# Deep learning-based analysis of dental X-ray images for forensic estimation of age and sex

**Milošević, Denis**

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Denis Milošević

# DEEP LEARNING-BASED ANALYSIS OF DENTAL X-RAY IMAGES FOR FORENSIC ESTIMATION OF AGE AND SEX

DOCTORAL THESIS

Zagreb, 2022.

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Denis Milošević

# DEEP LEARNING-BASED ANALYSIS OF DENTAL X-RAY IMAGES FOR FORENSIC ESTIMATION OF AGE AND SEX

DOCTORAL THESIS

Supervisor: Professor Marko Subašić, PhD

Zagreb, 2022.

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Denis Milošević

# ANALIZA STOMATOLOŠKIH RENDGENSKIH SNIMAKA ZA FORENZIČKU ESTIMACIJU DOBI I SPOLA TEMELJENA NA DUBOKOM UČENJU

DOKTORSKI RAD

Mentor: prof. dr. sc. Marko Subašić

Zagreb, 2022

This doctoral thesis was completed at the University of Zagreb Faculty of Electrical Engineering and Computing, Department of Electronic Systems and Information Processing. This research has been supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

Supervisor: Professor Marko Subašić, PhD

The dissertation has: 153 pages

Dissertation number: _____

# O mentoru

Prof. dr. sc. Marko Subašić rođen je 13. travnja 1976. godine u Zagrebu. Diplomirao je na Fakultetu elektrotehnike i računarstva (FER) u Zagrebu, 1999. Magistrirao je elektrotehniku na FER-u u Zagrebu, 2003. Doktorirao je elektrotehniku na FER-u u Zagrebu 2007.

U znanstveno-nastavno zvanje docent na FER-u izabran je 2010. U znanstveno-nastavno zvanje izvanredni profesor na FER-u izabran je 2017.

Dr. Subašić bavi se istraživanjima u području digitalne obradbe i analize slika s primjenama u medicini, prometu, udaljenim mjerenjima te industriji te neuronskim mrežama, strojnim učenjem i drugim metodama umjetne inteligencije. Dr. Subašić je član grupe za digitalnu obradu slike pod vodstvom prof. Lončarića. Član je slijedećih profesionalnih organizacija: IEEE (Institute of Electrical and Electronics Engineers) i IEEE Computer Society, Znanstvenog centra izvrsnosti za znanost o podacima i kooperativne sustave, Centra izvrsnosti za računalni vid i Hrvatskog društva za medicinsku i biološku tehniku. Dr. Subašić je aktivno sudjelovao u organizaciji nekoliko međunarodnih konferencija (International Symposium on Image and Signal Processing and Analysis 2000, 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, 2019 i 2021). Sudjelovao je u nekoliko znanstvena projekta Ministarstva znanosti, Hrvatske zaklade za znanost, kompetitivnih EU projekata te komercijalnih projekata.

# About the Supervisor

Prof. dr. sc. Marko Subašić was born on April 13, 1976 in Zagreb. He graduated from the Faculty of Electrical Engineering and Computing (FER) in Zagreb in 1999. He received his master's degree in electrical engineering from FER in Zagreb, 2003. He received his PhD in electrical engineering from FER in Zagreb in 2007.

He was elected to the title of assistant professor at FER in 2010. In 2017, he was elected associate professor at FER.

Dr. Subašić is conducting research in the field of digital image processing and analysis with applications in medicine, transport, remote sensing and industry, as well as neural networks, machine learning and other methods of artificial intelligence. Dr. Subašić is a member of the group for digital image processing led by prof. Lončarić. He is a member of the following professional organizations: IEEE (Institute of Electrical and Electronics Engineers) and IEEE Computer Society, the Scientific Center of Excellence for Data Science and Cooperative Systems, the Center of Excellence for Computer Vision and the Croatian Society for Medical and Biological Engineering. Dr. Subašić has actively participated in the organization of several international conferences (International Symposium on Image and Signal Processing and Analysis 2000, 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, 2019, and 2021). He has participated in several scientific projects of the Ministry of Science, the Croatian Science Foundation, competitive EU projects and commercial projects.

# Acknowledgement

I would like to express sincere gratitude to my mentor, Professor Marko Subašić, Ph.D., for his unwavering support and guidance during all these years of research, as well as for the opportunity to work on such an interesting and exciting topic. I would also like to thank Professor Sven Lončarić, Ph.D., for the opportunity to work on the DATACROSS project, which enabled my research. I'm also grateful to Associate Professor Marin Vodanović, Ph.D., for enthusiastically lending his expertise whenever needed. Many thanks to my friends who had to hear more about neural networks and teeth than they ever imagined they would. To my loving mother and father, you have my endless gratitude for all the support you have given me over the years and for always having my back no matter what, rain or shine.

