Automated Sex Assessment of Individual Adult Tooth X-Ray Images

Milošević, Denis; Vodanović, Marin; Galić, Ivan; Subašić, Marko

Automated Sex Assessment of Individual Adult Tooth X-Ray Images

Denis Milošević*, Marin Vodanović†, Ivan Galić‡, Marko Subašić*
*Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
†School of Dental Medicine, University of Zagreb, Croatia
‡University Hospital Centre Split, Croatia

***Abstract*—Sex assessment is an important step of the forensic process. Dental remains are often the only remains left to examine due to their resistance to decay and external factors. Contemporary forensic odontology literature describes multiple methods for sex assessment from mandibular parameters, all of which require manual measurements and expert training. This study aims to explore the applicability of deep learning and image analysis methods to automate this task, thus allowing for easier reproducibility of assessments, reduction of the time experts lose on repetitive tasks, and potentially better performance. We have evaluated state-of-the-art deep learning models and components on the largest dataset of individual adult tooth x-ray images, consisting of 76293 samples. This study also explores the usage of decayed or structurally altered teeth, with which contemporary methods struggle. Two types of models are constructed, a family of models specialized for specific tooth types, and a general model that can assess the sex from any tooth type. We examine the performance of those models per tooth type and age group, as well as the impact of decayed and structurally altered teeth. The specialized models achieve an overall accuracy of 72.40%, and the general model reaches an overall accuracy of 72.68%.**

***Index Terms*—forensic odontology, x-ray image analysis, convolutional neural network, deep learning, machine learning, image processing**

I. Introduction

Sex assessment is one of the first steps in the forensic process [1]. Current forensic odontology literature proposes methods based on manual measurements of mandibular parameters. Mastering those methods requires years of training and education, and is performed only by trained experts. Furthermore, this allows for human error to creep into the results and reduces the overall reproducibility.

Skeletal remains can be used to assess the sex with near 100% accuracy [2]. Despite that, estimation from dental remains is an important toolbox of a forensic expert, as those remains show high resistance to external factors and decay, and are often the only remains left to examine. External factors include blunt force, fire, bacterial decomposition, and other degenerative processes [3], [4], all of which have a lower impact on dental remains. Those methods are used in a wide variety of fields, ranging from legal proceedings to exploration of demographic changes of historically important sites.

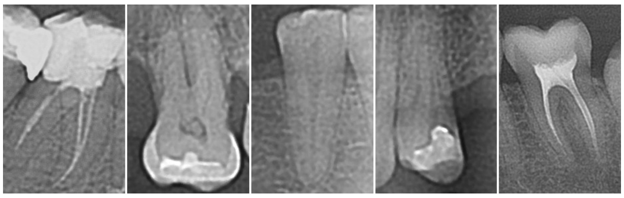

Fig. 1. **Samples of individual teeth.** Individual teeth are clipped from the full panoramic dental x-ray image as indicated by expert annotators.

In this study, we specifically focus on sex assessment from x-ray images of individual teeth. To that end, we've collected the largest dataset in forensic odontology literature, which consists of 76293 unique adult tooth images. This dataset contains images of imperfect teeth that can contain various illnesses, deformations, and dental alterations. As contemporary methods require manual measurements, and therefore a high time investment, we're exploring the applicability of deep learning and image analysis methods for the automation of this task. Deep learning models consisting of well-tested and proven elements, like state-of-the-art convolutional neural network architectures as feature extractors and the attention mechanism, are evaluated for this task. Specialized models for each tooth type are trained, as well as a general model that assesses the sex from any tooth type. Additionally, we've annotated a subset of the dataset with status annotations, and we've evaluated the impact of dental alterations, decay, and illness on the results.

II. Related works

Sex assessment in forensic odontology is done through measurements of different morphometric and nonmetrical parameters of the mandible. Early works have explored this topic [5]–[9], as well as in contemporary literature [10].

