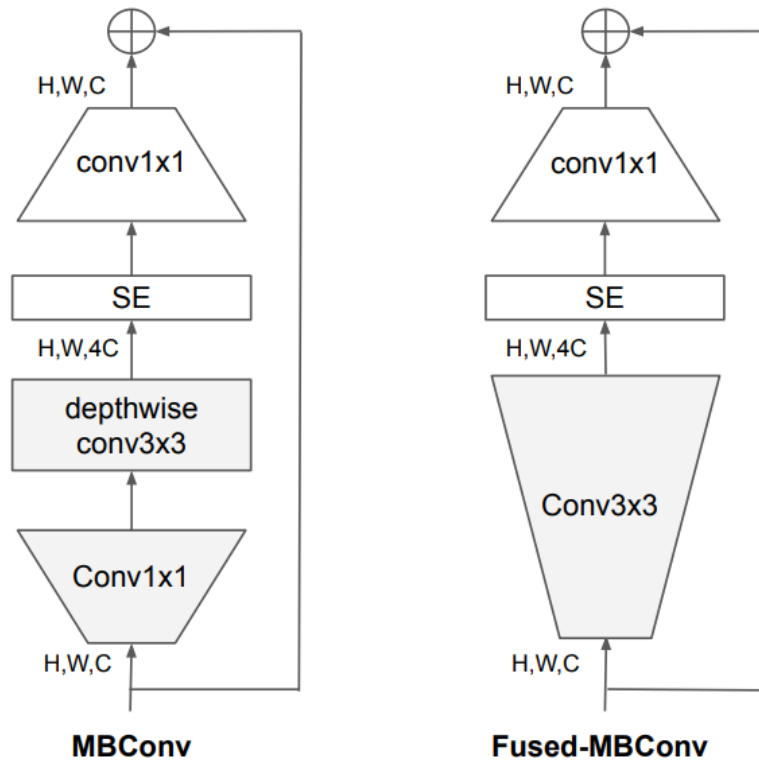
Model	Optimizer	Loss function	Learning rate
EfficientNet-B0	AdamW (betas=(0.9,0.999), amsgrad=False, maximize=False, foreach=None, capturable=False)	Cross-entropy	1,00E-05
EfficientNet-B7			
EfficientNetV2-Small			
EfficientNetV2-Large	Adam (betas=(0.9,0.999), amsgrad=False, maximize=False, foreach=None, capturable=False)		

## 4.2. Hyperband

Hyperband’s method divides the parameters into various combinations and trains several models in parallel, each with a different configuration. Then the models are evaluated, and the worst parameters are eliminated so that only the best-performing parameters are retrained. This method also helps avoid overfitting. This method resulted in the following hyperparameters: AdamW with the default parameters offered by Pytorch like betas=(0.9,0.999), amsgrad=False, maximize=False, foreach=None and capturable=False to EfficientNet-B0, EfficientNet-B7 and EfficientNetV2-Small and Adam with the same default parameters to EfficientNetV2-Large; cross-entropy loss function and a learning rate of  $1e-5$  to all networks in all networks (these parameters are summarized in Table 3).

## 4.3. Trained models

Our experiments started with a transfer learning approach using pre-trained weights provided by the EfficientNets, which were initially trained with the ImageNet data set. All



**Figure 2. Structure of MBConv and Fused-MBConv. Taken from [Tan and Le 2021].**

models used an approach to get better results, including a small learning rate of  $1e-5$  and 120 epochs to train and validate. The last part of the network consists of two fully connected layers with two neurons at the end.