# Abstract

Forensic odontology is an important branch of forensic medicine focused on the analysis of dental remains. Dental remains are one of the most stable in regard to blunt force, fire, bacterial decomposition, and other external factors, making them a prime candidate for forensic analysis. The forensic analysis aims to identify human remains, with the most common steps being the estimation of age and assessment of sex. Assessment of sex, and especially age estimation in adults, are complex problems that require years of training to acquire the necessary expertise. Current methods of forensic odontology are based on time-intensive, precise manual measurements paired with reference tables and atlases, which can lead to human error, as minor differences can lead to a reproducibility problem. Those measurements can be taken directly on the remains, from casts, or from radiographic images, the latter being the most widely used approach due to the ease of handling and the ubiquitousness of panoramic dental x-ray imaging. With the rise of artificial intelligence and deep learning-based methods, computer vision has shown unprecedented success in many domains, including medical image analysis. Age estimation, sex assessment, and the additional tasks required to accomplish those tasks are prime candidates for computer vision and automation. This thesis explores the intersection of deep learning-based methods for computer vision and forensic odontology. Two neural network architectures are proposed, one based on current state-of-the-art image processing architectures used as feature extractors and attention, and one from the ground-up built around the BiFPN module for age estimation and sex assessment from dental x-ray images. The hyperparameter space is exhaustively examined, and the strengths and weaknesses of the best models are analyzed and compared to the current state of literature. Models are developed for age estimation, sex assessment, tooth detection and type determination, and segmentation of teeth and dental alterations (tooth decay, crowns, fillings, and root canal fillings) for panoramic and individual tooth x-ray images. Multiple variants are trained for each task, either as a different approach to achieving the same goal or with the intent of understanding the model's inner workings. Unlike classic forensic odontology methods, the proposed methods extend to imperfect teeth that can have alterations, and the impact of those alterations on the performance is extensively examined. All the proposed methods are evaluated on a dataset of panoramic or individual dental x-ray images of adults between the ages of 19 to 90 years. The results suggest that the proposed methods perform as well or better than current methods in literature while being fully automated, reproducible, and multiple orders of magnitude faster.



**Keywords**: Computer vision, Artificial Intelligence, Neural network, Forensic odontology, Age estimation, Sex assessment, Tooth type determination, Teeth detection, Segmentation of teeth and dental alterations

# Analiza stomatoloških rendgenskih snimaka za forenzičku estimaciju dobi i spola temeljena na dubokom učenju

Forenzička odontologija je grana medicinske forenzike koja se bavi dentalnim posmrtnim ostacima s ciljem identificiranja osobe. Prvi koraci u forenzičkom procesu je određivanje demografskih podataka osobe - dobi i spola. Iako ne uvijek, u velikom broju slučajeva su te dvije informacije dovoljne za identificiranje osobe. Određivanje dobi i spola je kompleksan problem koji zahtjeva višegodišnje usavršavanje stručnjaka. Metode određivanja dobi i spola su kompleksne, te zahtijevaju značajnu količinu vremena i precizno ručno mjerenje dimenzija zuba i zubnih struktura. Mjerenja se mogu raditi na samim dentalnim posmrtnim ostacima, na gips-kopijama ili na radiografskim snimkama, pri čemu su radiografske snimke, posebice panoramske dentalne rendgenske snimke, najpopularnije. Te mjere se koriste u kombinaciji s raznim linearnim modelima, referentnim tablicama i atlasima, ovisno o dodatnim procjenama koji stručnjak forenzičke odontologije radi tijekom analize. Male razlike u mjerama i pretpostavkama, i male greške kod usporedbe s referentnim tablicama mogu dovesti do pogrešne procjene, što u konačnici uvodi potencijalnu ljudsku grešku u forenzički proces. Umjetna inteligencija i metode bazirane na dubokom učenju su doveli do revolucije na području računalnog vida. Primjena dubokih neuronskih mreža se proširila po mnogobrojnim područjima, što uključuje i medicinu, gdje modeli dubokog učenja vrše analizu medicinskih snimki. Forenzička odontologija, posebice procjena dobi, spola i popratnih zadataka, su odličan kandidat za automatiziranje pomoću metoda dubokog učenja. Ova doktorska disertacija istražuje upravo to interdisciplinarno područje umjetne inteligencije i metoda forenzičke odontologije za procjenu dobi i spola.

Prvo poglavlje je uvod u disertaciju, i daje opis i širu sliku problema, te objašnjava motivaciju za pristup za istraženi pristup. Objašnjeni su i znanstveni doprinosi disertacije, i pregled na koji način su ostvareni. Znanstveni doprinosi disertacije se fokusiraju na izradu modela temeljenih na dubokom učenju za zadatke forenzičke odontologije za procjenu dobi i spola. Također je u prvom poglavlju dan pregled strukture disertacije.