Many studies in the literature are based on mandibular parameters and with datasets of various sizes. A sample size of 40 can be found in [11], where three different approaches yield accuracies of 72.5% to 95%. On the other hand, another study with a sample size of 419 [12] achieves an overall accuracy of 70.9%. Despite the larger sample size, [12] has an imbalanced dataset that is skewed toward female samples. An accuracy of 80.2% is achieved again using mandibular parameters in [1], but their dataset is heavily biased toward male samples. Different mandibular parameters and combinations of those parameters are used in literature, with [13] having an accuracy of 85%, and [14] claiming an accuracy between 94% to 99%. A different approach is the usage of geometric morphometric methods in addition to 10 mandibular parameters, by

2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA) | 978-1-6654-2639-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISPA52656.2021.9552124

which [11] achieve an accuracy of 95% with a dataset of 40 individuals. Those methods and parameters are verified on different populations across the globe, and they match the performance of the methods they're based on [1], [15]–[21]. All cited studies worked with intact mandibles, without any pathology, loss of mandibular molars, or anomalous molars and teeth [12]. Deep learning-based approaches have also been proposed for panoramic dental x-ray images, attaining remarkable performance [22], [23].

Forensic odontology literature suggests that assessment of sex from individual teeth is not recommended [24]. As per literature, teeth are only a useful supplement for sex assessment, as the accuracy of those methods is not sufficient [25], not being able to reach even 80% [24]. A systematic review of contemporary sex assessment literature can be found in [26].

III. Data

Our dataset consists of 76293 individual tooth x-ray images which have been taken from 2683 panoramic dental x-ray images. Those 2683 panoramic dental x-ray images have been taken from 2680 male and female individuals in the age range from 19 to 85 years. Experts have annotated those images with bounding boxes for each individual tooth and their designation as per the FDI World Dental Federation notation (ISO 3950) [27]. The dataset contains 44321 images of female teeth, and 31972 images of male teeth, which gives a female to male ratio of 58% to 42%. The data is not evenly distributed in all age groups, as the majority of samples fall in the age range of 20 to 55 years old. A detailed overview of the data per age group and sex for individual tooth images and source panoramic dental x-ray images can be seen in Table I. Examples of the images in the dataset can be seen in Fig. 1.

TABLE I
Detailed overview of data samples per age group.

Age group	**Orthopantomographs**		**Individual teeth**	
	Female	Male	Female	Male
[18, 20)	12	8	367	250
[20, 25)	205	105	6239	4600
[25, 30)	257	147	7688	4409
[30, 35)	259	161	7641	4751
[35, 40)	232	168	6652	4832
[40, 45)	177	127	4986	3566
[45, 50)	115	97	3104	2667
[50, 55)	110	80	2851	2158
[55, 60)	76	61	1922	1540
[60, 65)	58	57	1464	1406
[65, 70)	26	33	647	840
[70, 75)	22	22	565	561
[75, 80)	6	11	152	299
[80, 85)	2	3	43	69
[85, 90)	0	1	0	24
Subtotal	**1557**	**1126**	**44321**	**31972**
Total	**2683**		**76293**	

The samples are collected from multiple locations in Croatia and belong to the collection of the Department of Dental Anthropology School of Dental Medicine University of Zagreb. The use of this collection for research purposes has been approved by the ethics committee School of Dental Medicine University of Zagreb. All samples are anonymized, and no personal data is stored alongside the images. An identity hash is used to differentiate between images of different people. The hash is not reversible, and no identifying information can be gathered from it.

No filtering of teeth by any anomalies or alterations has been done to the data. Teeth can have various changes, be they natural (decay) or artificial (dental interventions). Those changes can obstruct the view of the tooth morphology, which makes them unsuitable to most current sex assessment methods. The teeth in our dataset can have fillings, endo-fillings, crowns, bridges, implants, carious lesions, appliances, and they can be left behind roots. Of the 2683 panoramic dental x-ray images, only 983 have alteration annotations.

This study uses two types of tooth notation systems. The dataset uses the FDI World Dental Federation notation (ISO 3950) [27]. As teeth on the same jaw side can be considered symmetrical, and as much of the literature in forensic odontology does, we adopt a notation where teeth are identified by their jaw side and position. For example, canines on the upper jaw (13 and 23 as per ISO 3950) are both labeled as ”Up-3”.

The dataset is split into three parts - train, validation, and test. The data in the train set is used to fit the model. The validation set is used to evaluate the performance of the trained model for research decisions. While performance on the validation set is used as an indicator for research decisions, the test set is used for the final reported results. In other words, the training dataset is used to fit the model, the validation dataset is used to determine model hyperparameters, and the test set is unseen by any experiment until the final model has been trained. The size ratio between the train, validation, and test set is 70% : 15% : 15%. Images with the same identity hash were assigned to the same data subset.