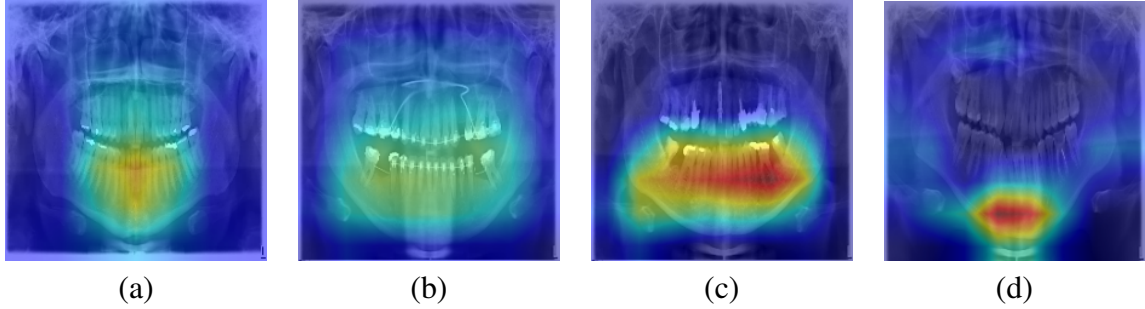
The second version of the EfficientNets has faster training than the corresponding models in the first version [Tan and Le 2021]. One of the main reasons EfficientNetsV2 is faster is due to the use of a new scale normalization method that allows the model to be trained with larger batches of images. In addition, EfficientNetsV2 uses a new type of basic building block called Fused-MBConv. This block combines the MBConv method and the channel-wise separation method in a single operation, significantly reducing the number of convolution operations (see Fig. 2 for an illustration of these two blocks). Notably, we chose just the lightest and heaviest models of each version of EfficientNet. It was done to establish a fast and efficient comparison between each version's smallest and the biggest architecture.

#### 4.4. Methodological evaluation

All results for each model were raised from the point of view of Grad-CAM. For the images that were classified correctly and those that the network made mistakes, we analyzed and defined a methodology that identifies where the network had more difficulty in getting the precise classification. The test sets remained unchanged throughout the experiments. Ultimately, we could study whether the network could generalize and learn how to work with new data from the weights acquired in the training phase, making the experiments reproducible.

**Table 4. Summary of the results found considering the smallest and largest model of each EfficientNet version.**

Model	F1 score
EfficientNet-B0	91.30% $\pm$ 0.47
EfficientNet-B7	90.00% $\pm$ 0.01
EfficientNetV2-Small	89.10% $\pm$ 0.67
<b>EfficientNetV2-Large</b>	<b>91.43% <math>\pm</math>0.67</b>



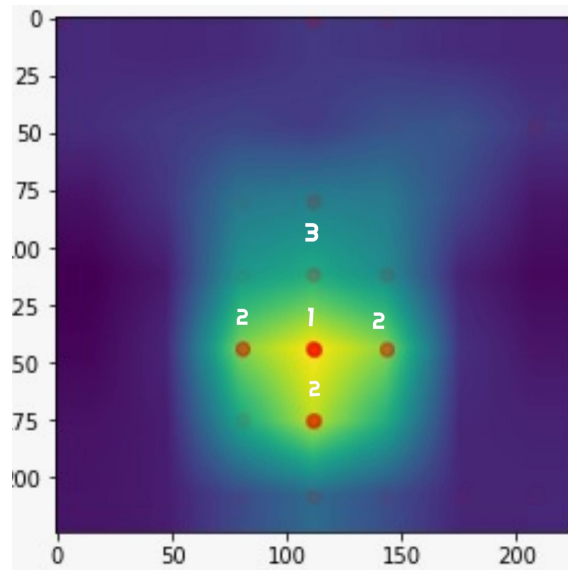
**Figure 3. The average attention region of the Grad-CAM method in all networks: (a) EfficientNet-B0, (b) EfficientNet-B7, (c) EfficientNetV2-Small and (d) EfficientNetV2-Large. A hotter tone indicates a greater contribution to the correct estimate.**

## 5. Result analysis

In the experimental analysis illustrated in Table 4, EfficientNetV2-Large presented the best result. However, in terms of the trade-off between cost and speed, the EfficientNet-B0 performed as well as the EfficientNetV2-Large. The F1-score of EfficientNet-B0 had only a decrease of 0.13 percentage points compared to EfficientNetV2-Large but also with a smaller standard deviation.