Drugo poglavlje postavlja temelje na kojima su građene metode zasnovane na dubokom učenju. Analiziran je perceptron, prototip gradivnih jedinica modernih neuronskih mreža. Perceptronova osnovna formulacija je snažan klasifikator, ali se tek u konfiguraciji beskonačne širine ili beskorisne dubine dokazuje teorem univerzalnog aproksimatora. Teorem univerzalnog aproksimatora kaže da se niz perceptrona u konfiguraciji beskonačne širine ili dubine može biti proizvoljno točan aproksimator bilo koje funkcije. Da bi se takva neuronske mreža mogla naučiti, potrebno je uvesti algoritam unazadne propagacije greške, tzv. *backpropagation* algoritam. Algoritam unazadne propagacije greške je primjena lančanog pravila derivacije. Pomoću tog algoritma se računa derivacija greške u odnosu na svaki parametar, te se koriste metode gradijentnog spusta za optimizaciju. Klasični gradijentni spust pomiče parametre neuronske

mreže u smjeru suprotnom od smjera gradijenta, prema minimumu. Kako se radi o iznimno nekonveksnom problemu, razvijene su metode kao Adam, koje uvode dodatne parametre i momente, te tako izbjegavaju neke probleme koje mogu nastati tijekom treniranja modela temeljenog na dubokom učenju. Ti novouvedeni parametri se zovu hiperparametri, i razlikuju se od parametara neuronske mreže po tome što se ne podešavaju tijekom treniranja niti pomoću optimizacijskog algoritma, već se moraju podesiti "ručno". Mogu se koristiti algoritmi pretraživanja za efektivno pronalaženje dobrih vrijednosti hiperparametara, kao što su Bayesovo pretraživanje, pretraživanje po rešetki, i slučajno pretraživanje. Konačno, dobiveni modeli su tzv. "black-box" modeli, modeli koje ne možemo objasniti konvencionalnim metodama. Metode prekrivanja i gradijentske metode kao GradCAM omogućavaju uvid u rad modela. U drugom dijelu poglavlja se radi analiza trenutnog stanja literature. Pregled razvoja područja forenzičke odontologije pokazuje kako se rano otkrilo da su kosti i zubi korisni indikatori za procjenu dobi i spola, te da su zubi značajno stabilniji na utjecaj okoline od drugih dijelova tijela. Rane metode su uništavale posmrtne ostatke, ali se proširivanjem tehnologije radiografskog snimanja taj problem uklonio. Danas se koristi 6 načina radiografskog snimanja dentalnog sustava, ali je najučestalija metoda panoramska dentalna radiografska snimka. Za procjenu dobi je došlo do razvoja tri fundamentalna pristupa, Kvaal et al., Drusini et al., i Cameriere et al., na kojima je većina današnjih metoda temeljena. Prvi pokušaji za procjenu dobi metodama temeljenim na dubokom učenju daju obećavajuće rezultate, ali su većinom grube kategorizacije u široke dobne skupine ili su fokusirane na djecu, što je značajno lakši problem. Za procjenu spola ne postoje fundamentalne metode, ali postoji širok raspon metoda temeljen na mandibularnim i drugim dentalnim mjerama i ne-metričkim indikatorima. Postoje i metode koje pokušavaju ostvariti procjenu spola pomoću metoda zasnovanih na dubokom učenju, koje koriste posebno razvijene jednostavne arhitekture neuronskih mreža, i koje daju obećavaju rezultate. Određivanje tipa zuba je standardiziran postupak u forenzičkoj odontologiji, a za metode zasnovane na dubokom učenju se obično kombinira detekcija zuba s određivanjem tipa koristeći razne klasifikacijske sustave za tip. Istraživanja o segmentacijskim modelima zasnovanim na dubokom učenju su još u početcima, gdje različita istraživanja dobivaju obećavajuće prve rezultate segmentacije zuba, dok segmentacija dentalnih intervencija nije značajno istražena.

Treće poglavlje se bavi analizom i pripremom podataka koji se koriste za izradu i evaluaciju metoda predloženih u ovoj disertaciji. Inicijalni skup slika se sastoji od 4035 panoramskih dentalnih rendgenskih snimki od ukupno 3994 osobe. Omjer ženskih i muških uzoraka je 58.7% : 41.3%, a dobni raspon je od 19 do 90 godina. Skup podataka je podijeljen na skup za učenje, validaciju i testiranje na temelju osoba, tako da se ista osoba ne može pojaviti u npr. skupu za treniranje i testiranje. U podatkovnom skupu 2899 slika ima oznake pozicija zuba, koje su iskorištene za izradu dodatnog skupa podataka rendgenskih slika pojedinih zubi koji se sastoji od 86495 slika. Podskup od 7630 slika ima dodatne anotacije zubnog statusa, koje se koriste

za evaluaciju utjecaja dentalnih intervencija na performanse modela. Također je dodatni podskup panoramskih dentalnih slika, njih 813, označeno s detaljnim segmentacijskim mapa koje uključuju svaki pojedini zub posebno, te sve dentalne intervencije koje su vidljive na snimci.