In addition to these three sets, another subset of teeth without alterations was constructed. While the models have to handle all kinds of alterations, it is still interesting to explore the model performance on perfect dentition. This subset consists of 80 images, with 5 images per tooth type. While this is a small sample size compared to the entire dataset, it still gives insight into the influence of alterations for sex assessment.

The original images are taken with a variety of orthopantomographs, resulting in an image with a width between 1127 px to 3260 px and a height between 553 px and 1536 px. From those images, individual tooth images are extracted as per their annotated bounding box. Each individual tooth image can be of a different size, but no image has a dimension larger than 512 px. All images are padded with black color to achieve a size of 512x512 px across all images in the dataset.

IV. Method

This study explores the capabilities of deep learning and image analysis for automated sex assessment of x-ray images of individual teeth. Deep learning, specifically convolutional neural networks, has shown amazing capabilities in the field of image analysis. To determine the potential of using deep learning for forensic odontology, this study uses only well-tested and proven architectures and techniques instead of designing a lower-capacity custom network. While those types of networks tend to produce acceptable results, they are often too unstable for real-world use. As our dataset contains 76293 images, transfer learning [28] was not required. Preliminary experiments have shown that using networks pretrained on ImageNet did not improve the overall result, nor did it significantly decrease training time.

The studied models consist of four main parts: the base state-of-the-art convolutional neural network architecture, an additional 1x1 convolutional layer to change the number of feature maps in the final layer of the feature extractor, an optional attention mechanism [29], and a two-layer fully-connected network with an adjustable number of units in the first fully connected layer.

For those components, reasonable hyperparameter spaces have been determined. Those hyperparameter spaces are in accordance with deep learning literature and other medical image analysis studies. Random search is used [30].

After the hyperparameter search, the best-performing model is selected for further training. The best model is chosen by its performance on the validation dataset. Fine-tuning is a model training process where training hyperparameters are adjusted to maximize the performance of the model and to achieve those final few percentage points. Those usually include adjustments of the optimizer, like the introduction of a learning rate schedule. Finally, the trained model is evaluated on the before unseen test dataset.

Two types of experiments have been done for this study. One type trained sex assessment models for a specific tooth type. For example, one model would be trained only on maxillary canines, and it would be evaluated only on maxillary canines. A model family for each tooth type has been trained. The other type of experiments explored sex assessment models that work with any tooth type.

A. The model

As already mentioned, the model consists of four parts. Each part is a proven and well-known component. The goal of this study is to explore the applicability of deep learning methods on the forensic odontology problem of sex assessment. While a custom-made model might achieve marginally better results, those solutions often provide brittle, overfit models.

The first part is the convolutional neural network architecture used as the feature extractor. The following architectures were tested: DenseNet201 [31], InceptionResNetV2 [32], ResNet50 [33], VGG16, VGG19 [34] and Xception [35]. The viability of using ImageNet pretrained network weights was evaluated in preliminary experiments. None of the tested models showed any kind of improvement when transfer learning was used. We have therefore decided to not use pretrained weight as a starting point for training. Instead, each network got randomly initialized as described in their original paper.

An important note on those state-of-the-art architectures is that they are designed with RGB images in mind. X-ray images are grayscale. In technical terms, the network architectures expect a 3-channel input, but our images are only single-channel. We duplicate the value of our images across all 3 channels, effectively achieving a 3-channel grayscale image, and thus allowing us to use SOTA network architectures.

The second part of the model consists of a single 1x1 convolutional layer. Our images are less diverse than images found in ImageNet. Having the same amount of filters in the final convolutional layer for ImageNet and for sex assessment might lead to overfitting. The 1x1 convolution is used to scale that final feature tensor in the channel dimension, allowing us to downsize the capacity at the end of the feature extraction network without having to do drastic changes to the state-of-the-art architecture itself.

The third part of the model is an optional attention module. Attention as a mechanism has been shown to be very successful in cutting edge computer vision and natural language processing [29], [36]. Given its success, it was included in the hyperparameter search. To determine if the attention mechanism is meaningfully contributing to the automation of this particular problem, attention is optional.

The fourth and last part is the classification network. It consists of two fully connected layers. The size of the first layer is a hyperparameter, while the size of the second layer is fixed to two units. The second layer represents the class probabilities, of which there are two for this problem. Batch normalization is used between those two layers.

All activations in the network are ReLU [37], except for the last fully-connected layer which used the softmax activation. The objective function is cross entropy [38].