The Grad-CAM method was applied during the network validation steps to evaluate the average region where all networks extracted features when the classification was correct. Figure 3 shows the results of the GradCAM method for each one of the networks used. In this case, it is noticeable that the common region of average maximum attention in all networks was in the central region under the mouth, where the structure of the mandible and the lower teeth is present. For example, the point of maximum attention of EfficientNet-B0 (Fig. 3(a)) was located around the lower teeth, specifically between the second premolars on the left and right sides and the mandible (around the red color in the heatmap produced by the Grad-CAM). As a matter of fact, this region is verified by the specialist to determine a person's sex once our mandibular canines exhibit a noticeable disparity in dimensions, with males presenting a greater contrast in size than their female counterparts [Sherfudhin et al. 1996]. The mandible is the skull's sturdiest and most long-lasting bone, depicting a significant level of sexual dimorphism [Saini et al. 2011]. In clinical practice, sex estimation can also be accomplished by evaluating the canines and the intercanine distance. Anderson and Thompson [Anderson and Thompson 1973] conducted a study that revealed greater mandibular canine width and intercanine distance in males than in females.

On the other hand, the network does not completely discard complementary re-



**Figure 4. Accounting for the average of all maximum attention points of the Grad-CAM method of all networks together when the network estimated sex correctly.**

gions from that one in the mandible. All the networks also extract some points outside the region of maximum attention in the mandible, contributing to the final classification. We called these points in question the least relevant. Figure 4 illustrates the less relevant points as the overall average regions of the maximum attention points of all networks when they correctly classified a radiograph. The most relevant point is still shown in the image, and it is fixed in the center of the mandibular bone (1); the second points (2) are placed in the lower teeth and between the lower second premolars on both the left and right sides; the third point (3) is located in the areas close to the upper bone structures in the anterior region of the maxilla and the upper incisor.

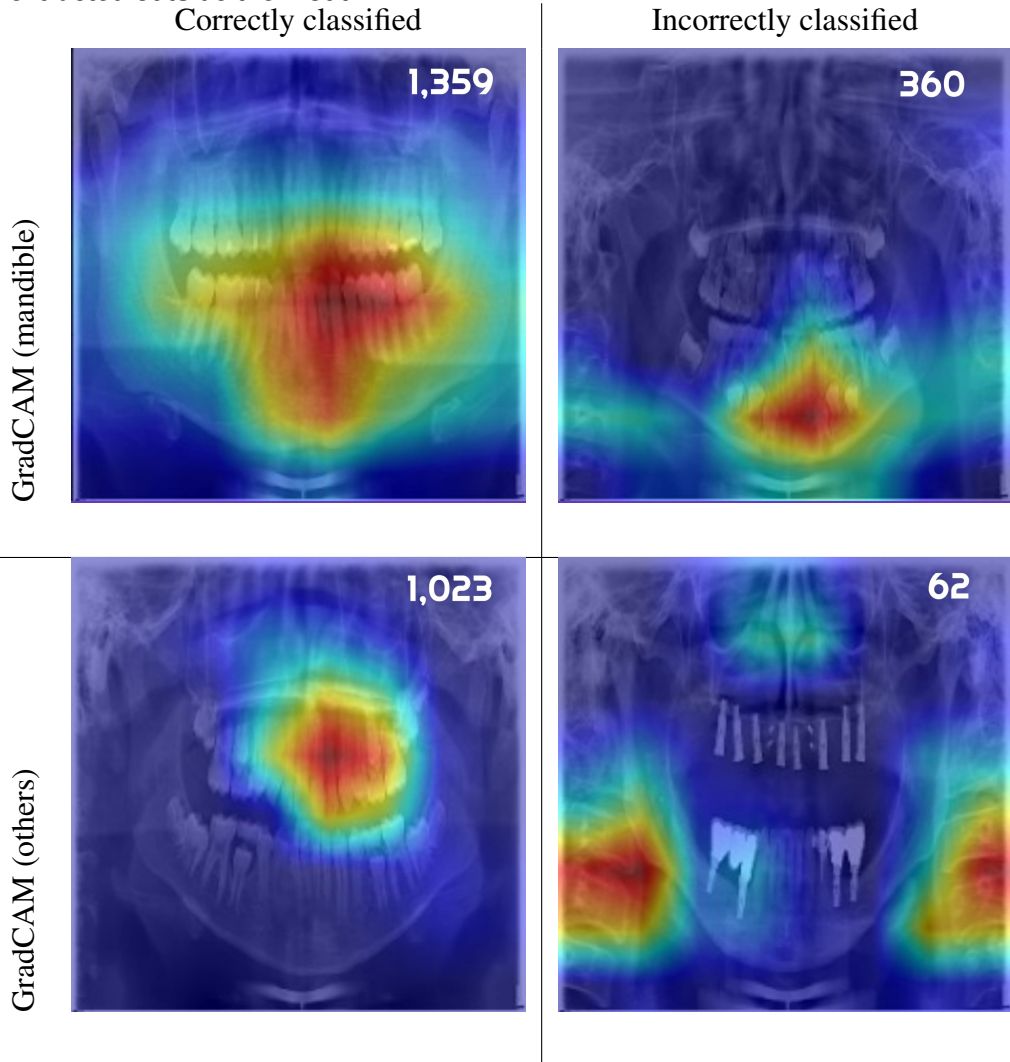
The structure of the maxilla is also important to estimate the sex of a patient as there are differences in the shape and size of the maxilla between men and women. The premolars, first and second molars, and the maxillary incisors exhibit noteworthy discrepancies and play a crucial role in sex estimation [Sherfudhin et al. 1996, Saini et al. 2011]. For example, women tend to have wider upper incisor teeth than men. Therefore, the analysis of these factors together with the structure of the mandible can provide hints to improve the task of sex estimation based on panoramic radiography.