Četvrto poglavlje opisuje i analizira arhitekture neuronskih mreža koje su korištene za realizaciju istraživanja. Prva arhitektura spaja mehanizam pažnje s najuspješnijim arhitekturama za analizu slika u literaturi, koji se koriste kao ekstraktor značajki. Ta arhitektura ima četiri hiperparametra: koja se arhitektura koristi za izdvajanje značajki, broj kanala u zadnjem konvolucijskom sloju, mogućnost isključivanja mehanizma pažnje, i broj neurona u potpuno povezanom sloju. Druga arhitektura je građena oko BiFPN modula, načina za korištenje značajki na više dimenzija za izračun konačne procjene. Ta arhitektura ima šest hiperparametara: dva koja definiraju veličinu početnih konvolucija, broj BiFPN slojeva u modulu, broj značajki koje BiFPN računa, broj kanala koji vraća mehanizam kompresije, te broj neurona u zadnjem potpuno povezanom sloju. Iako obije arhitekture imaju isti zadatak, druga arhitektura je definirana sa značajno manjim brojem podesivih parametara, njih 2-5% u odnosu na prvu arhitekturu, što je čini manje zahtjevnom i bržom za računanje.

Peto poglavlje daje detaljan uvid u metodologiju disertacije, pregled kombinacija algoritama pretraživanja i detalje pristupa za svaki pojedini forenzički zadatak. Bayesov algoritam pretraživanja može smanjiti ukupni broj evaluiranih modela koristeći Bayesov teorem, ali se evaluacije izvršavaju uzastopno. Kako je svaka evaluacija zapravo treniranje modela do konvergencije, taj postupak zahtjeva ogromnu količinu vremena. Predložena je kombinacija algoritma pretraživanja po rešetki i slučajnog pretraživanja. Algoritam pretraživanja po rešetki može dati dobar pregled ograničenog potprostora pretraživanja, a izvršavanje se može vršiti paralelno. Slučajno pretraživanje se čini neintuitivnim, ali se analizom vjerojatnosti može pokazati da je za 60 iteracija vjerojatnost da je pronađen rezultat unutar 5% od najboljeg rezultata u ograničenom potprostoru jednaka 95%. Povećavanjem brojem evaluiranih modela se povećava vjerojatnost ili smanjuje tolerancija odstupanja, što omogućava efektivno prilagođavanje metode pretraživanja dostupnim resursima. Kombinacijom slučajnog pretraživanja i pretraživanja po rešetki se dobiva iterativni algoritam koji može većinu evaluacija izvršavati paralelno, a vjerojatnosna analiza pokazuje vjerojatnosti pronalaska modela zadovoljavajućih performansi. Predložena su dva načina treniranja, brzi i spori. Brzi način treniranja koristi adaptivne optimizacijske metode i veću stopu učenja, čime se dobivaju modeli lošijih performansi, ali jednakih relativnih performansi. Drugim riječima, ako dva modela treniramo brzo i sporo, model koji je bio bolji tijekom brzog treniranja će biti bolji i nakon sporog treniranja, iako će oba modela imati bolje performanse od brzo treniranih pandana. Spori pristup koristi osnovni gradijenti spust, jer istraživanje pokazuje da ne-adaptivni optimizatori dovode do modela koji bolje generaliziraju. Također se koristi funkcija koja računa stopu učenja ovisno o epohi, tzv. "cosine annealing with warm restarts", koja zajedno s gradijentnim spustom daje u prosjeku modele boljih rezultata.

Za sporo treniranje je dodana i augmentacija slika, postupak dodavanja šuma i generiranja više raznolikosti u skupu podataka. Svim forenzičkim zadacima se pristupa na dva načina - procjena iz cijele panoramske dentalne snimke, i procjena iz rendgenske snimke pojedinog zuba. Za procjenu dobi se radi procjena regresijom, ali se radi i procjena pomoću distribucije vjerojatnosti. Na temelju procijenjene distribucije se računa očekivana vrijednost, koja se uzima kao procjena dobi, a standardna devijacija distribucije se koristi kao regularizacijski faktor. Za dob se također treniraju modeli za tri anatomske regije interesa radi utvrđivanja doprinosa pojedinog dijela dentalnog sustava procjeni, te se provjerava doprinosi li stratifikacija po dobi boljim rezultatima. Procjeni spola i određivanju tipa se pristupa kao klasifikacijskom problemu. Određivanje tipa se radi samo na temelju rendgenske slike pojedinog zuba, ali se treniraju modeli za četiri različita klasifikacijska sustava, 32-klasni, 16-klasni, 8-klasni i 4-klasni sustav. Za segmentaciju zubi i dentalnih intervencija su evaluirane 3 arhitekture (UNet, FCN, i DeepLab v3) i njihove varijante, što je ukupno 12 arhitektura. Pronađeni su najuspješniji modeli za svaki segmentacijski slučaj (zubi i četiri intervencije). Testira se segmentacija cijele panoramske dentalne rendgenske snimke, rendgenske snimke pojedinog zuba, i segmentacija intervencija na rendgenskoj snimci pojedinog zuba s uklonjenim okolnim tkivom. Za detekciju zuba je evaluirana trenutno najuspješnija mreža za detekciju objekata, YOLO v5. Evaluirane su sve varijante tog modela, njih osam, te su testirane i uspješnosti temeljene na veličini ulazne slike. Analizirana je varijanta kada se samo pronalazi zubi, ali su dodatno evaluirani modeli detekcije koji paralelno s detekcijom rade i procjenu tipa zuba, za svaki od 4 klasifikacijska sustava.