B. Hyperparameter search

Random search is used for the search algorithm [30]. Random search has benefits and guarantees that are sufficient for this study. The approach is ”embarrassingly parallel”, allowing for parallel execution of the experiments. Additionally, random search gives us an easy to understand probability that the solution is within a certain percentage of the best possible solution in the selected search space. For the hyperparameter search space, it is intuitively clear that there is some point that performs best compared to all other points in the search space. That solution does not have to be the global optimum, but it's the optimum within that selected subspace. So for a random point, there's a probability of p_1 that the point is within p_1 percent around the best solution. For n points, the probability that all of them miss that subspace is $(1-p_1)^n$. Therefore, to achieve a probability of p_2 that at least one point is within p_1 percent of the best solution, $1-(1-p_1)^n > p_2$ holds true. For $p_1 = 0.05$ and $p_2 = 0.95$, $n \geq 60$. In this study, 256

TABLE II
OVERVIEW OF MODEL HYPERPARAMETERS USED FOR THE GRID SEARCH.

Hyperparameter	Search space	Best value
Pretrained feature extractor	DenseNet201, InceptionResNetV2, ResNet50, VGG16, VGG19, Xception	VGG16
1x1 convolution channel size	Between 5 to 1000	40
Attention mechanism	Present or not present	Not present
Fully-connected size	Between 1 and 2048	128

experiments were done, indicating that the model found has a high likelihood to be very close to the best solution in the selected search space. Hyperparameters in this study determine the model as described in Section IV-A. An overview of all model hyperparameters can be seen in Table II.

Hyperparameter search is a computationally expensive operation. To train the model sufficiently for the validation set evaluation to be indicative of the final performance, we have empirically determined that, for this problem, it is enough to train the model on the entire train dataset with Adam [39] as the optimizer, with a learning rate of $3.24 \cdot 10^{-4}$ for 100 epochs.

C. Model fine tuning

Once the best model has been found, it is trained again with a different training regime to achieve the best possible performance. Models take longer to train this way, but they eventually perform better than their quickly trained counterparts. Different training strategies were tested, but the best performing regime uses SGD as the optimizer with a learning rate schedule. The learning rate schedule used is cosine annealing with warm restarts [40]. For each epoch, the learning rate is determined as follows:

$$\eta_t = \eta_{min} + \frac{1}{2}\left(\eta_{max} - \eta_{min}\right)\left(1 + cos\left(\frac{T_{cur}}{T_i}\pi\right)\right)$$

Where η_t is the learning rate in epoch t, η_{min} and η_{max} are the minimum and maximum learning rate respectively, T_{cur} is the current epoch in the period, and T_i is the number of epochs in a period.

For this study, the maximum learning rate is 10^{-3}, the minimum learning rate is 10^{-7}, the period is set to 100 epochs, and the model is trained for 1000 epochs in total.

D. Evaluation

Two different datasets are used for evaluation. For results that are used to determine research decisions (for example, hyperparameter selection), the performance is measured on the validation dataset. The size of the validation dataset differs between experiment types. For specialized models, each validation dataset is of a different size, as there is a different amount of images for each tooth type. Still, the validation set is 15% of all images of that specific tooth type. For the general model, the dataset is much bigger in total numbers, but the validation set is still 15% of the total dataset. The results reported in this study are obtained from the test dataset. Again, the size in the number of images differs between tooth types, but it is important to note that those images were not used during any part of this study except for the final evaluation.

Additionally, the general model performance is evaluated on a separate test set consisting of only healthy, unaltered teeth. As that dataset contains only 5 samples per tooth type, evaluation on the specialized models would not be indicative and is therefore not performed. That dataset was generated from the status annotations that some images contain, but they were furthermore verified by dental experts. The evaluation metric used for classification is accuracy.

V. RESULTS AND DISCUSSION

Hyperparameter search resulted in models of accuracy between 58% and 71%. The most successful model uses VGG16 as its feature extractor, it uses 40 feature maps in the final convolutional layer, it uses no attention mechanism and it has 128 units in the first fully-connected layer.

When fine-tuned, models specialized by tooth type range in accuracy from 69.04% to 76.66% for mandibular teeth, and from 60.84% to 77.10% for maxillary teeth. For mandibular teeth, the best performing teeth are canines (Down-3), and the worst-performing teeth are lateral incisors (Down-2). For maxillary teeth, the best performing teeth are again canines (Up-3), and the worst-performing teeth are first molars (Up-6). The specialized models achieve an overall accuracy of 72.40%. Detailed results per age group are shown in Table III.

