A Comprehensive Exploration of Neural Networks for Forensic Analysis of Adult Single Tooth X-Ray Images

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

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 June 2022, accepted 28 June 2022, date of publication 4 July 2022, date of current version 11 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3187959

RESEARCH ARTICLE

A Comprehensive Exploration of Neural Networks for Forensic Analysis of Adult Single Tooth X-Ray Images

DENIS MILOŠEVIĆ[1], MARIN VODANOVIĆ[2], IVAN GALIĆ[3,4], AND MARKO SUBAŠIĆ[1], (Member, IEEE)

[1]Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia
[2]Department of Dental Anthropology, School of Dental Medicine, University Hospital Centre Zagreb, University of Zagreb, 10000 Zagreb, Croatia
[3]School of Medicine, University of Split, 21000 Split, Croatia
[4]University Hospital of Split, 21000 Split, Croatia

Corresponding author: Denis Milošević (denis.milosevic@fer.hr)

This work was supported in part by the European Regional Development Fund (DATACROSS) under Grant KK.01.1.1.01.0009; and in part by the Croatian Science Foundation, Croatia, under Project IP-2020-02-9423.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the School of Dental Medicine University of Zagreb, and by the Ethics Committee of the Faculty of Electrical Engineering and Computing University of Zagreb.

ABSTRACT Determining the demographic characteristics of a person post-mortem is a fundamental task for forensic experts, and the dental system is a crucial source of those information. Those characteristics, namely age and sex, can reliably be determined. The mandible and individual teeth survive even the harshest conditions, making them a prime target for forensic analysis. Current methods in forensic odontology rely on time-consuming manual measurements and reference tables, many of which rely on the correct determination of the tooth type. This study thoroughly explores the applicability of deep learning for sex assessment, age estimation, and tooth type determination from x-ray images of individual teeth. A series of models that use state-of-the-art feature extraction architectures and attention have been trained and evaluated. Their hyperparameters have been explored and optimized using a combination of grid and random search, totaling over a thousand experiments and 14076 hours of GPU compute time. Our dataset contains 86495 individual tooth x-ray image samples, with a subset of 7630 images having additional information about tooth alterations. The best-performing models are fine-tuned, the impact of tooth alterations is analyzed, and model performance is compared to current methods in forensic odontology literature. We achieve an accuracy of 76.41% for sex assessment, a median absolute error of 4.94 years for age estimation, and an accuracy of 87.24% to 99.15% for tooth type determination. The constructed models are fully automated and fast, their results are reproducible, and the performance is equal to or better than current state-of-the-art methods in forensic odontology.

INDEX TERMS Age estimation, sex assessment, tooth type determination, tooth numbering, convolutional neural network, deep learning, forensic odontology, dental x-ray, image processing, medical image analysis.

I. INTRODUCTION

Dental remains are regularly the only evidence left in the wake of violent crime and disaster scenarios. Forensic odontology specializes in the analysis of dental remains to collect evidence and help identify the victims in those situations. Those methods can also be used in other fields - for example, archaeology, where those methods are used to study the demographic makeup of a population through time, where written records are missing or unreliable. Age and sex are expressed in the human body and the dental system through different indicators, which are primarily based on develop-

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

ment, maturity, genetics, decay, and wear. No significant changes happen to those indicators post-mortem, allowing forensic experts to determine a person's demographic information at their time of death. In younger individuals, particularly children, developmental indicators alone are sufficient to determine the age within the range of a few months. In adults, after all elements have fully matured, developmental indicators cannot be used to determine the age past a certain age, making the forensic tasks much more challenging to solve. Likewise, indicators of sexual dimorphism decay with age and wear, making it harder to distinguish the different sexes in older individuals.

Current state-of-the-art methods for these problems in forensic odontology are based on manual measurements and reference charts. Those reference charts need to be created for each population, and they need to be regularly updated as changes to a population occur. Measurements are taken manually, which additionally can introduce human error. Taking those measurements also takes time, as an expert in the field has to measure multiple indicators per human remain. For methods that work with individual teeth, it is necessary to determine the type of the tooth to properly apply current state-of-the-art methods, which can be an issue if only individual, displaced teeth are available.

In this study, we examine the use of convolutional neural networks for the task of age estimation, sex assessment, and tooth type determination. While we have done some preliminary research [1]–[3], this study solely focuses on individual tooth x-ray images. Individual tooth images naturally contain less information, as the jaw bone and other surrounding tissue that can aid in age estimation is missing. This holds true both for sex assessment and age estimation. We have collected one of the most extensive datasets of individual tooth x-ray images in literature, and we uniquely have a test set with annotated tooth status information. To develop automated, accurate, and fast image analysis models, we have trained and evaluated 1570 models which use state-of-the-art architectures as their feature extractor paired with attention, thereby leveraging years of vision-model research and evaluating the applicability of modern, highly complex models for single-tooth forensic odontology tasks. In addition to tackling individual tooth x-ray images with direct sex assessment and age estimation, which is a topic not well explored for automated image analysis in forensic odontology, we have developed accurate models for tooth type determination for four different classification systems, which can be used in conjunction with classical forensic odontology methods. We also show that models specialized per tooth type, which is common in current forensic odontology methods, and multi-task models, which are common in medical image analysis, do not perform better than general, single-task models. Attention was also tested, as current image analysis literature reports performance improvements, and we show that in our use-case, attention models underperformed. In a departure from current forensic odontology research, our models are trained to handle non-perfect teeth. With the test set having tooth status annotations, we have done a detailed analysis of the impact and influence of alterations on model performance. Finally, we compare our results with current state-of-the-art manual and automated forensic odontology methods.

This paper is structured as follows. First, in Section II we present the current state of forensic odontology for age estimation, sex assessment, and tooth type determination in literature. In Section III, we give an overview of our dataset. After that, in Section IV, we explain our methodology, which includes an explanation of model construction, model training, hyperparameter search, and fine-tuning. We conclude this study with a showcase of the achieved results (Section V) and the analysis of those results together with a comparison to the current state-of-the-art (Section VI).

II. RELATED WORKS

Forensic odontology has a long history, particularly for age estimation and sex assessment tasks. Tooth type determination is less storied, as the question of precise tooth type arose with newer scientific methods. Early works back in 1837 show that dental indicators are helpful for age estimation [4]. Research has shown that the sex of a person can be determined with 100% accuracy when the entire skeleton is present [5], which is sadly a luxury in the field of forensics. Post-mortem changes, decay, and decomposition are the slowest for dental tissue, and the dental system is one of the most resilient parts to external force, making those remains the prime candidate for forensic analysis [6]. Early work has a major drawback – they use a destructive approach that either requires the destruction of remains (and therefore evidence), or it requires the removal of a tooth from a living person, which is unacceptable [7]. With the rise of radiographic imaging, destructive approaches have fallen out of favor, and radiography-based approaches outperform their destructive counterparts in performance [8]–[12]. As noted in more recent research, forensic odontology methods are developed and analyzed from the point of view of perfectly healthy teeth, namely ''intact mandibles, without any pathology, loss of mandibular molars, or anomalous molars and teeth'' [13].

Age estimation research in forensic odontology was first heavily researched during the introduction of child labor protection laws, as children were forced to work and those laws were frequently avoided. Early work describes development charts in either ten [14] or eight [15] stages, which shows what the expected age of a child with a given development status would be. There have also been attempts at modeling a linear relationship between some orthopedic measurements and age [16]. These approaches worked, and research later confirmed that the dental development schedule is defined strictly genetically, which is why age in children can be estimated with an error measured in months [17]–[19]. This highlights an observation – age estimation in children (up to 17/18 years) is a solvable problem with guaranteed low errors. Most research on age estimation in children confirms this, with their error being measured in months rather than years [14]–[16], [20].

This is not the case for adult samples, as all development ceased at that point. In modern forensic odontology for adults, three methods can be considered foundational. Most research in ''classical'' forensic odontology is based on one of those methods, with either minor improvements or the determination of optimal parameters for those methods given a particular population. The first foundational study [21] describes a model for adult age estimation based on the measurements of the dental pulp cavity. They discovered that secondary dentine deposits slowly reduce the dental pulp cavity with age, and this correlation can be used to estimate the age. The second foundational study is based on the tooth-coronal index [22], which again correlates the coronal pulp cavity with chronological age [23]. The third and nowadays most used method [24] establishes a model for age based on measurements of a single-rooted tooth, specifically the single-rooted maxillary right canine. Those methods formulate a linear model and then determine the optimal parameters for age estimation based on their data. Population specific research shows that those models hold true, but that different populations have a different set of optimal parameters [25]–[31]. This implies that those models not only have to be tailored to specific populations, but also have to be maintained and updated over time as systemic changes in the population happen due to advances in dental health. A systematic review of dental age estimation research confirms that modern forensic odontology age estimation research can be categorized into those three studies [32].

Sex assessment, too, has significantly advanced in recent times. Most methods are based on mandibular parameters, for early work to more recent, and therefore require the entire jaw to properly assess the sex [33]–[37]. Moreover, while the jaw and dental tissue are very resilient to decay and external force, oftentimes, teeth are the only evidence left behind [6]. Modern forensic odontology literature suggests that sex assessment from individual teeth is not independently feasible and should only be used in conjunction with other methods to form a strong consensus [38], [39]. Specifically, [38] claims that single-tooth sex assessment methods cannot exceed 80% accuracy. A systematic review of sex assessment in forensics literature [40] shows that methods either use the significant parts of the skeleton, biochemical analysis of remains, or dental remains (odontometric). For dental remains, measurements are taken directly, from a cast, or from radiographic imaging. While reported results collected in the systematic review vary in range, population, age, and sample size, most adhere to the findings of [38], especially those studies with higher Quality Assessment Scores. Specifically for sex estimation based on x-ray images, one study achieved an accuracy between 68% and 80% [39] with a sample size of 200 by using measurements of mandibular teeth, and another study [41] achieved an accuracy of 83.3% with a sample size of 60 by using diagonal measurements. While the systematic review is from 2017, newer studies use a similar approach and achieve similar results [42]. Similarly to age estimation, the parameters of the established models can be adjusted to improve accuracy on different populations [43]–[50].

Deep learning has revolutionized the field of computer vision. This revolution has not gone unnoticed in medical research, with deep learning being used to create tools that improve healthcare outcomes and diagnosis success [51]. These methods are slowly reaching dentistry, too, with early research showing promise with tooth detection [52], segmentation [53], [54], and a multitude of other tasks, as can be seen in a recent review study [55]. There are also approaches that explore the usage of optical coherence tomography for dental diagnosis, which can provide tooth status information in an noncontact and noninvasive manner [56]. Another approach estimates dental parameters directly, albeit for orthodontic assessment [57]. Still, this highlights the trend and need for automation of manual measurements and estimations in dentistry and dentistry-related fields. Some research into deep learning has been done for forensics, too, with studies developing models for estimating the age of entire skeletons [58]. Other approaches tackle human identification directly by matching panoramic dental x-ray images using neural networks and attention [59].

Age estimation from dental remains has also been tackled in recent studies. A study with a custom-designed neural network and a dataset mainly consisting of child and adolescent panoramic dental x-ray images achieves a mean absolute error of 2.84 ± 3.75 years [60]. Another study evaluated the feasibility of deep learning for the classification of archaeological samples from the 11th century into age groups, achieving an accuracy of 73% [61]. In our previous study, we designed models for age estimation in adults based on panoramic dental x-ray images, achieving a mean absolute error of 3.96 years and achieving mean absolute errors in the range between 6.30 to 8.68 years on a family of models adapted for single tooth images [2]. Approaches that classify samples into age groups instead of estimating the age directly are also becoming more popular, with studies focused on the first molar [62] achieving an accuracy between 89.05% to 90.27% with their three-class approach (''children and adolescents (ages 0–19), young adults (ages 20–49), and older adults (age $>$ 50 years)''). Studies classifying child samples [63] into Stages D to H of the Demirjian developmental groups [15], achieving an accuracy of 82.50%. A new study combines the classical approach of manual measurements of indicators with the modeling capabilities of deep learning [64], achieving an error between 2.34 and 4.61 months for their child and adolescent samples.

Sex assessment has also been explored. In our earlier study, we have constructed a sex assessment model based on panoramic dental x-ray images, which achieves an accuracy of $96.87\% \pm 0.96\%$ [1]. Another study evaluated their deep learning model, achieving an accuracy between 90% and 96% for adult samples and an accuracy of 84% for child and adult samples cumulatively [65]. A multiple feature fusion model was later developed, achieving an accuracy of $94.6\% \pm 0.58\%$ on their dataset of 19976 panoramic

dental x-ray images [66]. Our initial study for the assessment of sex based on individual tooth x-ray images [67] achieves an accuracy of 72.68% on all samples, and preliminary results on a small subset of clinically perfect teeth achieved an accuracy of 85% with the same model. With a sample size of 1000 panoramic dental x-ray images, a study employing ResNet50 [68] and denoising claims an accuracy of 98.27% [69].

Tooth type determination is often done as part of the output of tooth detection models. For example, [52] is primarily concerned with tooth detection. However, they still achieve an accuracy of over 90% with their modified variant of AlexNet in a three-class approach (canine & incisors, premolars, and molars), with a sample size of 100 panoramic dental x-ray images. Another example is [70], where the authors use projection profile analysis and achieve an overall accuracy of 92.54% classifying teeth into four types. In a similar vein, [71] use Faster R-CNN with a dataset of 1250 panoramic dental x-ray images to achieve a tooth type determination accuracy between 71.5% and 91.7%. In our preliminary study, we achieved an accuracy between 91.13% and 97.83% for individual tooth x-ray images, for the 4, 8, 16 and 32 class approach [3].

This study is a continuation and exhaustive examination that started with our previous research, which focuses on individual tooth images. Our age estimation study [2] was primarily focused on age estimation from panoramic dental x-ray images. In that study, individual tooth x-ray images were evaluated too, and while the model was trained on individual tooth x-ray images, its hyperparameters were optimized for panoramic dental x-ray images. In this study, the sole focus is on individual tooth x-ray images. While the approaches to model discovery have similarities, this study extends the search with more experiments. Those experiments work solely with individual tooth x-ray images instead of panoramic dental x-ray images, and the basic grid search strategy has been extended with random search. Age estimation from panoramic dental x-ray images and individual tooth x-ray images are similar tasks applied in different scenarios. Estimation from individual teeth performs worse than the panoramic dental x-ray counterpart due to significantly less information in the input image. Additionally, in contrast to [2], this study not only performs hyperparameter tuning specifically for individual tooth x-ray images but also takes into account age estimation of imperfect, damaged, or otherwise altered teeth and provides a detailed analysis of the impact tooth alterations have on age estimation performance. For sex assessment, our preliminary study has been extended with the evaluation of a broader range of models with more experiments [67]. Finally, our preliminary study for the determination of tooth type [3] has in this study been fully extended, constructing a model on a much broader toolset, and using a significantly larger dataset. This study also improves upon the analysis of all our previous work with the examination of the impact of tooth alteration on prediction performance.