### **5.1. Explaining which image regions were determinant for the classification**

To dive deep into the matter of which image region best determines an individual's sex, we also analyzed the results using a confusion matrix containing the correct and incorrect classifications and the average region used to classify the feature vector (see Table 5). To this end, we used the best model of EfficientNetV2-Large trained in the cross-validation procedure and predicting over a previously separated test data set containing 2,804 images.

Noticeably, more images are correctly classified when the network extracts features in the mandible area (48.4%). However, looking at the other regions of the panoramic radiograph, we have a nearly amount of correctly classified images, with only

**Table 5. Confusion matrix with the average image region used to classify sex, and the correct and incorrect classification by EfficientNetV2-Large. A mandible region showing a correct classification is expected, and it occurred in 48.5% of the images. Correct classification with features extracted in other regions rather than the mandible occurred in 36.5%. Incorrect classifications happen in 12.8% of the images and inside the mouth, while circa 2.2% of the images, features are extracted outside the mouth.**



336 images less than when focusing on the jaw (corresponding to 36,5% of the images in the test data set). This shows that other regions, such as the maxilla, upper incisor teeth, premolars, first and second molars, and maxillary incisors, are also recognized to present significant differences and contributions to the estimation of sex.

When the network incorrectly classifies an input image, and the GradCAM points out that the network extracted features outside the jaw, this crass mistake made by the network completely hinders the classification; it occurs in less than 2.2% of the images and could be circumvented by an attention model if integrated into the EfficientNetV2-Large. Finally, 12.6% of the images are incorrectly classified by the network although the features fall inside the mouth and in a region close to the jaw, indicating that the region was not discriminating enough to the classifier.

## 6. Discussion and final remarks

In this paper, we conducted a study with the largest publicly available data set in the literature, containing 16,824 images, and the experiments showed that the EfficientNetV2-Large achieved superior performance. Our methodological study reinforced the use of an appropriate metric for evaluating the models, as was the case of F1-score, opposing the other studies in the literature that used accuracy even with an imbalanced data set. Although we found an F1-score of 91,43%  $\pm$ 0,67 for EfficientNetV2-Large as the highest one, EfficientNet-B0 is a much lighter architecture, which achieved a very close F1-score. For that, we strongly recommend using EfficientNet-B0 for practical applications of sex estimation.

Grad-CAM played an important role in our study, providing a way to understand how the EfficientNets estimated the sex and identifying the important regions for estimation in the validation step. This helped identify some limitations of the models, which can aid future studies to tackle those problems. The hyperband method also helped as a hyperparameter optimization technique that was used to find the best combination of the learning rate and optimizer. The use of cross-validation favored the analysis of the statistical significance of the results, which presented a very low standard deviation for all networks. The applied methodology including the use of Grad-CAM to identify the best regions, the hyperband to select the best parameters, the largest publicly available in the literature, and a more diversified data set and the discovery of the mandible. The regions adjacent to the upper bone structures in the anterior part of the maxilla and upper teeth worked as crucial areas in our sex estimation to allow for explaining the problem toward the improvement of new future studies in the field. Finally, we can state that this study not only strongly validates the application of EfficientNets for sex estimation in dental panoramic radiographs but also sets a new benchmark for future research in the field of forensic science.