The general model can assess the sex from any tooth type. The accuracy per tooth type for this model range from 72.22% to 77.41% for mandibular teeth, and from 68.05% to 74.44% for maxillary teeth. For mandibular teeth, the best performing teeth are first molars (Down-6), and the worst-performing teeth are central incisors (Down-1). For maxillary teeth, the best performing teeth are second molars (Up-7), and the worst-performing teeth are first premolars (Up-4). The general model achieves an overall accuracy of 72.68%. Detailed results per age group are shown in Table IV.

When evaluated on the subset of pristine teeth, 33/40 samples of the mandibular teeth got correctly classified, and 35/40 samples of maxillary teeth got correctly classified. In other words, mandibular accuracy is 82.5%, maxillary accuracy is 87.5% and overall accuracy is 85%. While this is a much smaller dataset, this result indicates that tooth decay and dental interventions significantly hinder the process of sex assessment.

As can be seen in Fig. 2 (left), there is no clear trend in assessment accuracy in regards to age groups. Conversely, as seen in Fig. 2 (right), a trend per tooth type is noticeable. For both the general model and the specialized models, canines seem most suitable for the task.

VI. CONCLUSION

In this study, we've shown that the forensic odontology task of sex assessment from single tooth x-ray images can be

TABLE III
ACCURACY OF SEX ASSESSMENT PER AGE GROUP OF THE MODELS SPECIALIZED TO SPECIFIC TOOTH TYPES. UP REFERS TO MAXILLARY TEETH, AND DOWN REFERS TO MANDIBULAR TEETH.

Age	1		2		3		4		5		6		7		8	
	Up	Down	Up	Down	Up	Down	Up	Down	Up	Down	Up	Down	Up	Down	Up	Down
[15, 20)	50.00%	100%	50.00%	83.33%	83.33%	66.67%	80.00%	100%	80.00%	83.33%	60.00%	80.00%	100%	66.67%	75.00%	100%
[20, 25)	73.47%	70.41%	78.35%	69.39%	77.55%	73.47%	75.00%	79.38%	65.26%	77.32%	69.15%	71.43%	75.51%	67.35%	78.75%	84.52%
[25, 30)	71.67%	73.95%	72.27%	69.49%	82.20%	85.71%	81.74%	78.81%	71.96%	81.25%	60.18%	80.00%	70.69%	81.42%	75.79%	77.08%
[30, 35)	78.23%	75.00%	72.58%	68.55%	81.97%	72.58%	75.86%	73.55%	73.04%	80.67%	57.94%	77.27%	70.34%	81.36%	70.65%	72.29%
[35, 40)	69.17%	62.71%	66.96%	69.75%	74.58%	82.20%	77.27%	72.03%	64.76%	70.69%	58.49%	60.92%	72.48%	74.23%	68.06%	63.38%
[40, 45)	72.63%	70.10%	78.35%	67.35%	71.88%	75.51%	74.71%	68.42%	75.61%	73.63%	65.28%	75.38%	66.67%	75.90%	70.00%	61.11%
[45, 50)	71.15%	65.38%	61.54%	59.62%	71.15%	76.47%	69.05%	71.15%	75.00%	63.27%	67.50%	68.75%	68.00%	75.56%	86.67%	73.33%
[50, 55)	75.00%	84.62%	71.15%	63.46%	66.00%	69.23%	72.34%	60.00%	50.00%	63.04%	45.45%	54.84%	76.09%	78.57%	61.90%	64.00%
[55, 60)	73.33%	62.50%	79.07%	68.75%	80.43%	72.92%	62.86%	73.33%	52.94%	88.10%	50.00%	64.00%	75.00%	76.32%	68.42%	55.00%
[60, 65)	83.72%	76.74%	82.50%	79.55%	78.95%	74.42%	56.25%	72.09%	86.67%	67.57%	70.37%	65.52%	64.29%	76.92%	73.33%	89.47%
[65, 70)	76.19%	72.73%	77.27%	68.18%	77.27%	65.00%	60.00%	66.67%	58.82%	84.21%	57.14%	72.73%	56.25%	46.15%	62.50%	50.00%
[70, 75)	78.57%	85.71%	92.86%	92.86%	85.71%	78.57%	69.23%	85.71%	80.00%	69.23%	75.00%	100%	55.56%	75.00%	50.00%	100%
[75, 80)	66.67%	83.33%	33.33%	66.67%	83.33%	100%	40.00%	83.33%	80.00%	50.00%	40.00%	75.00%	66.67%	40.00%	100%	50.00%
Total	73.62%	71.59%	73.19%	69.04%	77.10%	76.66%	73.86%	73.41%	69.09%	75.17%	60.84%	71.76%	70.83%	75.58%	72.85%	72.11%