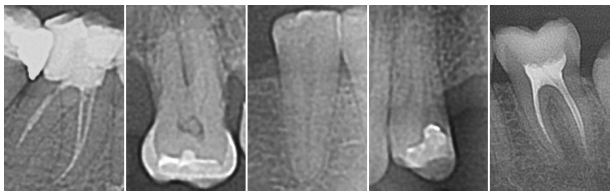


FIGURE 1. Randomly selected samples of individual teeth. Individual teeth are clipped from the panoramic dental x-ray image by their bounding box.

III. DATA

Our dataset consists of 86495 individual tooth images originating from 2899 panoramic dental x-ray images. The ratio of female to male samples is 59.03% to 40.97%. The dataset contains only adult samples, with ages ranging from 19 to 86 years. The samples are moderately biased towards younger samples, with the average age being 38.41 years. Some examples of individual tooth images and their possible alterations can be seen in Figure 1. The distribution of samples across age and sex can be seen in Figure 2.

The samples 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.

Dentistry experts have provided annotations of individual tooth position and type for each tooth in those panoramic dental x-ray images. The position is given as a bounding box around each tooth. Tooth type information is provided using the FDI dual notation system (ISO-3950 notation) [72], which is the standard used in dentistry and forensic odontology. Additionally, for a subset of 7630 images, status annotations were provided. A tooth status annotation contains the information about any alterations, namely tooth decay, fillings, root canal fillings, crowns, bridges, tooth germs, leftover roots, dental implants, missing teeth, and crowns. The distribution of alterations is not uniform, with missing teeth, fillings, root canal fillings, and tooth decay being the most numerous. To avoid inconclusive or misleading results due to the small sample size, we have grouped tooth germs, leftover roots, dental implants, bridges, and crowns into one category, referred to as ''Other'' in this study. A tooth can have one alteration, multiple alterations, or no alterations at all. In our dataset, 66.37% have no alterations, 27.93% of all teeth have at least one alteration, 5.10% have at least two alterations, 0.59% have three alterations, and 0.01% have four alterations.

From a technical point of view, the raw x-ray readings are converted into 8-bit images in JPEG format. As the samples are taken with different orthopantomographs, the resulting images have a resolution of 1127 px to 3260 px in width dimension and 553 px to 1536 px in the height dimension. Individual tooth images are then clipped out of the panoramic dental x-ray image and stored separately. Teeth come in different shapes and sizes, with premolars and molars

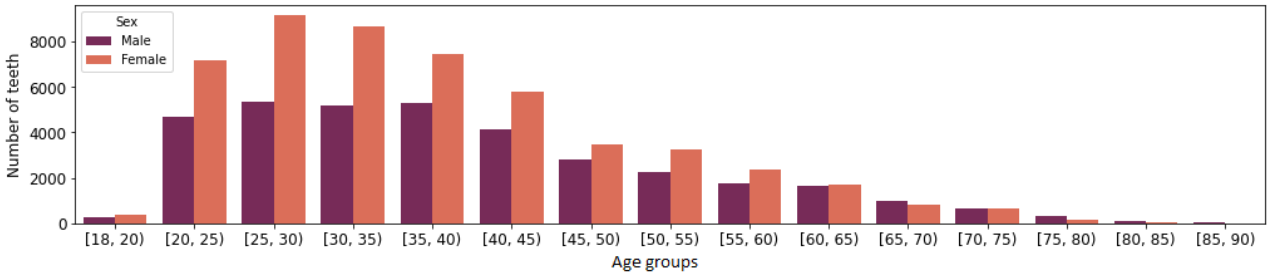


FIGURE 2. Distribution of samples in dataset per age and sex. A decrease in samples with age can be observed, as well as a slight bias towards female samples (59.03% to 40.97% overall female to male ratio). The difference in sample count per sex diminishes with age.

having an aspect ratio around 1:1, while incisors and canines have an aspect ratio heavily in favor of height. Nevertheless, no individual tooth exceeds a size of 528 px in either the width or height dimension.

Given that natural size limitation, for training each individual tooth image is placed in a 528 px by 528 px empty image randomly, in such a manner that the entire tooth is visible. This ensures that no part of the tooth is left out and that the original information of the image is not distorted by rescaling, while at the same time allowing for additional variance in the dataset to prevent overfitting. We have chosen the image background to be black, seeing that our previous experiments show that the choice of background does not influence performance [1]. For evaluation, tooth images are placed centrally in the 528 by 528 px canvas, ensuring that the validation and test set are always the same, and that results are comparable between epoch, and between experiments. In our previous studies [1], [2] we applied no procedural variation, and images were resized to 512 by 512 px.

The target data (age, sex, tooth type) is transformed into the appropriate form required for training. For classification tasks (sex, tooth type), the annotation is encoded as a one-hot vector. For age estimation, which is a regression task, the age is represented as a floating-point number of years. This number is, for our dataset, in the range of 19 to 86. Preliminary experiments have shown that normalizing the age to a number between 0 and 1 does not influence performance, and we have therefore chosen not to rescale the ground-truth data.

Tooth type classification presented another challenge. While FDI dual notation system (ISO-3950) [72] is the standard for tooth numbering, research in forensic odontology uses a few different classification approaches. As per ISO-3950, each tooth in adults is identified by two numbers - the quadrant it resides in and its position within that quadrant, resulting in 32 classes. Another numbering system used is the differentiation between maxillary and mandibular teeth and their position within their quadrant, without a left-right differentiation [21], [24]. For example, by this classification system both maxillary canines are considered the same class. This results in 16 classes total. The differentiation between maxillary and mandibular teeth is sometimes dropped, resulting in an 8-class system. Finally, teeth can be differentiated into four basic types: incisors, canines, premolars, and molars. As every choice has its merits, and for the sake of completeness and comparability, we have chosen to design models and evaluate the performance of every mentioned classification system.

The dataset is separated into multiple parts. One part is the training dataset, which is used to learn the parameters of the deep model. It consists of 80% of the data. The remaining data is separated into a validation and a test dataset, which comprise 11.18% and 8.82% of the data, respectively. The validation set is used to evaluate the feasibility of a deep learning approach, be it the model used or any hyperparameter involved in the experiment. The test set is a subset of the data evaluated only once, after the best models have been selected and fine-tuned. These results are used to report the final performance, as this data could not have influenced any research decision and thereby biased the results. The unusual split between validation and test is due to the availability of annotated status data. There is no overlap in the samples between sets. Annotations of tooth status are limited in number, as their creation requires experts in the field, and it requires a lot of time. We have thus decided to use all 7360 samples with tooth status annotations as the test set, which totals in 8.82% of the dataset. As we have a very limited amount of status data, we cannot verify, without a loss of data, that the distribution of alterations is equal in the train, validation, and test set. However, given the sample size, the measured impact of alterations should not be influenced by this. A few pairs of images belong to the same person but are taken some time apart. Those pairs have all been placed in the train set to avoid any kind of bias or data leakage in the results.

A detailed overview of tooth status annotations is given in Table 1.

IV. METHOD

In this study we examine the performance of deep learning models for the forensic tasks of age estimation, sex assessment and tooth type determination from x-ray images of adult individual teeth. Those x-ray images are extracted from panoramic dental x-ray images, and then processed as described in Section III. A deep convolutional neural network model with optional attention is used as the base for

TABLE 1. Detailed overview of data samples per age group. While the entire dataset consists of 86465 individual tooth images, the test set with status annotations is a subset of 7630 images. As can be seen, the number of samples, and therefore status annotations, diminishes with age.

Age group	Female	Male	Missing	Root canal fillings	Filling	Tooth decay	Other alterations
[18, 20)	376	255	0	0	0	0	0
[20, 25)	7187	4709	50	56	207	37	11
[25, 30)	9149	5327	133	85	414	89	23
[30, 35)	8635	5170	100	43	313	44	17
[35, 40)	7437	5272	76	60	261	32	45
[40, 45)	5803	4123	135	75	251	57	57
[45, 50)	3441	2805	24	10	29	8	5
[50, 55)	3271	2275	38	26	76	7	46
[55, 60)	2348	1777	44	7	38	12	7
[60, 65)	1715	1623	0	0	0	0	0
[65, 70)	791	981	0	0	0	0	0
[70, 75)	668	665	0	0	0	0	0
[75, 80)	180	331	0	0	0	0	0
[80, 85)	55	96	0	0	0	0	0
[85, 90)	0	30	0	0	0	0	0
Total	**51056**	**35439**	**600**	**362**	**1589**	**286**	**211**

exploration. We have invested 14076 hours of GPU compute time to search for the best performing model hyperparameters. We evaluate the difference between models specialized by tooth type and ''general'' models (models that estimate on any single tooth image), as well as the performance of so-called joint models - models that perform two or more of the forensic tasks simultaneously. As shown later, the joint models did not perform as well as expected, and this study is therefore focused on single task models. The best performing models are then fine-tuned, and then evaluated with metrics conventionally used in forensic odontology studies. The results are then analyzed per age group, tooth type and tooth alterations to determine any weak points or biases of the models. A visual summary of the entire research process an be seen in Figure 3.

A. MODELS

All models developed during this study follow the same ''meta-model.'' Each model consists of a feature extractor, followed by a 1×1 convolutional layer and optional attention mechanism, converted into a feature vector by flattening or global average pooling, ending in two fully-connected layers. All activation functions in the models are ReLU [73], except for sex assessment and tooth type determination, which use the softmax activation for the final layer, as they are classification tasks.

The feature extractor is an architecture proven to work well in literature. Specifically for this study, the following architectures were evaluated: DenseNet201 [74], Inception-ResNetV2 [75], ResNet50 [68], VGG16, VGG19 [76] and Xception [77]. Transfer learning was taken into consideration [78]. Given the large size of our dataset, we have done preliminary experiments to determine if transfer learning is beneficial. Results have shown that there is no significant difference between transfer learning and learning from random initialization if the feature extractor parameters are updated during training, nor is there a difference in time to convergence. In [2], we have shown that transfer learning does improve overall performance on full panoramic dental x-ray images. We assume that the difference arises from the size of the dataset. The feature extraction architectures have a high enough capacity to solve the posed problems, but the domain of individual tooth x-ray images is too different from the domain of natural images found in ImageNet, leading to a position where the pretrained weights are ''effectively random'' for this specific use case.

The following 1×1 convolutional layer is used to change the number of channels in the final feature map. While the added value of this layer is diminished by the fact that the feature extractor weights are updated during training (contrasted to [2]), its presence still allows for an additional dimension in the hyperparameter space that significantly impacts the performance of the model. The 1×1 convolutional layer is followed by an optional attention mechanism [79]. In this case, optional means that the presence of the attention mechanism is a hyperparameter.

The final two fully-connected layers are the final estimator of the model. The size of the first fully-connected layer is a hyperparameter, and it is tuned specifically for each task. The size of the second fully-connected layer is determined by the output shape of the task target. For age estimation, the last fully-connected layer is of size 1, for sex assessment it is of size 2, and for tooth type determination it is either size 4, 8, 16, or 32 depending on the classification system used.

In previous research, we have evaluated the applicability of so-called specialized models. Those models are specialized in the sense that they process only images of a particular tooth type. The motivation behind this was two-fold. We wanted to test the assumption that different tooth types have a different enough morphology that the extracted features might conflict and thereby decrease performance. Current state-of-the-art forensic odontology research also defines different parameters for different tooth types, as seen in Section II, further reinforcing the motivation to evaluate this approach. In this

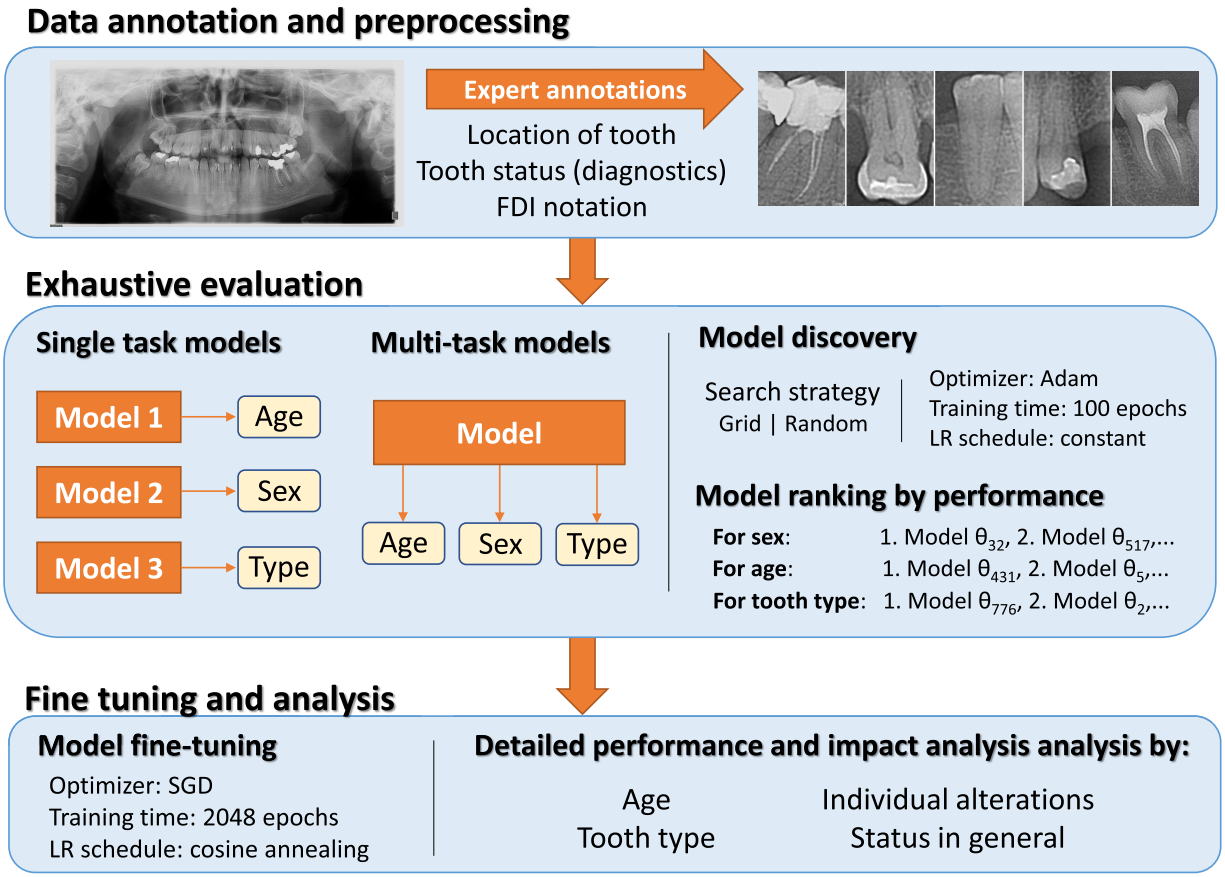


FIGURE 3. A summary of the overall research process. The dataset of individual tooth images is prepared from expert annotations of panoramic dental x-ray images, with a subset of images having tooth status annotations. A wide search space of model constructions is exhaustively evaluated using grid and random search. The best models are then fine-tuned, and then carefully analyzed in general, and per different clinically important properties.

study, with a large dataset and additional annotation about tooth status, we determined that this specialization is unnecessary and, in some cases, decreases overall performance.