U šestom poglavlju se daje pregled rezultata istraživanja, radi se iscrpna analiza rezultata i performansi modela, te se uspoređuje uspješnost s trenutnim stanjem u literaturi. Za direktnu procjenu dobi iz panoramske dentalne rendgenske slike se pokazuje da model s augmentacijom i bez stratifikacije radi najbolje, i da se postiže prosječna greška od 3.96 godina, i medijalna greška od 2.95 godina. S porastom godina greška raste. Zubi i okolno koštano tikovo nezavisno doprinose procjeni, a zubi su indikativniji. Procjenom pomoću distribucije vjerojatnosti se postižu bolji rezultati, te prosječna greška pada na 3.60 godina, a medijalna greška pada na 2.76 godina. Ti rezultati pokazuju da predložena metoda daje bolje rezultate od trenutnog stanja u literaturi. Procjena dobi iz rendgenske snimke pojedinog zuba radi lošije, ostvarujući prosječnu grešku od 6.55 godina i medijalnu grešku od 5.32 godine. Za zdrave zube bez intervencija ta greška pada na 6.15 godine (prosjek) i 4.94 godine (median). Intervencije u prosjeku pogoršavaju rezultate, a ispun korijena to čini najviše. U odnosu na stanje u literaturi, predloženi pristup ostvaruje jednake rezultate, ali je istovremeno u potpunosti automatiziran i podržava analizu zuba s intervencijama. Za procjenu spola se vidi sličan trend s godinama, gdje su rezultati za starije uzorke u prosjeku lošiji. Model temeljen na BiFPN modulu postiže bolje rezultate, unatoč tome što je 45 puta manji od modela s pažnjom. GradCAM analiza interpretabilnosti pokazuje da je model naučio prepoznati spol na temelju anatomski relevantnih

regija. U odnosu na trenutno stanje u literaturi, predloženi pristup ostvaruje najbolje rezultate. Za panoramske dentalne rendgenske slike postignuta je točnost od 97.04%. Procjena spola iz slike pojedinog zuba je kompleksnija zadaća, jer slika zuba sadrži značajno manje informacija. To se reflektira i u rezultatima s općom točnosti od 75.44%, i s točnosti od 76.41% na zdravim uzorcima bez intervencija. Intervencije pogoršavaju rezultate, ponajviše ispuni korijena. Grad-CAM analiza pokazuje da se najviše pažnje posvećuje kruni zuba. U odnosu na trenutno stanje u literaturi, predloženi pristup postiže jednake rezultate, ali je u potpunosti automatiziran i nekoliko redova veličine brži. Evaluacijska segmentacijskih modela je pokazala da je velika varijanta UNeta gotovo uvijek najuspješnija za sve zadatke, i da to vrijedi za segmentaciju panoramske rendgenske snimke i rendgenske snimke pojedinog zuba u obije inačice. Intervencija koja se najbolje segmentira je kruna. U usporedbi s trenutnim stanjem literature, rezultati segmentacije pojedinog zuba su jednaki ili bolji od dosadašnjih istraživanja, rezultati segmentacije karijesa su lošije, a istraživanja segmentacija drugih intervencija trenutno nema u literaturi. Detekcija zuba je najuspješnija s modelom srednje složenosti i najvećom mogućom slikom ortopana. Ako se istovremeno određuje tip, rezultati se poboljšavaju, pri čemu su najbolji rezultati za 4-klasni pristup. Slično vrijedi i za određivanje tipa zuba iz rendgenske snimke pojedinog zuba, gdje pristupi s manje klasa imaju bolje rezultate od pristupa s više klasa. U odnosu na trenutno stanje u literaturi, rezultati predložene metode su jednaki ili bolji.

Sedmo i zadnje poglavlje je zaključak. Iz rezultata istraživanja se može zaključiti da su predložene arhitekture valjane, te da je BiFPN arhitektura gotovo jednako uspješna unatoč tome što je i do 40 puta manja od uobičajenih. Pristup pretraživanja i podešavanja hiperparametara kombinacijom pretraživanja po rešetki i slučajnog pretraživanja je uspješan. Razlike između brzog i sporog treniranja su značajne, pri čemu je međusobni poredak uspješnosti modela podjednak, što ubrzava kompletni proces pretraživanja i treniranja. Rezultati procjene dobi su u svakom slučaju bolji od trenutnog stanja u literaturi, a pristup s procjenom vjerojatnosne distribucije i regularizacije pomoću standardne devijacije te distribucije daje značajno bolje rezultate. Procjena spola je također uspješna, pri čemu je procjena iz panoramske snimke značajno bolja od procjene iz pojedine rendgenske snimke zuba. Tip se također može uspješno odrediti neuronskom mrežom, pri čemu je uspjeh veći za klasifikacijske pristupe s manje klasa. Isti je slučaj za detekciju. Intervencije u prosjeku pogoršavaju točnost svih procjena, posebice ispun korijena. Segmentacijski modeli su također uspješni, kako za pojedini zub, tako i za dentalne intervencije. Svi predloženi pristupi, osim segmentacije karijesa, rade jednako dobro ili bolje od trenutnog stanja u literaturi, dok su istovremeno u potpunosti automatizirani, reproducibilni, i nekoliko redova veličine brži od trenutnih metoda.