TABLE IV
ACCURACY OF SEX ASSESSMENT PER AGE GROUP OF THE GENERAL MODEL. UP REFERS TO MAXILLARY TEETH, AND DOWN REFERS TO MANDIBULAR TEETH.

Age	1		2		3		4		5		6		7		8	
	Up	Down	Up	Down	Up	Down	Up	Down	Up	Down	Up	Down	Up	Down	Up	Down
[15, 20)	83.33%	83.33%	66.67%	100%	66.67%	100%	100%	100%	80.00%	83.33%	100%	80.00%	100%	83.33%	100%	75.00%
[20, 25)	66.33%	67.35%	61.86%	68.37%	71.43%	73.47%	66.67%	70.10%	66.32%	73.20%	64.89%	75.82%	77.55%	73.47%	73.75%	73.81%
[25, 30)	74.17%	70.59%	77.31%	76.27%	79.66%	75.63%	81.74%	79.66%	81.31%	79.46%	78.76%	83.81%	78.45%	85.84%	70.53%	79.17%
[30, 35)	68.55%	79.03%	74.19%	73.39%	78.69%	66.94%	68.10%	76.86%	80.00%	73.11%	69.16%	79.09%	68.64%	76.27%	65.22%	62.65%
[35, 40)	62.50%	70.34%	70.43%	72.27%	65.25%	73.73%	62.73%	66.10%	61.90%	68.10%	55.66%	70.11%	68.81%	61.86%	62.50%	73.24%
[40, 45)	71.58%	72.16%	79.38%	76.53%	72.92%	78.57%	66.67%	75.79%	71.95%	78.02%	80.56%	73.85%	75.86%	73.49%	76.67%	74.07%
[45, 50)	76.92%	71.15%	71.15%	73.08%	75.00%	62.75%	69.05%	67.31%	65.91%	65.31%	72.50%	78.12%	80.00%	73.33%	73.33%	70.00%
[50, 55)	71.15%	76.92%	73.08%	71.15%	68.00%	78.85%	65.96%	68.00%	67.50%	71.74%	61.36%	70.97%	69.57%	76.19%	61.90%	64.00%
[55, 60)	77.78%	68.75%	72.09%	75.00%	76.09%	68.75%	68.57%	66.67%	76.47%	66.67%	60.00%	76.00%	75.00%	78.95%	57.89%	85.00%
[60, 65)	88.37%	69.77%	72.50%	84.09%	84.21%	79.07%	65.62%	76.74%	63.33%	83.78%	77.78%	82.76%	78.57%	76.92%	66.67%	78.95%
[65, 70)	80.95%	68.18%	81.82%	72.73%	63.64%	65.00%	30.00%	80.95%	41.18%	84.21%	78.57%	72.73%	62.50%	69.23%	25.00%	41.67%
[70, 75)	92.86%	78.57%	64.29%	71.43%	78.57%	100%	53.85%	92.86%	50.00%	92.31%	100%	100%	88.89%	100%	75.00%	100%
[75, 80)	100%	83.33%	66.67%	100%	66.67%	83.33%	100%	83.33%	60.00%	50.00%	40.00%	100%	100%	40.00%	100%	100%
Total	71.98%	72.22%	72.68%	74.28%	73.79%	73.65%	68.05%	73.54%	70.54%	73.97%	69.66%	77.41%	74.44%	75.00%	68.46%	72.31%