In addition to the in-depth exploration of deep learning models for the individual forensic tasks, preliminary experiments have been done to test some alternative approaches. To further explore the idea of shared features, we have evaluated the feasibility of multi-task models. Multi-task models are similarly structured to the models already mentioned, with one convolutional feed-forward network for feature extraction followed by a two-layer fully-connected subnetwork. This way, the model is able to discover shared features across the tasks that might either enhance the performance or at least reduce the number of computations required. Two different variants have been evaluated for the fully-connected part. One variant shares one intermediate fully-connected layer for both decision layers, while the other has independent intermediate fully-connected layers. Direct assistance from the other demographic information was also evaluated. A series of experiments were performed where the model would estimate the age based on the x-ray image of the tooth and the sex of the person. However, those experiments showed no improvement in the best case and a decrease in performance in general.

While there is a huge variablity of model structures, the overall ''meta-model'' structure consists of 4 functional modules: feature extractor subnetwork, feature map depth scaling, attention, and the fully-connected subnetwork. The overall structure of these model variants are given on the example of what was evaluated as the best model, which uses VGG16 as the feature extractor Figure 4.

B. MODEL TRAINING

Models can be trained in two different regimes. One is a fast approach, used to get results close to the best performance the model can achieve. This includes a higher effective learning rate and fewer epochs to train. The other approach is ''slow-and-steady,'' where a model is given enough time to carefully adjust its weights to achieve optimal performance. The models are given significantly more time to train, and the effective learning rate is lower. An in-depth description of the fast approach can be seen in Section IV-C, and for the slow-and-steady approach in Section IV-D.

Different loss functions are used for different tasks. Sex assessment and tooth type determination are classification tasks, and the loss function is categorical cross-entropy (CCE). Age estimation is a regression task, and mean square

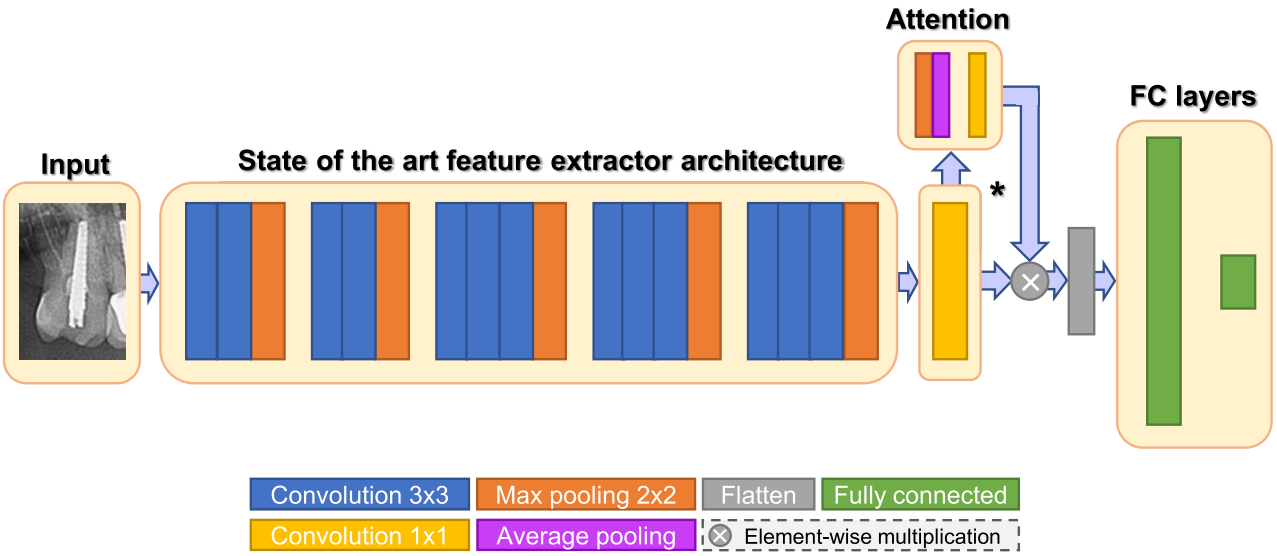


FIGURE 4. Overall model structure, demonstrated with the feature extractor architecture of the on the best performing model. All models consist of the same four functional modules: feature extractor subnetwork, feature map depth scaling (marked with * in the figure), attention [80], and the fully-connected subnetwork. Averege pooling is using the same kernel size as max pooling (2 × 2). The figure demonstrates this structure on using the single task variant and VGG16 as the feature extractor, as those performed best. Attention is show for demonstration purposes, and did not feature in the final models.

error (MSE) is used as the loss function. For the multi-task model, the respective loss functions are used, but they are weighted and summed by individual hyperparameters. For example, the total loss to train the multi-task model for sex assessment and age estimation would be $\mathcal{L} = CCE(x, y) + \lambda MSE(x, y)$. The hyperparameter λ has been exhaustively tested in the range [0, 1] for the multi-task model. As single task models outperformed multi-task models, we do not report multi-task model results, nor the optimal λ choice.

C. HYPERPARAMETER SEARCH

Hyperparameters are parameters of the deep learning model that cannot be learned during training. Research has shown that hyperparameter optimization plays a critical role in optimizing deep learning models [81]. We employ two different search strategies in this research - grid search and random search. Both have their benefits and drawbacks. Grid search [82] was used for age estimation and tooth type determination, while sex assessment used random search [83].

Grid search evaluates a set point of equidistant points, while random search evaluates a set of randomly-uniformly sampled points. Both methods search for the best-performing model in a limited search space. To get a good overview of the entire search space, grid search is the better approach. It allows for a detailed examination and discovery of well-performing subspaces. On the other hand, it is extremely computationally expensive to evaluate such a high amount of models.

Random search too evaluates points in a limited space, but the points are chosen randomly. While that subspace may not contain the global optimum of the solution, it will contain solutions of different efficacy. Looking at this situation from a purely probabilistic point of view, the probability of finding a point within a certain percentage difference range in relation to the best possible solution in that subspace depends on the number of sampled points. This assumes that small differences in hyperparameters result in small differences in performance, which for neural networks holds true [81]. Given these assumptions, to achieve a probability of p_2 that at least one point is within p_1 tolerance range of the best solution, $1 - (1 - p_1)^n > p_2$ holds true. For $p_1 = 0.05$ and $p_2 = 0.95$, $n \geq 60$.

Bayesian optimization was considered, but the sequential nature of that approach made it infeasible for us to use. Both mentioned search strategies have the benefit of points being generated independently. This allows for the experiments to be run in parallel, which drastically decreases the total real-time required.

In our initial studies [1]–[3], [67] some form of hyperparameter search was done, either directly or indirectly. The knowledge about the search space and sensible parameter ranges were learned from our initial study on sex assessment from panoramic dental x-ray images [1]. For age estimation, in addition to running grid search over the entire search space like in [2] for individual teeth x-rays only, we have evaluated an additional 381 experiments with random search. For sex assessment, we have evaluated an additional 256 experiments with grid search to complete the random search we have done in [67]. Tooth type determination in [3] was done on a smaller dataset, and the search consisted of 64 experiments. In this study, tooth type determination experiments were extended with a random search of 256 experiments and a random search

of 128 experiments. Additionally, for the multi-task model, over 549 experiments were conducted, which resulted in a total of 7796 hours of GPU time. As those models tended to be bigger and more complex, their average execution time per experiment was around 14 hours.

All these experiments are run with the fast training approach. The effective learning rate is higher than during fine-tuning, the models are trained for fewer epochs, and a faster optimizer was used. To be precise, we used the Adam optimizer [84] with a learning rate of $3.24 \cdot 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, $\lambda = 0$, and with training being limited to 100 epochs for all three tasks. The number of epochs is determined empirically, as preliminary results show that convergence is reached in 40 to 60 epochs.

D. FINE-TUNING

Research shows that adaptive methods generalize worse than stochastic gradient descent [85]. Additionally, research also shows that different learning rate schedules can improve performance when training with stochastic gradient descent. In this study, we chose cosine annealing with warm restarts [86] as the learning rate schedule. In preliminary experiments, results for the same models were consistently better when trained with this approach. In essence, that means that for each epoch, the learning rate is determined with the following expression:

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

Here, 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.

The fine-tuning of the best performing candidate models for each task was done by using Stochastic Gradient Descent (SGD) and cosine annealing with warm restarts as the training method, trained for 2048 epochs with $\eta_{min} = 10^{-7}$, $\eta_{max} = 10^{-3}$, and $T_i = 50$. The number of epochs is determined empirically, as preliminary results have shown that convergence is reached within 600 to 800 epochs for panoramic dental x-ray images, and within 1000 to 1600 epochs for individual tooth x-ray images.

E. EVALUATION APPROACH

The results presented in this study are the model evaluations on the test set, which contains detailed tooth status annotations. This hold-out set was not used during any experiments and was used only for the final evaluation of the fine-tuned models. The test set is made of all the teeth images with annotated interventions that we could obtain, which totals to 7630 images.

For age estimation, mean and median absolute error is used as the evaluation metric. The mean absolute error shows the statistically expected value of the error, while the median absolute error gives insight into the model's performance without the influence of extreme outliers. Both metrics are regularly used in forensic odontology research related to age estimation.

Both sex assessment and tooth type determination accuracy are classification tasks. To properly evaluate their performance, accuracy is used as the evaluation metric. Accuracy describes the ratio of correctly classified samples and the total number of samples. The F1 score is also calculated, which allows further insight into the performance with regard to any potential data imbalance. It is calculated as the harmonic mean of precision and recall.

In addition to the general performance of the model, in this study, we additionally examine the performance of each model per age group, per tooth type, and per tooth status, as well as the difference in performance on the complete test set and the subset of teeth that have no imperfections. By examining the model performance under all these different conditions, we can determine the performance's robustness and identify the developed models' weaknesses and capabilities.

V. RESULTS

In this study, we have developed models for the tasks of age estimation, sex assessment, and tooth type determination based on x-ray images of individual teeth. We have evaluated the performance of those models on a hold-out test set that features not only a high sample count but also supplementary tooth status annotations. The metrics are analyzed on the entire test set and on subsets per age group, tooth alteration, and tooth type.

Table 2 gives a summary of the overall performance and performance per age group for each task and model for the entire test set, and Table 3 gives the same summary but only for healthy teeth without alterations. A detailed overview of performance metrics per task can be seen in Table 5 and Table 6 for age estimation, in Table 7 for sex assessment, and Table 8 for tooth type determination. An overview of the results for each task is given in the following subsections, and detailed analysis and discussion of the results can be found in Section VI. For classifications tasks, an overview of F1 scores per alterations and overall is given in Table 4. Overall, the F1 scores show that there are no significant differences in performance, and that the slight inbalance of the dataset for sex assessment shown in Section III does not affect the classifier results.

A. AGE ESTIMATION

For age estimation, the best performing model has the following hyperparameter values: the feature extractor is VGG16, a final feature map depth of 662 channels, no attention mechanism, and a fully-connected layer of 1931 units. It achieves an overall absolute mean error of 6.55 years and an overall absolute median error of 5.32 years on the entire test set. When altered teeth are removed from the test set, the overall mean absolute error falls to 6.15 years (6.3% improvement), and the overall absolute median error falls to 4.95 years

TABLE 2. An overview of the performance of all task-specific models on both perfect and imperfect teeth. μ is the mean absolute error, $\wp$ is the median absolute error, and both are measured in years. Sex and tooth type performance is shown as accuracy. The performance for tooth type is shown for each type classification approach.

Age group	Age		Sex	Tooth type			
	μ	$\hat{y}$	Accuracy	Accuracy (4 types)	Accuracy (8 types)	Accuracy (16 types)	Accuracy (32 types)
[20, 25)	8.04	6.95	81.64%	98.95%	96.10%	95.91%	88.77%
[25, 30)	6.11	4.63	72.08%	98.29%	93.90%	93.33%	85.36%
[30, 35)	5.78	4.55	77.95%	98.81%	95.08%	94.66%	85.33%
[35, 40)	5.76	4.97	72.65%	98.21%	92.60%	93.27%	87.44%
[40, 45)	6.17	5.06	74.12%	96.96%	85.71%	85.95%	75.29%
[45, 50)	7.87	6.84	85.71%	96.10%	89.61%	88.31%	76.62%
[50, 55)	8.22	7.74	74.90%	95.29%	84.71%	84.71%	75.69%
[55, 60)	10.40	10.45	64.57%	91.34%	78.74%	78.74%	60.63%
Overall	6.55	5.32	75.44%	97.99%	92.40%	92.23%	83.74%

TABLE 3. An overview of the performance of all task-specific models on healthy, unaltered teeth only. μ is the mean absolute error, $\wp$ is the median absolute error, and both are measured in years. Sex and tooth type performance is shown as accuracy. The performance for tooth type is shown for each type classification approach.

Age group	Age		Sex	Tooth type			
	μ	$\hat{y}$	Accuracy	Accuracy (4 types)	Accuracy (8 types)	Accuracy (16 types)	Accuracy (32 types)
[20, 25)	7.32	6.38	82.09%	99.87%	97.73%	97.23%	90.29%
[25, 30)	5.27	4.03	72.54%	98.36%	94.21%	94.52%	85.76%
[30, 35)	5.31	4.24	77.89%	99.63%	96.89%	96.65%	88.20%
[35, 40)	5.61	5.02	74.49%	100.00%	97.97%	97.97%	91.50%
[40, 45)	6.53	5.57	76.52%	99.35%	91.96%	92.39%	81.30%
[45, 50)	9.29	9.28	87.64%	97.75%	92.13%	91.01%	84.27%
[50, 55)	10.14	9.58	76.77%	97.98%	92.93%	91.92%	86.87%
[55, 60)	12.37	13.04	60.78%	94.12%	90.20%	88.24%	76.47%
Overall	6.15	4.94	76.41%	99.15%	95.53%	95.46%	87.24%

TABLE 4. An overview of F1 scores for each classification task per alteration. This is some more text to make it look more professional with a bolded start. Multiple lines would be nice, and long enough so that it fills the page properly. A bit more text would be perfect, but if there isn't any, there's nothing we can do about.

Forensic task	Filling	Root canal filling	Missing	Tooth decay	Other imperfections	No imperfections	Overall
Sex	75.65%	78.46%	66.89%	79.28%	69.52%	75.77%	74.90%
Type-4	99.54%	98.30%	89.08%	99.59%	87.41%	99.15%	97.99%
Type-8	93.25%	87.40%	73.92%	94.36%	68.27%	95.53%	92.42%
Type-16	93.13%	88.28%	71.05%	96.02%	70.67%	95.46%	92.22%
Type-32	87.29%	83.92%	54.23%	89.46%	62.35%	87.26%	83.75%

(7.41% improvement). A detailed overview of performance per tooth alteration and tooth type can be seen in Table 5 (mean) and Table 6 (median).