**Keywords**: Umjetna inteligencija, Neuronske mreže, Forenzička odontologija, Procjena dobi, Procjena spola, Određivanje tipa zuba, Detekcija zuba, Segmentacija zuba i dentalnih

intervencija

# Contents

# Chapter 1

# Introduction

## 1.1 Overview

Forensic odontology is a subspecialty of dentistry that focuses on the identification of a person through the analysis of dental remains. The methods of forensic odontology are standard tools in the forensic process, as they can reliably provide demographic information, age and sex, about the remains of a person. Human dental remains are among the strongest and most resilient, showing very high resistance to external damage, be it anything from direct application of force to exposure to fire, and additionally also being very resistant to bacterial decomposition, making them a prime candidate for forensic examination [1, 2, 3]. While often used in judiciary cases, forensic odontology has a broader application, as it is a permanent fixture in archaeology and anthropology. Age and sex are expressed differently in different people, yet some commonalities are found across all humans, serving as the basis for forensic analysis. During our early days, dental development is on a strict, genetically defined schedule. Minor variations do occur from person to person and even from population to population, but those discrepancies never deviate more than a few months. However, once the teeth mature and their development stops, all developmental indicators cease being useful. Nevertheless, estimating the age and assessing the sex is a necessity in modern forensics, and dental remains are the prime candidate for this task. By extension, determining the tooth type is a sub-task of both as observations differ between different tooth types, and tooth alterations need to be ascertained to avoid their negative impact on the forensic analysis.

All current forensic odontology methods for age estimation and sex assessment, be they applied directly on dental remains, dental casts, or radiographic images (most often panoramic dental x-ray images), require manual measurements. Those measurements are painstakingly taken by highly trained forensic odontology experts, which can take between 25 to 35 minutes per sample. Once the measurements are completed, they are used to calculate indicative features, like dimension ratios of different anatomical parts, which are then compared to vari-

ous reference tables based on some other observations and assumptions. Nonmetrical methods avoid those repetitive and tedious by directly comparing the morphology to some reference atlases, usually yielding worse results. Reference tables, atlases, and the linear models developed alongside them, while in principle equal, also differ from population to population. Still, all approaches suffer from their common element - human error. Measurements can be taken slightly differently, and visual matching to reference atlases is subjective, ultimately leading to minor discrepancies that can make results hard to reproduce. Current studies in literature also form their models, for deep learning scales, on small sample sizes, and they have strict exclusion criteria based on any alterations a tooth can have. The need to establish if and which alterations a tooth has and the need to determine the tooth type to apply known methods properly can further exacerbate the complexness of forensic odontology analysis.

With the breakthrough success of convolutional neural networks, deep learning-based methods have started to make their way into medical applications like medical image analysis [4, 5]. The cornerstone methods of contemporary forensic odontology for estimating age and assessing sex are based on radiographic imaging. Highly trained experts take the required mandibular and dental anatomy measurements from panoramic dental x-ray images. Such a medical image analysis problem is an outstanding candidate for modern computer vision approaches. Specifically, age estimation can be formulated as a regression problem, sex assessment and tooth type determination can be formulated as a classification problem, tooth alteration analysis can be formulated as a segmentation problem, and individual teeth can be isolated if the task is formulated as a detection problem. More importantly, all those tasks can be expressed as supervised learning problems, reducing the amount of data needed and allowing for proper, exhaustive, and accurate performance evaluation. Deep learning-based approaches for age estimation, sex assessment, tooth type determination and detection, and tooth alteration segmentation would empower forensic odontology experts to focus on more complex tasks, while the tedious and repetitive tasks are done by the deep learning model in a fraction of the time.

## 1.2 Scientific contributions

This thesis focuses on the forensic odontology tasks of age estimation and sex assessment from panoramic dental x-ray images using deep learning-based methods. As sturdy as dental remains are, situations where only some teeth are available instead of the entire mandible can occur. Therefore, multiple deep learning-based models for age estimation and sex assessment are developed, which can perform their task either on full panoramic dental x-ray images or x-ray images of individual teeth. Two neural network architectures are proposed, one based on the recently very successful attention mechanism [6], and the other engineered around the BiFPN module. To comprehensively analyze the performance of those models, result trends,

strengths, weaknesses, biases, and to ultimately gain a complete understanding of their capabilities, additional models have been trained focused on the determination of tooth type, the segmentation of teeth and dental alterations, and the detection of individual teeth in panoramic dental x-ray images. The work in this thesis expands its aim by including teeth with dental alterations, which are conventionally excluded from studies as they are deemed unfit for forensic analysis. The tooth type is an important data point that experts require to apply contemporary forensic odontology methods; automation of this forensic task can improve the efficiency and reliability of already existing methods. In addition to the direct regression approach, an approach that estimates the age's probability distribution is proposed for age estimation. This approach effectively introduces the models' certainty in its proposed result, which has a positive knock-on effect on overall performance. To conclude, the scientific contributions of this thesis can be summarized as:

1.Classification model for sex assessment from dental x-ray images based on deep convolutional neural network.