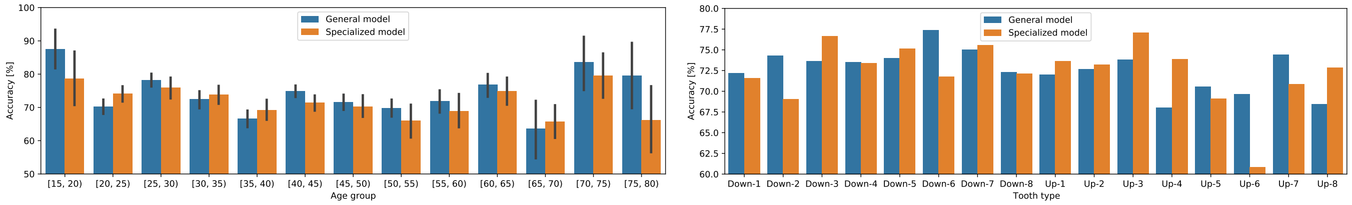

Fig. 2. **Accuracy of general and specialized models per age group and tooth type.** An overview that shows a) the accuracy of assessment for each tooth type and b) he accuracy of assessment for each age group, as well as vertical bars showing the variance. The left bar represents the accuracy of the general model, while the right bar shows the accuracy of each specialized models.

automated using deep learning. We've also shown that both specialized models and a generalized model can perform well on this task. Those models have been trained and verified on the largest dataset of individual tooth x-ray images in literature. The models have no quality requirements on the teeth. They can have illnesses, decay, or dental alterations, which makes this approach suitable for real-world usage. When controlling for changes in the structures of a tooth, experiments seem to indicate that better performance is achievable, with the accuracy reaching 85% without retraining of models. With an overall accuracy of 72.68% for the general model and 72.40% for the specialized models, without any quality requirements, and an accuracy of 85% when controlling for tooth quality, the proposed method matches human performance, shows great potential for application, and opens up a new category of damaged teeth for forensic tasks.

ACKNOWLEDGMENT

This research has been supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

We would like to thank Tea Šarić and Viktorija Sajković for lending their expertise to annotate the dataset.

References

[1] V. Saini, R. Srivastava, R. K. Rai, S. N. Shamal, T. B. Singh, and S. K. Tripathi, "Mandibular Ramus: An Indicator for Sex in Fragmentary Mandible*," *Journal of Forensic Sciences*, vol. 56, no. s1, pp. S13–S16, 2011.

[2] W. M. Krogman, "The human skeleton in forensic medicine. I.," *Postgraduate medicine*, vol. 17, no. 2, pp. A–48, 1955.

[3] K.-S. Hu, K.-S. Koh, S.-H. Han, K.-J. Shin, and H.-J. Kim, "Sex determination using nonmetric characteristics of the mandible in Koreans," *Journal of forensic sciences*, vol. 51, no. 6, pp. 1376–1382, 2006.

[4] P. C. Srivastava, "Correlation of odontometric measures in sex determination," *J Indian Acad Forensic Med*, vol. 32, no. 1, pp. 56–61, 2010.

[5] E. S. Martin, "A study of an Egyptian series of mandibles, with special reference to mathematical methods of sexing," *Biometrika*, vol. 28, no. 1/2, pp. 149–178, 1936.

[6] G. M. Morant, M. Collett, and N. K. Adyanthaya, "A biometric study of the human mandible," *Biometrika*, vol. 28, no. 1/2, pp. 84–122, 1936.

[7] A. Hrdlička, "Mandibular and maxillary hyperostoses," *American Journal of Physical Anthropology*, vol. 27, no. 1, pp. 1–67, 1940.

[8] H. De Villiers, "Sexual dimorphism of the skull of the South African Banu-speaking Negro," *South African Journal of Science*, vol. 64, no. 2, p. 118, 1968.

[9] L. T. Humphrey, M. C. Dean, and C. B. Stringer, "Morphological variation in great ape and modern human mandibles," *Journal of Anatomy*, vol. 195, pp. 491–513, Nov. 1999.

[10] A. P. Kanya, B. Kiswanjaya, B. N. Makes, and H. H. B. Iskandar, "Estimating Sex in an Indonesian Population Using the Mean Value of Eight Mandibular Parameters in Panoramic Images," *Journal of International Dental and Medical Research*, vol. 10, pp. 417–422, 2017.

[11] D. Franklin, P. O'Higgins, C. E. Oxnard, and I. Dadour, "Determination of Sex in South African Blacks by Discriminant Function Analysis of Mandibular Linear Dimensions: A Preliminary Investigation Using the Zulu Local Population," *Forensic Science, Medicine and Pathology*, vol. 2, no. 4, pp. 263–268, 2006.