B. SEX ASSESSMENT

For the hyperparameters for the sex assessment models, the values are: the feature extractor is VGG16, a final feature map depth of 40 channels, no attention mechanism, and a fully-connected layer of 128 units. The overall accuracy is 75.44% on the entire test set and 76.41% on the subset of healthy unaltered teeth. The model performs best in the age group of 45 to 50 year old samples, with an accuracy of 85.71% and 87.64% for all teeth and healthy unaltered teeth, respectively. A detailed overview of the sex assessment results per tooth alteration and tooth type can be seen in Table 7.

C. TOOTH TYPE DETERMINATION

Tooth type determination has different success rates based on the classification system used. For the 4-type model, an overall accuracy of 97.99% is achieved on the entire test set and an overall accuracy of 99.15% on only healthy unaltered teeth. For the 8-type model, the overall accuracy is 92.40% for the entire dataset and 95.53% for the subset of healthy unaltered teeth. The 16-type model follows the same downwards trend in overall accuracy is noticeable, with the accuracy reaching 92.23% on the entire test set and 95.45% on the subset of healthy unaltered teeth. Finally, the 32-type model achieved an overall accuracy of 83.74% and 87.24% on the entire test set and the subset of healthy unaltered teeth, respectively. Despite the differences in performance, all the hyperparameters for all model variants are the same, with their value being: the feature extractor is VGG16, a final feature map depth of

TABLE 5. Overview of the mean absolute error, measured in years, per tooth type and alteration. Some combinations of tooth type and alteration did not have sufficient samples to produce a valid metric and have therefore been marked as –.

Tooth type	Filling	Root canal filling	Missing	Tooth decay	Other imperfections	No imperfections	Overall
Incisor	6.97	5.70	14.72	3.62	6.56	7.09	7.30
Canine	8.31	9.62	13.33	7.16	10.42	6.70	6.67
Premolar	5.43	6.71	9.51	6.33	8.30	5.78	6.47
Molar	7.52	8.39	8.97	7.83	9.83	4.95	6.05
X1	7.88	9.62	11.58	6.56	10.60	7.17	7.30
X2	8.68	9.62	17.71	7.64	10.28	7.01	7.31
X3	6.97	5.70	14.72	3.62	6.56	6.70	6.67
X4	7.77	8.27	9.03	9.11	9.65	6.21	6.72
X5	7.36	8.44	8.95	7.08	9.97	5.26	6.22
X6	5.35	6.87	10.43	6.10	8.55	4.33	5.98
X7	5.49	6.48	7.62	7.09	7.78	4.55	5.25
X8	5.65	5.67	9.48	5.17	7.95	5.56	6.92
Down-1	11.22	9.92	–	7.04	5.20	7.44	7.46
Down-2	14.25	23.13	29.17	5.60	6.19	6.99	7.07
Down-3	5.12	–	12.50	2.27	1.36	7.13	6.98
Down-4	7.23	10.34	7.35	8.36	7.42	6.35	6.48
Down-5	7.29	9.57	7.59	6.86	8.53	5.15	5.72
Down-6	4.85	5.59	10.31	5.52	7.07	3.96	5.91
Down-7	5.32	6.59	8.19	7.34	7.27	4.50	5.29
Down-8	6.01	6.13	9.82	4.66	5.55	5.15	6.88
Up-1	7.68	9.61	11.58	6.49	11.17	6.81	7.13
Up-2	8.47	9.05	6.25	8.10	10.61	7.03	7.55
Up-3	7.20	5.70	16.94	5.64	7.36	6.23	6.36
Up-4	7.92	8.09	9.49	10.04	10.18	6.00	6.98
Up-5	7.39	8.09	9.65	7.35	10.47	5.41	6.72
Up-6	5.81	7.94	10.77	6.91	9.73	4.60	6.06
Up-7	5.68	6.32	6.83	6.63	8.24	4.59	5.21
Up-8	5.15	3.81	9.10	5.56	11.14	5.93	6.96
11	7.00	9.59	5.36	5.97	10.58	7.42	7.33
12	8.42	7.95	6.25	7.63	9.58	7.03	7.31
13	7.42	4.72	16.94	10.45	11.62	6.63	6.88
14	8.23	6.69	8.06	7.11	9.84	6.15	6.90
15	6.65	6.50	8.64	8.59	9.04	5.69	6.58
16	5.10	7.45	9.83	8.87	10.71	4.87	5.58
17	5.43	4.00	7.41	5.49	4.26	4.52	5.05
18	5.04	3.81	9.54	5.28	7.54	6.23	7.34
21	8.32	9.63	15.72	7.00	11.60	6.18	6.94
22	8.52	10.16	–	8.51	11.42	7.03	7.78
23	7.00	5.98	–	4.68	4.71	5.82	5.83
24	7.69	8.59	10.93	14.92	10.35	5.85	7.05
25	8.11	9.61	11.25	5.69	12.13	5.13	6.87
26	6.54	8.48	11.41	5.86	9.18	4.27	6.54
27	5.92	8.01	6.13	8.16	9.38	4.65	5.37
28	5.24	–	8.60	6.54	18.34	5.65	6.59
31	13.02	–	–	10.01	4.08	7.14	7.18
32	20.59	23.13	–	4.73	0.81	7.06	7.11
33	6.73	–	–	2.67	0.31	6.65	6.54
34	6.59	15.71	13.08	9.43	13.26	6.03	6.30
35	7.64	11.29	9.46	6.86	6.68	5.45	6.02
36	5.08	4.94	10.44	4.51	5.62	3.75	5.84
37	5.53	7.64	8.43	7.79	–	4.29	5.30
38	6.14	4.25	9.53	–	4.08	4.72	6.69
41	10.62	9.92	–	5.05	6.31	7.74	7.74
42	1.57	–	29.17	7.04	11.56	6.92	7.03
43	3.52	–	12.50	1.94	2.40	7.61	7.42
44	7.97	4.97	3.52	7.29	3.53	6.67	6.65
45	6.96	9.26	6.13	6.88	10.01	4.86	5.41
46	4.61	6.29	10.20	6.09	9.25	4.22	5.97
47	5.10	5.75	7.87	6.99	7.27	4.70	5.27
48	5.95	6.76	10.20	4.66	6.04	5.58	7.08
Overall	6.32	7.66	9.51	6.73	9.18	6.15	6.55

TABLE 6. Overview of the median absolute error, measured in years, per tooth type and alteration. Some combinations of tooth type and alteration did not have sufficient samples to produce a valid metric and have therefore been marked as –.

Tooth type	Filling	Root canal filling	Missing	Tooth decay	Other imperfections	No imperfections	Overall
Incisor	5.50	2.54	14.72	2.72	4.35	5.88	6.10
Canine	7.62	7.99	12.23	6.38	9.42	5.70	5.65
Premolar	4.50	5.65	9.10	5.46	7.12	4.68	5.24
Molar	6.18	7.19	7.83	6.16	8.98	4.03	4.80
X1	7.64	9.07	12.23	5.99	8.89	5.88	6.09
X2	7.38	7.17	17.71	6.45	9.96	5.86	6.13
X3	5.50	2.54	14.72	2.72	4.35	5.70	5.65
X4	6.81	5.97	9.13	6.80	9.36	5.06	5.50
X5	5.79	7.28	7.55	5.68	8.72	4.19	4.94
X6	4.26	5.72	9.38	5.38	7.69	3.16	4.61
X7	4.60	5.53	7.18	5.78	7.12	3.78	4.27
X8	5.31	4.25	9.05	4.41	4.08	4.79	5.71
Down-1	10.97	9.92	–	4.87	5.20	5.89	5.95
Down-2	18.04	23.13	29.17	5.57	6.19	5.99	5.90
Down-3	5.35	–	12.50	2.40	1.36	5.92	5.81
Down-4	6.84	10.34	6.62	7.29	7.95	5.34	5.43
Down-5	5.79	7.36	6.37	6.03	6.82	4.10	4.71
Down-6	3.96	4.43	9.45	4.24	5.18	3.13	4.44
Down-7	4.75	4.51	9.86	5.46	7.64	3.89	4.37
Down-8	6.22	5.20	9.59	4.08	3.84	4.37	5.57
Up-1	7.49	8.89	12.23	6.25	11.16	5.85	6.32
Up-2	7.29	6.75	6.25	7.32	9.96	5.83	6.38
Up-3	5.85	2.54	16.94	5.99	5.43	5.16	5.16
Up-4	6.63	5.97	9.69	6.31	9.75	4.64	5.55
Up-5	5.77	7.19	8.38	5.43	10.38	4.64	5.27
Up-6	4.61	6.39	8.98	6.22	8.49	3.31	4.67
Up-7	4.51	6.24	6.01	6.25	7.12	3.68	4.13
Up-8	4.72	3.81	8.00	4.75	14.39	5.07	5.82
11	6.44	9.35	5.36	5.09	11.50	6.65	6.53
12	7.71	5.93	6.25	6.53	5.75	5.85	6.55
13	8.02	4.72	16.94	10.45	8.02	6.03	6.05
14	5.77	5.61	8.15	6.45	7.23	4.58	5.48
15	5.55	5.55	6.12	8.14	9.51	4.73	5.22
16	4.13	5.83	7.52	9.07	8.89	3.51	4.33
17	4.34	3.38	6.94	5.34	4.26	3.29	4.01
18	4.76	3.81	8.02	3.81	7.54	5.39	6.15
21	7.64	8.70	14.86	6.83	8.89	4.97	6.09
22	6.38	7.90	–	7.99	11.20	5.69	6.35
23	3.85	2.54	–	4.75	3.34	5.00	4.78
24	6.87	7.67	10.45	6.16	9.98	4.68	5.80
25	5.89	7.95	13.79	3.67	11.61	4.41	5.35
26	5.30	7.78	10.21	5.79	7.15	3.16	5.30
27	4.52	6.36	5.14	7.25	7.45	3.88	4.27
28	4.29	–	7.84	6.54	18.34	5.02	5.62
31	13.02	–	–	10.01	4.08	5.88	5.88
32	20.59	23.13	–	5.50	0.81	6.05	5.90
33	6.73	–	–	1.86	0.31	5.63	5.58
34	6.84	15.71	13.08	7.43	13.26	5.10	5.44
35	5.87	11.29	8.07	6.18	6.42	4.44	4.95
36	3.96	3.52	9.46	2.81	4.11	2.52	4.17
37	4.64	4.51	9.86	8.14	–	3.04	4.41
38	5.79	4.25	8.78	–	4.08	4.09	5.23
41	10.54	9.92	–	4.34	6.31	6.07	6.20
42	1.57	–	29.17	7.29	11.56	5.86	5.93
43	3.52	–	12.50	2.40	2.40	6.34	6.19
44	6.73	4.97	2.32	4.66	2.54	5.59	5.38
45	5.14	7.36	5.27	4.53	7.49	3.52	4.19
46	3.98	5.65	8.76	4.85	11.02	3.46	4.66
47	4.79	5.12	7.60	5.37	7.64	3.94	4.36
48	6.22	6.14	10.36	4.08	3.59	4.96	6.20
Overall	5.12	6.24	9.01	5.73	7.99	4.94	5.32

TABLE 7. Overview of the accuracy of sex assessment, per tooth type and alteration. Some combinations of tooth type and alteration did not have sufficient samples to produce a valid metric and have therefore been marked as –.

Tooth type	Filling	Root canal filling	Missing	Tooth decay	Other imperfections	No imperfections	Overall
Incisor	75.00%	100.00%	100.00%	86.67%	60.00%	75.25%	75.62%
Canine	80.13%	73.58%	71.43%	86.08%	60.42%	80.83%	80.28%
Premolar	78.04%	79.61%	66.82%	74.76%	72.46%	76.53%	74.37%
Molar	70.46%	76.54%	60.00%	77.55%	68.85%	74.66%	74.42%
X1	80.00%	78.57%	60.00%	88.57%	66.67%	74.77%	75.45%
X2	80.25%	68.00%	100.00%	84.09%	55.56%	75.74%	75.80%
X3	75.00%	100.00%	100.00%	86.67%	60.00%	80.83%	80.28%
X4	69.60%	76.00%	73.91%	94.44%	69.23%	76.88%	75.55%
X5	71.00%	76.79%	53.19%	67.74%	68.57%	76.10%	73.20%
X6	77.59%	77.23%	67.92%	72.92%	68.89%	77.39%	75.86%
X7	79.26%	84.78%	64.58%	79.49%	88.24%	77.45%	77.66%
X8	74.00%	80.00%	66.79%	68.75%	57.14%	71.11%	69.72%
Down-1	25.00%	100.00%	–	80.00%	100.00%	75.65%	75.32%
Down-2	100.00%	100.00%	100.00%	87.50%	100.00%	77.57%	78.12%
Down-3	25.00%	–	100.00%	88.89%	100.00%	79.05%	78.88%
Down-4	69.23%	100.00%	60.00%	100.00%	80.00%	73.49%	73.66%
Down-5	78.95%	76.92%	56.25%	76.47%	55.56%	77.86%	76.98%
Down-6	81.77%	78.26%	69.62%	71.43%	70.00%	84.34%	79.34%
Down-7	81.44%	77.78%	57.14%	76.00%	75.00%	76.00%	77.04%
Down-8	72.41%	100.00%	66.43%	85.71%	75.00%	72.43%	70.59%
Up-1	83.33%	77.78%	60.00%	90.00%	63.16%	73.59%	75.57%
Up-2	79.49%	66.67%	100.00%	83.33%	52.00%	73.11%	73.47%
Up-3	81.25%	100.00%	100.00%	83.33%	53.85%	82.80%	81.68%
Up-4	69.70%	73.91%	77.78%	87.50%	66.67%	81.86%	77.49%
Up-5	66.13%	76.74%	51.61%	57.14%	73.08%	73.60%	69.35%
Up-6	73.76%	76.36%	62.96%	75.00%	68.00%	72.41%	72.41%
Up-7	76.92%	94.74%	75.00%	85.71%	100.00%	78.71%	78.28%
Up-8	76.19%	0.00%	67.19%	55.56%	33.33%	69.92%	68.86%
11	87.50%	72.73%	50.00%	93.33%	50.00%	70.83%	73.74%
12	83.78%	66.67%	100.00%	82.35%	63.64%	72.26%	74.49%
13	73.33%	100.00%	100.00%	100.00%	40.00%	84.09%	82.23%
14	69.05%	66.67%	66.67%	100.00%	57.14%	82.26%	77.66%
15	68.85%	76.19%	52.63%	62.50%	64.29%	72.73%	69.04%
16	75.00%	79.31%	81.82%	71.43%	66.67%	73.02%	73.74%
17	75.32%	87.50%	72.73%	100.00%	100.00%	78.22%	77.78%
18	66.67%	0.00%	70.59%	57.14%	0.00%	69.91%	69.04%
21	79.41%	81.25%	66.67%	86.67%	72.73%	76.43%	77.44%
22	75.61%	66.67%	–	84.21%	42.86%	74.02%	72.45%
23	88.24%	100.00%	–	80.00%	62.50%	81.44%	81.12%
24	70.18%	76.47%	88.89%	66.67%	71.43%	81.42%	77.32%
25	63.49%	77.27%	50.00%	50.00%	83.33%	74.49%	69.68%
26	72.48%	73.08%	50.00%	76.92%	68.75%	71.70%	71.07%
27	78.48%	100.00%	77.78%	66.67%	100.00%	79.21%	78.79%
28	83.33%	–	63.33%	50.00%	100.00%	69.92%	68.69%
31	0.00%	–	–	100.00%	100.00%	76.56%	76.53%
32	100.00%	100.00%	–	80.00%	100.00%	76.72%	77.16%
33	50.00%	–	–	75.00%	100.00%	79.47%	79.19%
34	64.29%	100.00%	100.00%	100.00%	100.00%	72.67%	73.33%
35	86.49%	100.00%	42.86%	69.23%	50.00%	75.36%	75.51%
36	81.19%	83.33%	72.97%	90.00%	75.00%	78.26%	79.70%
37	81.40%	66.67%	56.25%	72.73%	–	75.58%	76.65%
38	80.00%	100.00%	65.38%	–	100.00%	70.37%	69.04%
41	33.33%	100.00%	–	66.67%	100.00%	74.74%	74.11%
42	100.00%	–	100.00%	100.00%	100.00%	78.42%	79.08%
43	0.00%	–	100.00%	100.00%	100.00%	78.61%	78.57%
44	75.00%	100.00%	33.33%	100.00%	66.67%	74.29%	73.98%
45	71.79%	72.73%	66.67%	100.00%	60.00%	80.28%	78.46%
46	82.35%	72.73%	66.67%	61.11%	62.50%	91.89%	78.97%
47	81.48%	86.67%	58.33%	78.57%	75.00%	76.40%	77.44%
48	68.42%	100.00%	67.74%	85.71%	66.67%	74.53%	72.16%
Overall	76.32%	78.31%	66.07%	79.67%	67.36%	76.41%	75.44%