2.Regression model for age estimation from dental x-ray images based on deep convolutional neural network.

3.Deep learning-based region proposal model for tooth status assessment from dental x-ray images.

## 1.3 Organization of the thesis

The thesis is divided into seven chapters. The first chapter introduces the forensic odontology problems of age estimation and sex assessment, as well as the auxiliary tasks of tooth type determination, tooth and dental alteration segmentation, and tooth detection. The scientific contribution of this thesis is discussed, and this very overview of the thesis structure is given. The second chapter is an introduction to deep learning-based methods, neural networks, modern computer vision, an overview of forensic odontology methods for age estimation, sex assessment, tooth type determination, tooth and dental alteration segmentation, and tooth detection, and a review of the current state of literature. This chapter lays the foundation upon which this thesis is built. Data is the basic building block to developing deep learning-based approaches successfully. The third chapter fully explores the data used in this thesis, discusses its features and biases, and gives a detailed breakdown of how the data was processed and prepared for this research. The fourth chapter presents the two proposed neural network architectures. The first section of Chapter 4 describes the construction of the neural network architecture based on state-of-the-art image analysis feature extractors and attention and the mechanisms built-in to scale the model's capacity effectively. The second section of Chapter 4 goes into detail about the motivation and construction of the BiFPN-based model, the neural network architecture developed for forensic

odontology image analysis entirely engineered around the BiFPN module. The fifth chapter meticulously explains the methodology used to develop the deep learning-based approaches and the developed approaches themselves. Finally, the sixth chapter demonstrates the results of the developed deep learning-based approaches. It exhaustively analyzes the models' results, performance, and interpretability, discusses the proposed methods' strengths and weaknesses, and compares them with current methods found in the literature. The seventh chapter serves as a conclusion to this thesis, giving a broad-strokes overview of the developed methodology and placing this thesis in the context of the current state of forensic odontology for the tackled forensic tasks.

# Chapter 2

# Computer vision, medical image analysis, and forensic odontology

## 2.1 Computer vision, image analysis, and neural networks

From its humble beginnings till now, computer vision is a ubiquitous technology used in many facets of modern life, be it in industrial applications, medical research, or even astrophysics research. From the early theoretical work in the mid-1960s during MIT's Project on Mathematics and Computation, over the first proper practical applications like the Viola-Jones detector [7], the eventual rise of deep learning-based methods like neural networks is in thanks to Rosenblatt's research and his formulation of the "perceptron" [8, 9]. The perceptron, in its basic form, is a binary classifier defined as:

$$f(\mathbf{x}) = \begin{cases} 1 & \mathbf{w} \cdot \mathbf{x} + b > 0 \\ 0 & otherwise \end{cases} \tag{2.1}$$

Where $\mathbf{x}$ is the input vector, $\mathbf{w}$ are the learnable parameters, and $b$ is the learnable bias. While this is a learning-based approach, this still does not qualify as "deep" learning, nor is the perceptron capable of learning any arbitrary task. There are two configurations of perceptrons that avoid this issue according to the universal approximation theorem, the "arbitrary width" [10] and "arbitrary depth" [11] cases. It might appear that multiple perceptrons would not improve the performance, as it could be interpreted as multiple linear operations in a row, which can be simplified to a single linear operation. Activation functions are functions that introduce non-linearity to the perceptron, like the sigmoid function and the Rectified Linear Activation function (ReLU [12, 13]). They non-linearly transform the weighted sum, thus preventing this reduction of linear operations. The two cases of the universal approximation theorem are enabled by the introduction of backpropagation [14], the learning algorithm which uses gradient

descent to propagate the errors through the network in order to adjust the learnable weights. The infinite case was shown to work in theory; it is, however, practically unfeasible. Soon after the initial universal approximation theorem studies, it was shown that neural networks with a minimized number of free parameters could enhance the generalization ability of the neural network [15].

All this opened the door to applying neural networks to different domains, which for images resulted in convolutional neural networks [16]. These theoretical fundamentals were proven early, but the required raw compute power was unavailable, resulting in a two-decade-long "AI winter." The great change came with the implementation of backpropagation for Graphical Processing Units (GPUs) [17]. Graphical Processing Units fundamentally differ from Central Processing Units (CPUs). Central Processing Units are general processors, initially with no parallelization, their performance mainly defined by their clock speed. Graphic Processing Units are specifically designed for the calculations required to construct an image. Computer graphics as a field of computer science is very complex, but to (overly) simplify, it relies on fast matrix multiplications. To achieve high performance for matrix multiplication, GPUs have an extremely high degree of parallelization, albeit at a slower clock speed. Computations in neural networks are essentially built on matrix multiplications and other highly parallelizable operations, making neural networks and GPUs a great match. Contemporary hardware is designed with neural networks in mind, extending their instruction sets to accommodate their requirements, all of which ultimately lead to a speedup of multiple orders of magnitude and helped to usher in the universal success of neural networks.