## References

- Abdul, N. S., Alhazani, L., Alruwail, R., Aldres, S., and Asil, S. (2019). Awareness of forensic odontology among undergraduate, graduate, and postgraduate dental students in riyadh, saudi arabia: A knowledge-, attitude-, and practice-based study. *Journal of forensic dental sciences*, (1):35–41.
- Anderson, D. and Thompson, G. (1973). Interrelationships and sex differences of dental and skeletal measurements. *Journal of Dental Research*, 52:8–431.
- Badran, D. H., Othman, D. A., Thnaibat, H. W., and M., A. W. (2015). Predictive accuracy of mandibular ramus flexure as a morphologic indicator of sex dimorphism in jordanians. *International Journal of Morphology*, 33(4):1248–1254.
- Dayal, P. (1998). *Textbook of Forensic Odontology*. Hyderabad Paras Medical Publishers, 1 edition.
- Govindaraj, S., Jayanandan, M., Vishnu Priya, V., Thirumal, R., and Shamsudeen, S. (2018). Knowledge and attitude among senior dental students on forensic dentistry: A survey. *World Journal of Dentistry*, 9:91–187.
- Ilić, I., Vodanović, M., and Subašić, M. (2019). Gender estimation from panoramic dental x-ray images using deep convolutional networks. *IEEE*, pages 1–5.

- Ke, W., Fan, F., Liao, P., Lai, Y., Wu, Q., Du, W., Chen, H., Deng, Z., and Zhang, Y. (2020). Biological gender estimation from panoramic dental x-ray images based on multiple feature fusion model. *Sensing and Imaging (2020)*, 21:1–11.
- Milošević, D., Vodanović, M., Galić, I., and Subašić, M. (2019). Estimating biological gender from panoramic dental x-ray images. *International Symposium on Image and Signal Processing and Analysis, ISPA*, 2019-September:105–110.
- Milošević, D., Vodanović, M., Galić, I., and Subašić, M. (2021). Automated sex assessment of individual adult tooth x-ray images. *International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 72–77.
- Milošević, D., Vodanović, M., Galić, I., and Subašić, M. (2022). A comprehensive exploration of neural networks for forensic analysis of adult single tooth x-ray images. *IEEE*, pages 70980–71002.
- Nagare, S. P., Chaudhari, R. S. and Birangane, R. S., and Parkarwar, P. C. (2018). Sex determination in forensic identification, a review. *Journal of forensic dental sciences*, [https://doi.org/10.4103/jfo.jfds\\_55\\_17](https://doi.org/10.4103/jfo.jfds_55_17), 10(2):61–66.
- Rajee, M. and Mythili, C. (2021). Gender classification on digital dental x-ray images using deep convolutional neural network. *Biomedical Signal Processing and Control*, 69:102–139.
- Saini, V., Srivastava, R., Rai, R. K., Shamal, S. N., Singh, T. B., and Tripathi, S. K. (2011). Mandibular ramus: an indicator for sex in fragmentary mandible. *Journal of forensic sciences*, 56:6–13.
- Sharma, A., Shokeen, S., Arora, R., and Dhaginakatti, S. (2015). Survey on knowledge, attitude and practice of forensic odontology among private dental practitioners in ghaziabad city, india. *Journal of Dental Specialities*, 3:7–43.
- Sherfudhin, H., Abdullah, M., and Khan, N. (1996). A cross-sectional study of canine dimorphism in establishing sex identity: Comparison of two statistical methods. *Journal of Oral Rehabilitation*, 23:31–627.
- Shivani, B., Arshroop, K., Karanprakash, S., Mahjeet, S., Navgeet, P., and Chitra, A. (2017). Perception of forensic odontology and its practice among local dentists of an institution. *Journal of Forensic Research*, 8:1–4.
- Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*, pages 6105–6114.
- Tan, M. and Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. *International Conference on Machine Learning*, pages 10096–10106.
- Williams, B. and Rogers, T. (2006). Evaluatiing the accuracy and precision of cranial morphological traits for sex determination. *Journal of forensic sciences*, 4:35–729.