423 channels, no attention mechanism, and a fully-connected layer of 421 units. A detailed overview of the performance of each model can be seen in Table 8.

D. PRELIMINARY EXPERIMENTS FOR ALTERNATIVE APPROACHES

As for preliminary experiments looking for alternative approaches, two were considered. One was a multi-task approach, where a model would estimate two or more of the target variables based on an x-ray image of an individual tooth. We have extensively tested the multi-task models for age estimation and sex assessment over the course of 549 experiments. The evaluation did not yield a model with better performance, with results barely reaching the performance of the solo-task models. Further research is required as to why the experiments show that a multi-task model cannot achieve better nor equal results than single-task models.

The other approach was the inclusion of additional demographic data as the input to a single-task model. Specifically for age estimation, the motivation was based on the differences in child development. While dental developmental markers are strongly genetically defined, to the point where they are regularly used to determine a child's age with an error measured in months, those developmental schedules slightly differ between female and male children. The idea was that the additional data might allow the model to separate ambiguous cases, thereby improving performance even for adults. The inclusion of that information did not improve results, at best achieving the same performance as their image-only counterparts.

VI. ANALYSIS AND DISCUSSION

The achieved results for each task are in-line or better than current state-of-the-art methods while simultaneously being fully automated. All models deliver improved performance when evaluated on a subset of healthy unaltered teeth, despite being trained on teeth of all conditions, as can be seen in Table 2 and Table 3. Some alterations heavily impact the informativeness of a sample, like dental implants that replace a tooth entirely. As seen in Section III, our samples are not uniformly distributed by age. Our test set, consisting of teeth images with annotated interventions, currently has no samples older then 60 years of age.

The distribution of alterations is not uniform across all teeth. Incisors and canines have fewer alterations than premolars and molars. This explains why some specific combinations of tooth type and alterations do not have results in the result tables, as some combinations do not have enough samples to get a valid result.

When compared to our previous research, the results of our experiments show that well-tuned general models at worst perform as well as models specialized for their task per tooth type. Therefore, all results shown are from models that work with any tooth type.

A. MODEL ATTENTION

The newest deep learning research is showing that attention [79] is significantly improving prediction performance and interpretability [87]. We have therefore included attention as an option in this research. Models were constructed with and without attention, with the goal to identify and measure the performance improvements brought by attention. However, across all our experiments and tasks, models with attention underperformed compared to their non-attention counterparts. To control for model capacity, a series of model pairs was trained where the only difference was the presence of attention. This resulted in conclusive results that models with attention underperform for this specific use-case. We assume that this difference in lack of expected performance improvement from attention stems from the lack of variability in this task's domain. While teeth can come in different shapes, sizes and positions, they are fundamentally aligned in one of two ways (as mandibular or maxillary teeth), and they are one of eight possible types. All images are similar in regards to appearance, with a centered tooth, in full view, and scaled as is the clinical standard. Therefore the same indicators are of roughly the same size and at roughly the same position, seemingly nullifying the usual benefit of attention.

B. MULTI-TASK MODEL

Another approach we tested was using one model with multiple outputs - one for each task. The motivation was that by sharing the feature extractor, and by having access to additional information, the model might achieve a better performance in some or all tasks, or at the very least, achieve comparable performance for less learnable parameters. However, the results indicate that such an approach does not yield a model with better performance, and to achieve comparable results the model requires higher complexity. Additionally, the training is less stable, which results in much longer training time until convergence. While some models reached comparable performance to a single task model in one task, no model achieved such a feat in two or more tasks. No detailed analysis per tooth type, sex, age group, or status is reported in this study, as it shares trends with single task models, and a detailed analysis would not yield any useful information. Therefore, all models in this study are single task models.

C. PERFORMANCE AND TRENDS

The performance of all models varies by age group. Each age group is defined as a five-year interval, starting from age 20. As can be seen in Table 2 and Table 3, the performance of each model decreases with age. This is due to multiple factors. With age, the variation in tooth damage and decay makes differentiating between two possible ages much harder. To simplify, while the model can distinguish between 23 and 28-year-old samples with a relatively minor error, it will have difficulty differentiating between 75 and 80-year-old samples. It is also more common in higher age

TABLE 8. Overview of the performance for tooth type determination, per tooth type and alteration. Some combinations of tooth type and alteration did not have sufficient samples to produce a valid metric and have therefore been marked as –. Each classification approach has an additional "Overall" row that shows the model's accuracy for that approach and the precision for every tooth type.

Tooth type	Filling	Root canal filling	Missing	Tooth decay	Other imperfections	No imperfections	Overall
Incisor	94.44%	88.89%	0.00%	100.00%	66.67%	99.39%	99.17%
Canine	100.00%	100.00%	85.71%	100.00%	91.67%	98.33%	97.33%
Premolar	99.87%	99.34%	88.86%	100.00%	94.20%	99.06%	97.74%
Molar	99.08%	96.30%	85.71%	97.96%	81.97%	99.51%	97.59%
Overall	**99.54%**	**98.31%**	**88.02%**	**99.59%**	**87.56%**	**99.15%**	**97.99%**
X1	97.14%	96.43%	80.00%	97.14%	90.48%	97.00%	96.82%
X2	92.59%	80.00%	0.00%	97.73%	66.67%	95.96%	94.65%
X3	100.00%	100.00%	0.00%	93.33%	46.67%	98.75%	97.46%
X4	92.00%	72.00%	52.17%	88.89%	46.15%	96.92%	93.01%
X5	94.00%	82.14%	63.83%	96.77%	74.29%	95.81%	92.40%
X6	93.16%	92.08%	56.60%	91.67%	66.67%	91.96%	86.40%
X7	92.26%	89.13%	54.17%	92.31%	82.35%	86.74%	87.18%
X8	90.00%	80.00%	86.57%	93.75%	85.71%	94.44%	91.35%
Overall	**93.20%**	**87.46%**	**72.65%**	**94.31%**	**68.39%**	**95.53%**	**92.40%**
Down-1	50.00%	100.00%	–	100.00%	0.00%	93.72%	92.88%
Down-2	66.67%	100.00%	0.00%	100.00%	0.00%	95.25%	94.40%
Down-3	100.00%	–	0.00%	100.00%	50.00%	97.88%	97.46%
Down-4	96.15%	50.00%	80.00%	100.00%	80.00%	98.85%	98.21%
Down-5	98.68%	100.00%	56.25%	100.00%	77.78%	97.86%	95.91%
Down-6	95.57%	93.48%	58.23%	89.29%	60.00%	93.98%	85.71%
Down-7	92.81%	88.89%	46.43%	88.00%	62.50%	88.57%	86.73%
Down-8	93.10%	50.00%	93.57%	100.00%	75.00%	98.60%	96.16%
Up-1	98.48%	92.59%	80.00%	96.67%	94.74%	97.18%	97.20%
Up-2	94.87%	87.50%	0.00%	100.00%	76.00%	98.11%	96.43%
Up-3	93.75%	100.00%	0.00%	100.00%	61.54%	99.42%	97.46%
Up-4	91.92%	86.96%	77.78%	100.00%	61.90%	94.94%	91.62%
Up-5	92.74%	79.07%	45.16%	100.00%	57.69%	94.42%	87.27%
Up-6	95.48%	98.18%	33.33%	95.00%	80.00%	89.66%	88.61%
Up-7	85.26%	63.16%	25.00%	85.71%	88.89%	85.64%	82.83%
Up-8	76.19%	100.00%	78.91%	100.00%	100.00%	91.53%	86.84%
Overall	**93.12%**	**88.47%**	**69.86%**	**95.93%**	**70.47%**	**95.46%**	**92.23%**
11	93.75%	90.91%	50.00%	100.00%	87.50%	94.44%	93.94%
12	81.08%	75.00%	0.00%	94.12%	63.64%	94.89%	91.84%
13	86.67%	100.00%	0.00%	100.00%	20.00%	97.16%	93.91%
14	92.86%	83.33%	44.44%	100.00%	57.14%	94.35%	90.43%
15	86.89%	85.71%	21.05%	75.00%	71.43%	92.93%	82.23%
16	90.18%	93.10%	18.18%	100.00%	77.78%	88.89%	84.85%
17	81.82%	62.50%	18.18%	87.50%	50.00%	85.15%	80.30%
18	77.78%	100.00%	61.76%	85.71%	50.00%	86.73%	77.16%
21	88.24%	93.75%	33.33%	93.33%	90.91%	93.57%	91.28%
22	82.93%	75.00%	–	94.74%	50.00%	87.40%	84.18%
23	94.12%	100.00%	–	100.00%	75.00%	94.01%	93.37%
24	89.47%	82.35%	66.67%	100.00%	57.14%	92.04%	87.63%
25	93.65%	95.45%	50.00%	100.00%	75.00%	89.80%	87.23%
26	84.40%	80.77%	18.75%	76.92%	56.25%	79.25%	75.13%
27	75.95%	54.55%	0.00%	66.67%	57.14%	79.21%	73.23%
28	66.67%	–	76.67%	100.00%	100.00%	86.99%	82.83%
31	0.00%	–	–	100.00%	0.00%	74.48%	73.98%
32	100.00%	100.00%	–	60.00%	100.00%	76.19%	76.14%
33	100.00%	–	–	100.00%	0.00%	82.11%	82.23%
34	78.57%	0.00%	100.00%	80.00%	50.00%	90.70%	89.23%
35	89.19%	50.00%	57.14%	92.31%	50.00%	91.30%	89.29%
36	88.12%	91.67%	45.95%	70.00%	50.00%	84.78%	76.14%
37	88.37%	75.00%	18.75%	72.73%	–	80.23%	78.68%
38	90.00%	0.00%	76.92%	–	100.00%	91.67%	85.79%
41	66.67%	100.00%	–	66.67%	0.00%	73.16%	72.59%
42	0.00%	–	0.00%	100.00%	0.00%	80.00%	79.08%
43	50.00%	–	0.00%	80.00%	0.00%	89.30%	87.76%
44	66.67%	0.00%	66.67%	100.00%	33.33%	93.14%	90.82%
45	100.00%	90.91%	44.44%	100.00%	60.00%	95.77%	93.33%
46	92.16%	86.36%	26.19%	83.33%	50.00%	89.19%	75.38%
47	88.89%	86.67%	16.67%	100.00%	62.50%	84.27%	81.54%
48	94.74%	66.67%	67.74%	100.00%	33.33%	83.02%	78.87%
Overall	**87.24%**	**84.07%**	**52.69%**	**89.02%**	**60.62%**	**87.24%**	**83.74%**

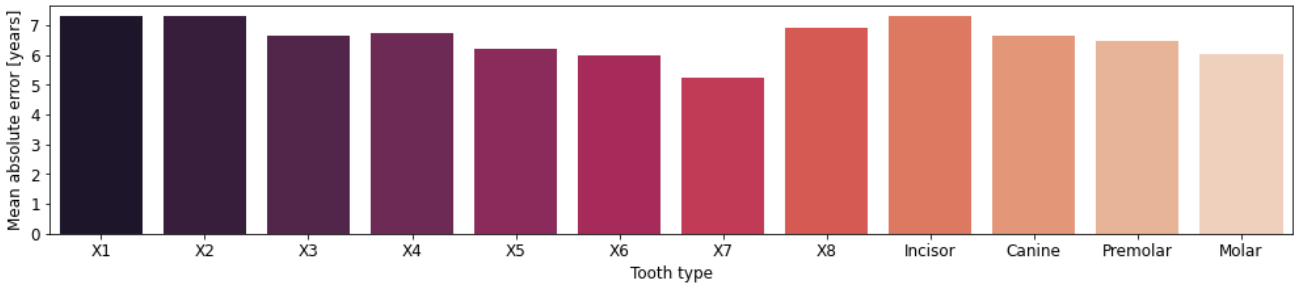


FIGURE 5. Age performance per tooth type. The trend of better performance towards the molars can be observed in the left figure, and a clear improvement in performance for premolars and molars can be observed in the right figure.

groups to have fewer teeth, which impedes the model from seeing and learning to recognize task-related variations in those age groups. Of those teeth present, a significant proportion have alterations, illnesses, or general decay. Those changes to the tooth destroy the naturally-grown tooth and, therefore, possible task-related indicators too.

For age estimation, the best performance can be observed with molars. The overall performance is best on the second molar, which we denoted in Table 5 and 6 as "X7". The upper left second molar has the best performance, with an average of 0.3 years advantage over their counterparts. When analyzing only healthy, unaltered teeth, molars still perform best, but the best performance is achieved by the first molar, closely followed by the performance on the second molar.