[12] D. H. Badran, D. A. Othman, H. W. Thnaibat, and W. M. Amin, "Predictive Accuracy of Mandibular Ramus Flexure as a Morphologic Indicator of Sex Dimorphism in Jordanians," *International Journal of Morphology*, vol. 33, pp. 1248–1254, Dec. 2015.

[13] E. Giles, "Sex determination by discriminant function analysis of the mandible," *American Journal of Physical Anthropology*, vol. 22, pp. 129–135, June 1964.

[14] S. R. Loth and M. Henneberg, "Mandibular ramus flexure: A new morphologic indicator of sexual dimorphism in the human skeleton," *American Journal of Physical Anthropology*, vol. 99, pp. 473–485, Mar. 1996.

[15] M. Dayal, M. Spocter, and M. Bidmos, "An assessment of sex using the skull of black South Africans by discriminant function analysis," *HOMO*, vol. 59, pp. 209–221, July 2008.

[16] A. P. Indira, A. Markande, and M. P. David, "Mandibular ramus: An indicator for sex determination - A digital radiographic study," *Journal of Forensic Dental Sciences*, vol. 4, no. 2, pp. 58–62, 2012.

[17] M. Marinescu, V. Panaitescu, and M. Rosu, "Sex determination in Romanian mandible using discriminant function analysis: Comparative results of a time-efficient method," *Romanian Journal of Legal Medicine*, vol. 21, pp. 305–308, Dec. 2013.

[18] T. Bhagwatkar, M. Thakur, D. Palve, A. Bhondey, and Y. Dhengar, "Sex Determination by Using Mandibular Ramus - A Forensic Study," *Journal of Advanced Medical and Dental Sciences Research*, vol. 4, no. 2, p. 6, 2016.

[19] K. N. Maloth, V. K. R. Kundoor, S. S. L. P. Vishnumolakala, S. Kesidi, M. V. Lakshmi, and M. Thakur, "Mandibular ramus: A predictor for sex determination-A digital radiographic study," *Journal of Indian Academy of Oral Medicine and Radiology*, vol. 29, no. 3, p. 242, 2017.

[20] T. Nagaraj, L. James, S. Gogula, N. Ghouse, H. Nigam, and C. K. Sumana, "Sex determination by using mandibular ramus: A digital radiographic study," *Journal of Medicine, Radiology, Pathology and Surgery*, vol. 4, no. 4, pp. 5–8, 2017.

[21] A. Alias, A. Ibrahim, S. N. A. Bakar, M. S. Shafie, S. Das, N. Abdullah, H. M. Noor, I. Y. Liao, and F. M. Nor, "Anthropometric analysis of mandible: an important step for sex determination," *La Clinica Terapeutica*, vol. 169, pp. e217–e223, Oct. 2018.

[22] D. Milošević, M. Vodanović, I. Galić, and M. Subašić, "Estimating biological gender from panoramic dental x-ray images," in *2019 11th international symposium on image and signal processing and analysis (ISPA)*, pp. 105–110, IEEE, 2019.

[23] 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," *Sensing and Imaging*, vol. 21, no. 1, pp. 1–11, 2020.

[24] 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*, vol. 4, no. 1, pp. 289–92, 2013.

[25] C. Capitaneanu, G. Willems, R. Jacobs, S. Fieuws, and P. Thevissen, "Sex estimation based on tooth measurements using panoramic radiographs," *International journal of legal medicine*, vol. 131, no. 3, pp. 813–821, 2017.

[26] C. Capitaneanu, G. Willems, and P. Thevissen, "A systematic review of odontological sex estimation methods," *The Journal of forensic odonto-stomatology*, vol. 35, no. 2, p. 1, 2017.

[27] ISO-3950:2016, "Dentistry — Designation system for teeth and areas of the oral cavity," standard, International Organization for Standardization, Geneva, CH, Mar. 2016.

[28] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A Survey on Deep Transfer Learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018*, Lecture Notes in Computer Science, (Cham), pp. 270–279, Springer International Publishing, 2018.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, Curran Associates, Inc., 2017.

[30] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization.," *Journal of machine learning research*, vol. 13, no. 2, 2012.

[31] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[32] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, (San Francisco, California, USA), pp. 4278–4284, AAAI Press, Feb. 2017.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[35] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Honolulu, HI), pp. 1800–1807, IEEE, July 2017.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv:2103.00020 [cs]*, Feb. 2021. arXiv: 2103.00020.

[37] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015. Publisher: Elsevier.

[38] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 1958.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[40] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.