On the other hand, sex assessment performance is best on canines, with a significant 4.4% lead. Premolars and molars perform about the same, with a difference of 0.05% in performance. Incisors perform slightly better than premolars and molars, achieving a modest 1% increase in performance. On healthy, unaltered teeth, canines still lead the pack. Interestingly, the difference in performance on healthy, unaltered and unhealthy, altered teeth for canines is small, amounting to a decrease of just 0.55%. Overall, maxillary teeth performance for sex estimation is better than mandibular teeth.

Tooth type estimation performs best when only classifying into 4 classes and, as expected, worst when classifying into 32 classes. The highest precision is achieved for incisors, as they have the most distinguished shape in relation to all teeth. Most misclassifications happen within the same morphological type (the 4-class system). In other words, if a misclassification occurs, it is very likely to be between neighboring teeth within the same group. For example, the model might mistake the first molar for the second molar, but it is very unlikely to confuse the first molar with an incisor or canine. These phenomena can best be observed in the confusion matrix, which is shown in Figure 9.

Another factor that significantly impacts performance is tooth type. The highest difference can be observed when taking into consideration the 4-type classification type, as can be seen in Tables 5, 6, 7, and 8. For age estimation, the best performance can be observed on premolars, as can be seen in Figure 5. For sex assessment, the best performance is seen on

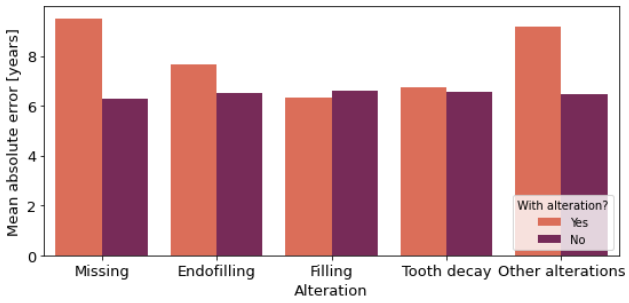


FIGURE 6. Comparison of age performance per alteration. A clear trend of higher absolute errors can be observed when alterations are present. Each bar shows the model performance with and without the alteration present.

incisors, shown on Figure 10. For type classification systems of a higher class count, while there are outliers, the trend from the 4-class system is observable.

Alterations can heavily impact model performance for every task, but it depends on the alteration and how much of the natural tooth is modified or removed. For age estimation, fillings, root canal fillings, and tooth decay reduce the model's accuracy to a similar degree. Fillings and tooth decay can vary in size, with larger sizes reducing model performance. Root canal fillings on their own do not have such a variance in size (and therefore model performance), as they are limited to root canals only. Other imperfections contain extremely destructive alterations, like replacing a tooth with an dental implant, or tooth germs, undeveloped teeth that lack the characteristics and indicators needed to perform the required tasks. These trends can be seen on Figure 6. Interestingly, a negative correlation between alterations and age estimation error can be observed for samples over 50 years. This implies that the model uses the damage and decay itself as an indicator. Such behavior is acceptable, as the frequency and intensity of those indicators truly correlate with age - older people have more dental work done due to the natural accumulation of damage and decay. A detailed overview of the estimation error bias can be seen in Figure 7.

A highly interesting category in our analysis is the missing tooth. As the tooth is missing, the models should not be able

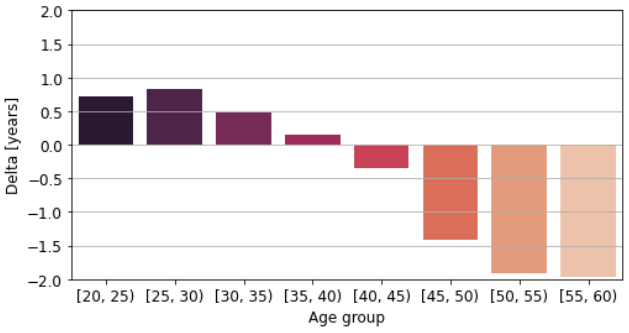


FIGURE 7. Difference between the mean absolute error of teeth with no alterations and teeth with alterations. For younger samples, in the age groups between 20 and 40 years, it can be seen that unaltered teeth perform better. For samples older than 40 years, alterations contribute positively to age estimation.

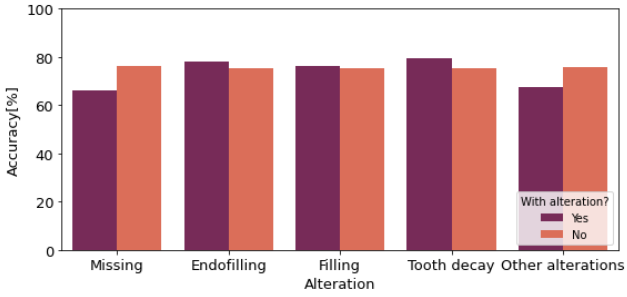


FIGURE 8. Comparison of sex assessment performance per alteration. More common alterations do not seem to impact overall accuracy significantly. However, a high impact can be observed on extreme or less common alterations (e.g., missing teeth, dental implants, crowns, bridges). Each bar shows the model performance with and without the alteration present.

to do any correct estimations, as the information necessary is simply not present. For age estimation, in Tables 5 and 6 we can see very high errors and variations in those errors, which supports our hypothesis about the presence and importance of the tooth. However, for sex assessment Table 7 shows better-than-random performance for missing teeth. As teeth are annotated with bounding boxes, neighboring teeth are commonly visible in parts of a sample image. For ''missing teeth'' samples, those bounding boxes encompass the gap in dentition, which can have neighboring teeth visible. Teeth can move, and if a gap is present, they will move towards it, leading to smaller gaps and effectively a ''better view of our neighborhood'' for our ''missing teeth'' samples. In essence, the sex assessment model does not rely solely on the tooth but also on the structures surrounding it. Future research needs to be done to determine the difference in performance when the surrounding structures are not present, for which either automatic or manual segmentation is required. Tooth type determination models do not suffer such a high decrease in performance when evaluated on ''missing tooth'' samples, indicating that they rely much more on the surrounding structure to determine the tooth type. With the increase of classes in different classification systems, the performance decreases, as is expected. However, we can observe that a tooth is easily defined by its neighborhood.

As for tooth type determination, we can observe in Table 8 that performance mostly decreases with the number of classes required. This trend does not hold true between the 8 and 16 class models. The difference between those two models is that a differentiation is made between maxillary and mandibular teeth. As tooth samples are not rotated during training, their orientation is an easily recognizable indicator of mandibular belonging. Thus, the model can look for the same indicators in the 8 and 16 class problems, with the addition of an orientation indicator for the 16 class model. The 32 class model does perform noticeably worse in differentiating between individual teeth positions, but it still differentiates well between the basic four classes. This can be seen in the confusion matrix in Figure 9, where we can observe scattering errors between neighboring classifications but very few long-distance errors.

D. COMPARISON WITH CURRENT METHODS

To properly understand the performance of this study, we need to compare it to the current state-of-the-art methods in forensic odontology literature. Age estimation methods have not changed much since the discovery of the correlation between age and secondary dentine deposits and the reduction of the pulp cavity. Newer studies either focus on a different part of the dental system (for example, age estimation from panoramic dental x-ray images [2], [60] or cone beam CT scans [88]), or they determine specific parameters for the tried-and-proven methods for different populations [25]–[31]. Sex assessment studies use a wide set of dental parameters, with newer studies achieving higher and higher accuracy. Moreover, while there are studies reporting a high variability in performance, studies with strict methodology and high sample sizes usually do not exceed 80%, as has been described as a ceiling for this sort of sex assessment from individual teeth [38]. Tooth type determination is taught early in dentistry education through dental morphology understanding. Morphology is taught from established textbooks [89], [90], and more focus in research is put towards effective teaching [91]. To the best of our knowledge, there are no studies researching automated age estimation and sex assessment from individual x-ray tooth images, and there are no studies directly researching the accuracy of dentistry experts for tooth type determination from x-ray images.

1) AGE

Single tooth age estimation in modern forensic odontology is based on three fundamental studies [21], [23], [24]. A common factor among all age estimation methods is the need for unaltered, healthy teeth. In other words, no decay, no fillings, no root canal fillings, or any other type of addition or subtraction of the dental tissue is permissible for those methods to work. As they require manual measurements and, therefore, a very high time investment, those methods are determined on

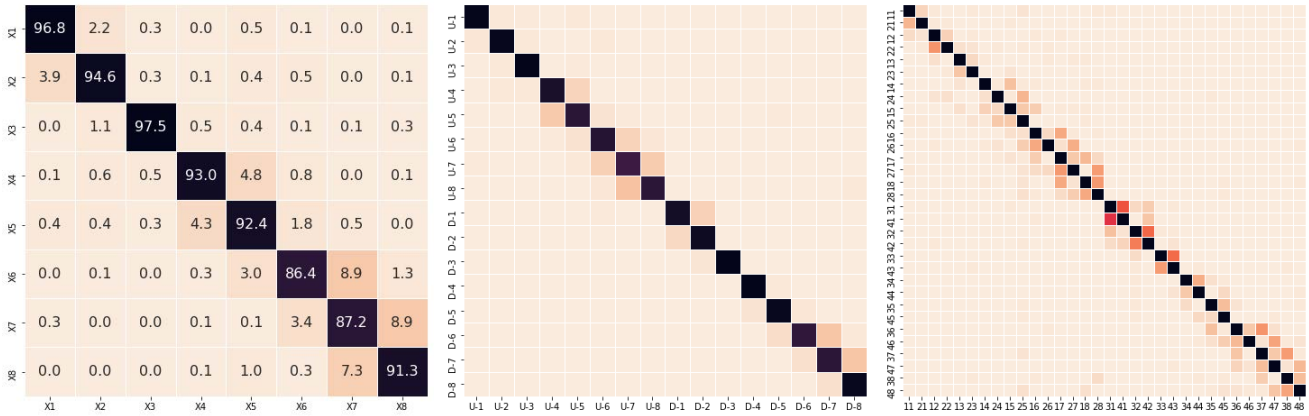


FIGURE 9. Confusion matrices for tooth type determination for the 8, 16, and 32 class approach. The x-axis represents the predicted labels, and the y-axis represents the true labels. For the 16-class case (middle figure), the prefixes "U-" and "D-" represent maxillary and mandibular teeth, respectively. The values in the left figure are the number of samples normalized by the number of true samples of its class (i.e., precision). As can be seen in the left figure, misclassifications mostly happen between teeth in the same morphological group (4-class system), with very rare instances of errors outside those groups. The middle figure shows the confusion matrix for the 16-class case. Again, misclassifications happen in the same morphological group, but with no errors mistaking mandibular and maxillary teeth. The right figure showing the 32-class case further reinforces that trend, with errors occurring cross-quadrant but within the same morphological group and the same jaw-side. For this qualitative overview, values are omitted in the middle and right figure for visual clarity.

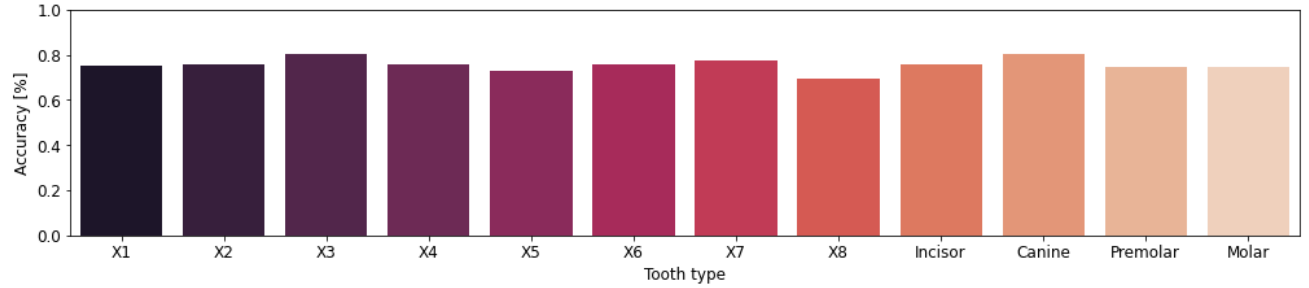


FIGURE 10. Sex performance per tooth type. The accuracy of sex assessment per tooth type in the 4 and 8 class classification system. Our model performs best on canines, which can be observed from both perspectives but is more clearly noticeable in the 4-class system.

TABLE 9. Performance comparison of the most used methods for age estimation from a single tooth in forensic odontology. The performance for our methods is given for intact, unaltered teeth (upper row) and performance on the entire dataset (lower row). Current methods in forensic odontology work only on intact, unaltered teeth. S.E. is the standard error, μ is the mean absolute error, and $\tilde{\varphi}$ is the median absolute error. All values are taken in their original form as given in the cited studies.

Study	Sample size	Error
Kvaal et al. [21]	100 individuals	S.E. = 8.6 to 11.5 years
Drusini et al. [23]	846 intact teeth	S.E. = 5.88 to 6.66 years
Cameriere et al. [24]	100 individuals	$\tilde{y}$ = 3.7 years
Ours	**86495 teeth**	**μ= 6.15, $\tilde{y}$ = 4.94, S.E. = 7.95**
		μ= 6.55, $\tilde{y}$ = 5.32, S.E. = 8.51

a relatively small sample size (less than a thousand samples) compared to our study (over 80000 samples). The specific performance values can be seen in Table 9. As this study explores the impact of tooth alterations, both results with and without alterations are given in the overview table.

Kvaal *et al.* [21] calculate linear regression models by tooth dimension ratios, and those models can be based on multiple or single tooth measurements. The models that take into account more teeth perform better, ranging in standard error between 8.6 and 11.5 years, while our study achieves a standard error of 7.95 years on healthy, unaltered single-tooth images and 8.51 years on any single tooth images. Drusini *et al.* [23] uses the coronal pulp cavity index [22] for molars and premolars, achieving a standard error between 5.88 and 6.66 years. While our overall standard error of 7.95 years on healthy, unaltered teeth is higher, our standard error for healthy molars is 6.37 years and 7.53 years for premolars, which is comparable in performance. The overall higher error can be explained by the inclusion of unhealthy, altered teeth, and higher variability of other teeth, which ultimately led Drusini *et al.* [23] to consider only premolars and molars in their research. Cameriere *et al.* [24] uses tooth dimension ratios of a single-rooted tooth to derive their model, specifically the right maxillary canine (tooth 23 according to the FDI numbering system). They achieve a median error of 3.7 years, compared to our overall median error of 4.94 years on healthy, unaltered teeth and 5.32 years on all teeth. While the difference is relatively big, our

TABLE 10. Performance comparison of the most used methods for sex assessment from a single tooth in forensic odontology. The performance for our methods is given for intact, unaltered teeth (upper row) and performance on the entire dataset (lower row). Current methods in forensic odontology work only on intact, unaltered teeth.

Study	Sample size	Accuracy
Karaman [41]	60 dental casts	83.3%
Capitaneanu et al. [39]	200 OPGs	69.0% to 72.5%
Neves et al. [42]	168 dental casts	75%
Ours	**86495 teeth**	**76.41%** **75.44%**

results are based on a much higher number of samples (over 80000 vs. 100 in [24]), and our model is made for any tooth type, therefore reducing the chance of any statistical overfitting. Our previous study [2] achieves a mean absolute error of 3.96 years for panoramic dental x-ray images. This is better than the results in this study; however, this is due to the difference in the analyzed image. Panoramic dental x-ray images contain more tissue and thus more age-related information; therefore, a model can achieve a lower overall error. The preliminary results for individual tooth x-ray images in our previous study achieve an overall mean absolute error of 7.49 years, while this study achieves an overall mean absolute error of 6.55 years and a mean absolute error of 6.15 years for teeth without alterations, which is a significant improvement in performance for individual tooth x-ray images. Considering these studies, we can conclude that our method performs in-line or better than other methods while simultaneously being fully automated and, up to a point, resistant to natural and artificial alterations.

2) SEX

Sex assessment of individual teeth, be it from x-ray images, casts, or physical teeth, is based on the correlation between different dimensions of the tooth. Like with age, these methods cannot work on unhealthy or altered teeth, as those modify the tooth shape and can lead to improper assessments. As already mentioned, current research with strict methodology and a high sample count has not been able to show assessment performance higher than 80% [38]. Additionally, as those methods require manual measurements, the analyzed sample size is significantly smaller than what could be achieved in our study with automation. There are methods based on measurements and analysis of the entire panoramic dental x-ray image which yield significantly better results but require an intact dental system [1], [65], [66]. Specific values for three representative methods can be seen in Table 10.

Mesiodistal and buccolingual measurements are commonly used for sex assessment, but Karaman [41] uses diagonal measurements and achieves an accuracy of 83.3% with a sample size of 60 dental casts. This is higher than our 75.44% and 76.41% accuracy for altered and unaltered teeth, and this is higher than the proposed 80% limit. While the method proves successful, we assume that the higher-than-usual reported accuracy is due to the very small sample size. Capitaneanu *et al.* [39] did not focus on a specific set of teeth or measurements and instead did a multivariate analysis of all tooth length and width related variables, which totals 212 variables in their study. On a sample of 200 panoramic dental x-ray images (100 female, 100 male), using principal component analysis (PCA), they have determined that measurements of a single tooth, sex can be assessed with an accuracy between 69.0% to 72.5%. This is slightly lower than our results but ultimately can be considered in-line when the sample size is considered. Neves *et al.* [42] is the newest study, and they developed a predictive model using teeth mesiodistal widths. They use 168 samples in their study (109 female, 59 male) and achieve an accuracy of 75%. This is in-line with our research, with Capitaneanu *et al.* [39], the proposed limit [38], and our research. We can therefore conclude that, while our method does not significantly outperform other single-tooth methods, it performs as well as others, it is verified on the largest dataset of its kind in literature, it is fully automated, it is infinitely reproducible, it can process inhumanly large workloads in seconds, and it can be reliably used with unhealthy and slightly altered teeth.

3) TYPE

Tooth type determination is often the prerequisite to applying other methods. As seen for sex assessment and age estimation, methods often work for measurements of specific teeth, so the final confidence of our predictions is not only predicated on the task-specific method but also the method of tooth type determination. Tooth type determination, either as part of another task or as a standalone task, has been tackled by other research into automated methods.

Oktay [52] determines the tooth type in the 4-class variant as part of tooth detection, and they achieve an accuracy of 92.84%. This is lower than our approach, which achieves an accuracy of 99.15% for healthy, unaltered teeth and an overall accuracy of 97.99% for the 4-class approach. This discrepancy can be explained by their focus on detection instead of tooth type determination, and their comparatively small dataset. Keerthana *et al.* [70] is primarily focused on the determination of tooth type. They use projection profile analysis for their model and achieve an accuracy of 92.54% with a 4-class approach on a dataset of 200 individual tooth x-ray images (50 images per tooth type). The difference in performance between our study and this one could come from the difference in sample size and the difference in model complexity. Neural networks have a much higher capacity than the proposed model in [70], which, combined with much denser (images vs. specific measurements) and more numerous data points (raw sample count), produces a more robust and accurate model. Chen *et al.* [71] is equally focused on detection and type determination. They use the 32-class approach, and their dataset consists of 1250 dental periapical films. The accuracy of their model varies between 71.5% to 91.7%. Their upper accuracy is very high but is a consequence of multiple post-processing steps. Information about

TABLE 11. Performance comparison of the most used methods for tooth type determination from a single tooth in forensic odontology. The performance for our methods is given for intact, unaltered teeth (upper row) and performance on the entire dataset (lower row). Current methods in forensic odontology work only on intact, unaltered teeth.

Study	Sample size	Classes	Accuracy
Oktay et al. [52]	100 OPGs	4 types	92.84%
Keerthana et al. [70]	200 tooth x-rays	4 types	92.54%
Chen et al. [71]	1250 dental periapical films	32 types	71.5% to 91.7%
Prados-Privado et al. [92]	8000 OPGs	*32/8 types	93.83%
Ours	**86495 teeth**	**4, 8, 16 and 32 types**	87.24% to 99.15%
		4, 8, 16 and 32 types	83.74% to 97.99%

neighboring teeth is incorporated in multiple stages, which ultimately corrects miss-classifications. While this approach is valid in some cases, a fair comparison is only possible when no post-processing is applied. Thus, our model outperforms this study with an accuracy of 87.24% for healthy, unaltered teeth and an overall accuracy of 83.74%. While relatively close in performance, we believe that our model performs better due to sample size, model capacity, and the solitary focus on tooth type determination. Prados-Privado *et al.* [92] determine the tooth type on panoramic dental x-ray images. They do not use any neighborhood-based post-processing of results, but they use symmetries of teeth to simplify their problem from the 32-class approach to the 8-class approach. Additionally, while not going into an exhaustive analysis, they have some insight into the impact of tooth alterations, claiming that ''the network is capable of correctly numbering teeth that contain metal parts, or any other treatment performed on it such as filled teeth, but in the case of the prosthetic crown, it detects a single tooth.'' [92] Our model achieves only an accuracy of 87.24% on healthy, unaltered teeth for the 32-class case and an overall accuracy of 83.74%. However, as [92] uses the 8-class approach, which is then post-processed into the 32-class case, we believe that comparison to our 8-class results is more appropriate. In that case, we achieve an accuracy of 95.53% on healthy, unaltered teeth and an overall accuracy of 92.40%. While we cannot come to a definitive conclusion given the differences in approaches and missing information about the properties of the dataset, and given the difference in sample size, we believe that the performance of both models is in-line with our method performing on par while having less information available to analyze.

VII. CONCLUSION

Age estimation and sex assessment are two cornerstone tasks of forensic odontology. Current state-of-the-art methods for those tasks rely on manual measurements, which are time-consuming, repetitive, and can introduce human error. It is also often necessary to know which tooth type is being analyzed for individual teeth. This tooth type determination is too done with manual measurements and estimations.

In this study, we have exhaustively analyzed deep learning as the approach to automate those tasks, speed them up, remove human error, and match or improve current performance. We propose three models based on deep convolutional neural networks that automatically analyze an x-ray image of a tooth, estimate the age, assess the sex, or determine the tooth type. Additionally, we have evaluated the possibility of designing a multi-task model, which would reduce the amount of computing resources necessary and potentially increase the overall performance across all tasks.

The models are designed and evaluated on one of the largest and most extensive datasets of individual tooth x-ray images in literature, containing 86495 images. The dataset also has an advantage over other datasets of this kind, as it offers a subset of not only imperfect dentition, which is notably mostly absent in forensic odontology research but also annotations of tooth status. These tooth status annotations have allowed us to analyze the models from a novel point of view, highlighting the performance of the models on perfect and imperfect dentition, as well as the impact specific alterations can have on the tasks.

The performance of the constructed models is equal to or better than current state-of-the-art methods in forensic odontology. In addition, this approach solves reproducibility and human error problems while simultaneously reducing the forensic estimation time to a minimum. This study explores a vast space of models, not only producing usable models but also showing which state-of-the-art feature extraction architectures perform best and in which conditions. These results are achieved and validated on one of the most extensive datasets in literature. This approach also expands the variety of samples that can be used in forensic analysis, as they work well with common tooth alterations, albeit with slightly decreased average performance.

In conclusion, the proposed approach and the designed and created models achieve an accuracy of 76.41% for sex assessment, a median absolute error of 4.94 years for age estimation, and an accuracy of 87.24% to 99.15% for tooth type determination, and this study shows that alterations on average hinder the correct classification and estimation with different magnitudes of impact for almost all cases, with an exception for age estimation of older samples, where the damage and decay are successfully used by the model as age indicators.

ACKNOWLEDGMENT

The authors would like to thank the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] D. Milošević, M. Vodanović, I. Galić, and M. Subašić, "Estimating biological gender from panoramic dental X-ray images," in *Proc. 11th Int. Symp. Image Signal Process. Anal. (ISPA)*, Sep. 2019, pp. 105–110.

[2] D. Milošević, M. Vodanović, I. Galić, and M. Subašić, "Automated estimation of chronological age from panoramic dental X-ray images using deep learning," *Expert Syst. Appl.*, vol. 189, Mar. 2022, Art. no. 116038.

[3] L. Banjsak, D. Milosevic, M. Subasic, H. Brkic, and M. Vodanovic, "Artificial intelligence implementation in tooth identification from X-ray images," *Int. Dental J.*, vol. 71, p. S35, Sep. 2021.

[4] E. Saunders, "The teeth a test of age, considered with reference to the factory children," *Amer. J. Dental Sci.*, vol. 7, no. 4, p. 330, 1837.

[5] W. M. Krogman, "The human skeleton in forensic medicine. I," *Postgraduate Med.*, vol. 17, no. 2, p. 48, 1955.

[6] J. C. Dudar, S. Pfeiffer, and S. Saunders, "Evaluation of morphological and histological adult skeletal age-at-death estimation techniques using ribs," *J. Forensic Sci.*, vol. 38, no. 3, pp. 677–685, 1993.

[7] S. Berg, M. Casey, F. Raasch, and T. Solheim, "Dental age estimation: An alternative technique for tooth sectioning," *Amer. J. Forensic Med. Pathol.*, vol. 5, no. 2, pp. 181–184, Jun. 1984.

[8] S. Singaraju and P. Sharada, "Age estimation using pulp/tooth area ratio: A digital image analysis," *J. Forensic Dental Sci.*, vol. 1, no. 1, p. 37, 2009.

[9] S. P. M. Carvalho, R. H. A. da Silva, C. Lopes-Júnior, and A. S. Peres, "Use of images for human identification in forensic dentistry," *Radiologia Brasileira*, vol. 42, no. 2, pp. 125–130, Apr. 2009.

[10] A. Panchbhai, "Dental radiographic indicators, a key to age estimation," *Dentomaxillofacial Radiol.*, vol. 40, no. 4, pp. 199–212, May 2011.

[11] P. G. Limdiwala and J. S. Shah, "Age estimation by using dental radiographs," *J. Forensic Dental Sci.*, vol. 5, no. 2, pp. 118–122, 2013.

[12] T. Y. Marroquin, S. Karkhanis, S. I. Kvaal, S. Vasudavan, E. Kruger, and M. Tennant, "Age estimation in adults by dental imaging assessment systematic review," *Forensic Sci. Int.*, vol. 275, pp. 203–211, Jun. 2017.

[13] 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," *Int. J. Morphol.*, vol. 33, no. 4, pp. 1248–1254, Dec. 2015.

[14] C. M. Nolla, "The development of permanent teeth," Ph.D. dissertation, School Dentistry, Univ. Michigan, Ann Arbor, MI, USA, 1952.

[15] A. Demirjian, H. Goldstein, and J. Tanner, "A new system of dental age assessment," *Hum. Biol.*, vol. 45, no. 2, pp. 211–227, May 1973.

[16] K. Haavikko, "The formation and the alveolar and clinical eruption of the permanent teeth. An orthopantomographic study," *Suomen Hammaslaakariseuran Toimituksia=Finska Tandlakarsallskapets Forhandlingar*, vol. 66, no. 3, pp. 103–170, 1970.

[17] I. Gleiser and E. E. Hunt, Jr., "The permanent mandibular first molar: Its calcification, eruption and decay," *Amer. J. Phys. Anthropol.*, vol. 13, no. 2, pp. 253–283, 1955.

[18] C. F. A. Moorrees, E. A. Fanning, and E. E. Hunt, "Age variation of formation stages for ten permanent teeth," *J. Dental Res.*, vol. 42, no. 6, pp. 1490–1502, Nov. 1963.

[19] R. Cameriere, L. Ferrante, and M. Cingolani, "Age estimation in children by measurement of open apices in teeth," *Int. J. Legal Med.*, vol. 120, no. 1, pp. 49–52, Jan. 2006.

[20] R. Cameriere, L. Ferrante, D. De Angelis, F. Scarpino, and F. Galli, "The comparison between measurement of open apices of third molars and Demirjian stages to test chronological age of over 18 year olds in living subjects," *Int. J. Legal Med.*, vol. 122, no. 6, pp. 493–497, Nov. 2008.

[21] S. I. Kvaal, K. M. Kolltveit, I. O. Thomsen, and T. Solheim, "Age estimation of adults from dental radiographs," *Forensic Sci. Int.*, vol. 74, no. 3, pp. 175–185, Jul. 1995.

[22] N. Ikeda, K. Umetsu, S. Kashimura, T. Suzuki, and M. Oumi, "Estimation of age from teeth with their soft X-ray findings," *Jpn. J. Legal Med.*, vol. 39, no. 3, pp. 244–250, 1985.

[23] A. G. Drusini, O. Toso, and C. Ranzato, "The coronal pulp cavity index: A biomarker for age determination in human adults," *Amer. J. Phys. Anthropol.*, vol. 103, no. 3, pp. 353–363, Jul. 1997.

[24] R. Cameriere, L. Ferrante, and M. Cingolani, "Variations in pulp/tooth area ratio as an indicator of age: A preliminary study," *J. Forensic Sci.*, vol. 49, no. 2, pp. 1–3, 2004.

[25] C. S. Farah, D. R. Booth, and S. C. Knott, "Dental maturity of children in Perth, Western Australia, and its application in forensic age estimation," *J. Clin. Forensic Med.*, vol. 6, no. 1, pp. 14–18, Mar. 1999.

[26] F. Ardakani, N. Bashardoust, and M. Sheikhha, "The accuracy of dental panoramic radiography as an indicator of chronological age in Iranian individuals," *J Forensic Odontostomatol*, vol. 25, no. 2, p. 25, 2007.

[27] A. Gulsahi, B. Yüzügüllü, P. Imirzalıoğlu, and Y. Genç, "Assessment of panoramic radiomorphometric indices in Turkish patients of different age groups, gender and dental status," *Dentomaxillofacial Radiol.*, vol. 37, no. 5, pp. 288–292, Jul. 2008.

[28] M. Babar, S. Iqbal, and A. Jan, "Essential guidelines for forensic dentistry," *Pakistan Oral Dental J.*, vol. 27, pp. 79–84, Jan. 2008.

[29] I. Galić, E. Nakaš, S. Prohić, E. Selimović, B. Obradović, and M. Petrovecki, "Dental age estimation among children aged 5–14 years using the Demirjian method in Bosnia–Herzegovina," *Acta Stomatologica Croatica*, vol. 44, no. 1, pp. 17–25, 2010.

[30] I. Galić, M. Vodanović, R. Cameriere, E. Nakaš, E. Galic, E. Selimovic, and H. Brkic, "Accuracy of Cameriere, Haavikko, and Willems radiographic methods on age estimation on Bosnian–Herzegovian children age groups 6–13," *Int. J. Legal Med.*, vol. 125, no. 2, pp. 315–321, 2011.

[31] A. Selmanagić, M. Ajanović, A. Kamber-Ćesir, L. Redžepagić-Vražalica, A. Jelešković, and E. Nakaš, "Radiological evaluation of dental age assessment based on the development of third molars in population of Bosnia and Herzegovina," *Acta Stomatologica Croatica*, vol. 54, no. 2, pp. 161–167, Jun. 2020.

[32] H.-M. Jeon, S.-M. Jang, K.-H. Kim, J.-Y. Heo, S.-M. Ok, S.-H. Jeong, and Y.-W. Ahn, "Dental age estimation in adults: A review of the commonly used radiological methods," *J. Oral Med. Pain*, vol. 39, no. 4, pp. 119–126, Dec. 2014.

[33] E. S. Martin, "A study of an Egyptian series of mandibles, with special reference to mathematical methods of sexing," *Biometrika*, vol. 28, nos. 1–2, pp. 149–178, Jun. 1936.

[34] G. M. Morant and N. K. Adyanthaya, "A biometric study of the human mandible," *Biometrika*, vol. 28, nos. 1–2, pp. 84–122, Jun. 1936.

[35] A. Hrdlička, "Mandibular and maxillary hyperostoses," *Amer. J. Phys. Anthropol.*, vol. 27, no. 1, pp. 1–67, Jun. 1940.

[36] H. De Villiers, "Sexual dimorphism of the skull of the South African Banu-speaking Negro," *South Afr. J. Sci.*, vol. 64, no. 2, p. 118, 1968.

[37] L. T. Humphrey, M. C. Dean, and C. B. Stringer, "Morphological variation in great ape and modern human mandibles," *J. Anatomy*, vol. 195, no. 4, pp. 491–513, Nov. 1999.

[38] A. P. Joseph, R. K. Harish, P. K. R. Mohammed, and R. B. V. Kumar, "How reliable is sex differentiation from teeth measurements," *Oral Maxillofacial Pathol. J.*, vol. 4, no. 1, pp. 289–292, 2013.

[39] C. Capitaneanu, G. Willems, R. Jacobs, S. Fieuws, and P. Thevissen, "Sex estimation based on tooth measurements using panoramic radiographs," *Int. J. Legal Med.*, vol. 131, no. 3, pp. 813–821, May 2017.

[40] C. Capitaneanu, G. Willems, and P. Thevissen, "A systematic review of odontological sex estimation methods," *J. Forensic Odonto-Stomatol.*, vol. 35, no. 2, pp. 1–19, Dec. 2017.

[41] F. Karaman, "Use of diagonal teeth measurements in predicting gender in a Turkish population," *J. Forensic Sci.*, vol. 51, no. 3, pp. 630–635, May 2006.

[42] J. A. Neves, N. Antunes-Ferreira, V. Machado, J. Botelho, L. Proença, A. Quintas, J. J. Mendes, and A. S. Delgado, "Sex prediction based on mesiodistal width data in the Portuguese population," *Appl. Sci.*, vol. 10, no. 12, p. 4156, Jun. 2020.

[43] M. R. Dayal, M. A. Spocter, and M. A. Bidmos, "An assessment of sex using the skull of black south Africans by discriminant function analysis," *HOMO*, vol. 59, no. 3, pp. 209–221, Jul. 2008.

[44] 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," *J. Forensic Sci.*, vol. 56, no. s1, pp. S13–S16, 2011.

[45] A. P. Indira, A. Markande, and M. P. David, "Mandibular ramus: An indicator for sex determination—A digital radiographic study," *J. Forensic Dental Sci.*, vol. 4, no. 2, pp. 58–62, 2012.

[46] M. Marinescu, V. Panaitescu, and M. Rosu, "Sex determination in Romanian mandible using discriminant function analysis: Comparative results of a time-efficient method," *Romanian J. Legal Med.*, vol. 21, no. 4, pp. 305–308, Dec. 2013.

[47] T. Bhagwatkar, M. Thakur, D. Palve, A. Bhondey, and Y. Dhengar, "Sex determination by using mandibular ramus—A forensic study," *J. Adv. Med. Dental Sci. Res.*, vol. 4, no. 2, p. 6, 2016.

[48] 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," *J. Indian Acad. Oral Med. Radiol.*, vol. 29, no. 3, p. 242, 2017.

[49] T. Nagaraj, L. James, S. Gogula, N. Ghouse, H. Nigam, and C. K. Sumana, "Sex determination by using mandibular ramus: A digital radiographic study," *J. Med., Radiol., Pathol. Surg.*, vol. 4, no. 4, pp. 5–8, 2017.

[50] 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, no. 5, pp. e217–e223, Oct. 2018.

[51] F. Piccialli, V. D. Somma, F. Giampaolo, S. Cuomo, and G. Fortino, "A survey on deep learning in medicine: Why, how and when?" *Inf. Fusion*, vol. 66, pp. 111–137, Feb. 2021.

[52] A. B. Oktay, "Tooth detection with convolutional neural networks," in *Proc. Med. Technol. Nat. Congr. (TIPTEKNO)*, Oct. 2017, pp. 1–4.

[53] G. Jader, J. Fontineli, M. Ruiz, K. Abdalla, M. Pithon, and L. Oliveira, "Deep instance segmentation of teeth in panoramic X-ray images," in *Proc. 31st Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 400–407.

[54] G. Silva, L. Oliveira, and M. Pithon, "Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives," *Expert Syst. Appl.*, vol. 107, pp. 15–31, Oct. 2018.

[55] D.-Y. Kang, H. P. Duong, and J.-C. Park, "Application of deep learning in dentistry and implantology," *Korean Acad. Oral Maxillofacial Implantol.*, vol. 24, no. 3, pp. 148–181, Sep. 2020.

[56] M. Machoy, J. Seeliger, L. Szyszka-Sommerfeld, R. Koprowski, T. Gedrange, and K. Woźniak, "The use of optical coherence tomography in dental diagnostics: A state-of-the-art review," *J. Healthcare Eng.*, vol. 2017, Jul. 2017, Art. no. 7560645.

[57] C.-H. Wu, W.-H. Tsai, Y.-H. Chen, J.-K. Liu, and Y.-N. Sun, "Model-based orthodontic assessments for dental panoramic radiographs," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 2, pp. 545–551, Mar. 2018.

[58] C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi, "Deep learning for automated skeletal bone age assessment in X-ray images," *Med. Image Anal.*, vol. 36, pp. 41–51, Feb. 2017.

[59] Y. Lai, F. Fan, Q. Wu, W. Ke, P. Liao, Z. Deng, H. Chen, and Y. Zhang, "LCANet: Learnable connected attention network for human identification using dental images," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 905–915, Mar. 2021.

[60] 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. Imag.*, vol. 39, no. 7, pp. 2374–2384, Jul. 2020.

[61] L. Banjšak, D. Miloševic, and M. Subašic, "Implementation of artificial intelligence in chronological age estimation from orthopantomographic X-ray images of archaeological skull remains," *Bull. Int. Assoc. Paleodontol.*, vol. 14, no. 2, pp. 122–129, 2020.

[62] S. Kim, Y.-H. Lee, Y.-K. Noh, F. C. Park, and Q.-S. Auh, "Age-group determination of living individuals using first molar images based on artificial intelligence," *Sci. Rep.*, vol. 11, no. 1, p. 1073, Jan. 2021.

[63] W. Upalananda, K. Wantanajittikul, S. Na Lampang, and A. Janhom, "Semi-automated technique to assess the developmental stage of mandibular third molars for age estimation," *Austral. J. Forensic Sci.*, pp. 1–11, Feb. 2021.

[64] M. Zaborowicz, K. Zaborowicz, B. Biedziak, and T. Garbowski, "Deep learning neural modelling as a precise method in the assessment of the chronological age of children and adolescents using tooth and bone parameters," *Sensors*, vol. 22, no. 2, p. 637, Jan. 2022.

[65] N. V. Blanco, R. R. Vilas, M. J. C. Nouche, and I. T. Carmona, "Towards deep learning reliable gender estimation from dental panoramic radiographs," in *Proc. 9th Eur. Starting AI Researchers' Symp. Located 24th Eur. Conf. Artif. Intell. (ECAI)*, vol. 2655, S. Rudolph and G. Marreiros, Eds. Santiago Compostela, Spain, Aug. 2020, pp. 1–8.

[66] 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. Imag.*, vol. 21, no. 1, pp. 1–11, Dec. 2020.

[67] D. Milosevic, M. Vodanovic, I. Galic, and M. Subasic, "Automated sex assessment of individual adult tooth X-ray images," in *Proc. 12th Int. Symp. Image Signal Process. Anal. (ISPA)*, Sep. 2021, pp. 72–77.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[69] M. V. Rajee and C. Mythili, "Gender classification on digital dental X-ray images using deep convolutional neural network," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102939.

[70] K. M. Keerthana, B. Rajeshwari, S. Keerthi, and H. P. Menon, "Classification of tooth type from dental X-ray image using projection profile analysis," in *Proc. Int. Conf. Signal Process. Commun. (ICSPC)*, Jul. 2017, pp. 394–398.

[71] H. Chen, K. Zhang, P. Lyu, H. Li, L. Zhang, J. Wu, and C.-H. Lee, "A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films," *Sci. Rep.*, vol. 9, no. 1, p. 3840, Mar. 2019.

[72] *Dentistry—Designation System for Teeth and Areas of the Oral Cavity*, ISO Standard 3950:2016, ISO Central Secretary, International Organization for Standardization, Geneva, Switzerland, 2016.

[73] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. IEEE Int. Conf. Mach. Learn. (ICML)*. Madison, WI, USA: Omnipress, Feb. 2010, pp. 807–814.

[74] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[75] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, San Francisco, CA, USA: AAAI Press, Feb. 2017, pp. 4278–4284.

[76] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[77] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1800–1807.

[78] 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* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2018, pp. 270–279.

[79] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017, pp. 5998–6008.

[80] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 11211, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 3–19.

[81] A. Koutsoukas, K. J. Monaghan, X. Li, and J. Huan, "Deep-learning: Investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data," *J. Cheminformatics*, vol. 9, no. 1, p. 42, Jun. 2017.

[82] S. M. LaValle, M. S. Branicky, and S. R. Lindemann, "On the relationship between classical grid search and probabilistic roadmaps," *Int. J. Robot. Res.*, vol. 23, pp. 673–692, Aug. 2004.

[83] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[84] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–15.

[85] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The marginal value of adaptive gradient methods in machine learning," 2017, *arXiv:1705.08292*.

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

[87] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021.

[88] A. Gulsahi, C. K. Kulah, B. Bakirarar, O. Gulen, and K. Kamburoglu, "Age estimation based on pulp/tooth volume ratio measured on cone-beam CT images," *Dentomaxillofacial Radiol.*, vol. 47, no. 1, Jan. 2018, Art. no. 20170239.

[89] R. G. Phulari, *Textbook of Dental Anatomy, Physiology and Occlusion*. JP Med. Ltd, New Delhi, India, Nov. 2013.

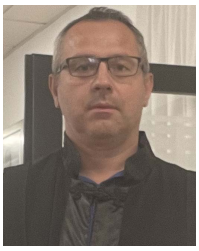
[90] R. C. Scheid, *Woelfels Dental Anatomy*, 9th ed. Philadelphia, PA, USA: Jones & Bartlett Learning, Mar. 2016.

[91] S. Risnes, Q. Khan, E. Hadler-Olsen, and A. Sehic, "Tooth identification puzzle: A method of teaching and learning tooth morphology," *Eur. J. Dental Educ.*, vol. 23, no. 1, pp. 62–67, Feb. 2019.

[92] M. Prados-Privado, J. G. Villalón, A. B. Torres, C. H. Martínez-Martínez, and C. Ivorra, "A convolutional neural network for automatic tooth numbering in panoramic images," *BioMed Res. Int.*, vol. 2021, pp. 1–7, Dec. 2021.

DENIS MILOŠEVIĆ was born in Böblingen, Germany, in 1993. He received the master's degree in computer science from the Faculty of Electrical Engineering and Computing, University of Zagreb, in 2017, with his graduate thesis titled "Object Tracking in Video Sequences Using Deep Networks," where he is currently pursuing the Ph.D. degree in medical image analysis with deep learning, under the mentorship of Dr. Marko Subašić. He attended the High School of Natural Sciences and Mathematics in Osijek. He is working as a Young Researcher in the Project DATACROSS at the Research Centre for Excellence in Data Science and Cooperative Systems.

MARIN VODANOVIĆ was born in Bochum, Germany, in 1975. He received the Ph.D. degree, in 2008. He specialized in endodontics and dental pathology. He is currently appointed as an Associate Professor and a Scientific Adviser with tenure at the School of Dental Medicine, University of Zagreb, where he has been serving as the Vice Dean, since 2015. He is also employed as a Specialist of dental pathology and endodontics at the University Hospital Centre Zagreb. His current research interests include anthropological aspects of teeth and mouth for age and sex estimation in forensic dentistry and bioarcheology.



IVAN GALIĆ was born in Sarajevo, Bosnia and Herzegovina, in 1973. He received the Ph.D. degree, in 2011. He specialized in oral surgery. He is currently appointed as an Assistant Professor with tenure at the School of the Medicine, University of Split. He is also employed as a Specialist of oral surgery at the University Hospital Centre Split. His main research interests include oral surgery, forensic dentistry, dental radiology, dental anthropology, and dental implantology. He is a member of the International Association for Paleodontology, Croatian Dental Society, Croatian Oral Surgery Society, Croatian Dental Implantology Society, and Croatian Dental Chamber.

MARKO SUBAŠIĆ (Member, IEEE) received the Ph.D. degree from the Faculty of Electrical Engineering and Computing, University of Zagreb, in 2007. Since 1999, he has been working at the Department for Electronic Systems and Information Processing, Faculty of Electrical Engineering and Computing, University of Zagreb, currently working as an Associate Professor. He teaches several courses at the graduate and undergraduate levels. His research interests include image processing and analysis and neural networks, with a particular interests in image segmentation, detection techniques, and deep learning. He is a member of the IEEE Computer Society, the Croatian Center for Computer Vision, the Croatian Society for Biomedical Engineering and Medical Physics, and the Centre of Research Excellence for Data Science and Advanced Cooperative Systems.