Learning-based approaches are *learning*-based because their parameters are adjusted, i.e. *learned* from some data. The most intuitive approach is supervised learning, where the model learns to map from the domain of the input to the domain of the target directly. Formally, a learning based model $f(\mathbf{x}|\theta)$ is defined as $f : \mathbb{R}^n \to \mathbb{R}^m$, where $\mathbf{x}$ is the input data, $\theta$ are the learnable parameters of the model, $n$ is the dimension of the input data, and $m$ is the dimension of the target data. For example, a deep learning-based model $f$ which estimates the age from a 256 by 256 px x-ray image of a tooth will have an input dimension of $n = 256^2$ and a target dimension of $m = 1$ (one number, the estimated age). To find the optimal model parameters $\theta$ for some objective function $L$, the expected loss $\mathscr{L}^*(\theta)$ needs to be defined over the data distribution $p_{\text{data}}$ as:

$$\mathscr{L}^*(\theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{\text{data}}} L(f(\mathbf{x}|\theta), \mathbf{y}) \tag{2.2}$$

However, it is in practice unfeasible to determine the data distribution $p_{\text{data}}$. Thus, the expected loss needs to be approximated. It is approximated on a finite set of independent and identically

distributed samples from the data distribution as follows:

$$\mathscr{L}(\theta) = \frac{1}{N}\sum_{i=0}^{N} L(f(\mathbf{x}_i|\theta), \mathbf{y}_i) \tag{2.3}$$

$N$ is the number of samples, and $(\mathbf{x}_i, \mathbf{y}_i)$ is the i-th sample from the finite dataset. The assumption that the samples are independent and identically distributed means that a) a sampled data point does not influence the data points sampled in the future, and b) that they are sampled from the same data distribution $p_{\text{data}}$. The more samples are available, the closer the approximation is to the optimal solution, leading to the optimal model parameters. As all operations in a neural network are differentiable, the backpropagation algorithm [14] can calculate gradients for all learnable weights, which is then used for optimization. The optimization problem is formally defined as:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \frac{1}{N}\sum_{i=0}^{N} L(f(\mathbf{x}_i|\theta), \mathbf{y}_i) \tag{2.4}$$

Originally, the most commonly used optimization algorithm was Stochastic Gradient Descent (SGD), but a wide variety of gradient descent optimization algorithms for neural networks have been developed [18]. SGD is the most intuitive approach, as it directly moves the model parameters in the opposite direction of the gradient towards a minimum. Formally, SGD can be summarized as:

$$\theta_{\mathbf{t+1}} = \theta_{\mathbf{t}} - \eta \nabla_\theta \mathscr{L}(\theta_{\mathbf{t}}) \tag{2.5}$$

In this formulation, $\theta_t$ are the learnable parameters of the model at step $t$, $\eta$ is the learning rate, and $\nabla_\theta \mathscr{L}(\theta_{\mathbf{t}})$ are the calculated gradients. The learning rate is important, as very large values will not allow for convergence, and very small values significantly slow down the optimization process and tend to converge in local minima that generalize poorly. Adaptive optimization methods, methods that modify the learning rate during the optimization process, have become more popular as they tend to produce sufficient results in less time and require less learning rate tuning. These approaches modify the effective learning rate in different ways, depending on the specific algorithm in question. The most popular and generally successful optimizer is Adam [19]. It utilizes exponentially moving averages and introduces multiple additional

hyperparameters alongside the learning rate. Formally, Adam is defined as:

$$m_t = \beta_1 m_t + (1-\beta_1) g_t \tag{2.6}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{2.7}$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \tag{2.8}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{2.9}$$

$$\theta_t = \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \tag{2.10}$$

In this formulation, $\beta_1$ and $\beta_2$ are hyperparameters that control the moving average, $g_t$ is the gradient at step $t$, and $\varepsilon$ is an arbitrarily small value used for numerical stability. Despite the additional control given by the inclusion of the additional hyperparameters, research shows that models trained with pure stochastic gradient descent with a properly tuned learning rate generalize better than adaptive methods [20]. However, research also shows that a fixed learning rate schedule improves overall model performance [21]. While this will not outperform a perfectly tuned learning rate for SGD, it will improve the average outcome. As with optimization algorithms based on gradient descent, many learning rate scheduler approaches have their own strengths and weaknesses. Cosine annealing with warm restarts is one such approach, borrowing motivations from simulated annealing [22] by introducing a period length in which the learning rate gradually decreases until it snaps back to its initial value. This thesis extensively uses this learning rate scheduler, and a deeper analysis of this learning rate scheduler is shown in Chapter 5, Section 5.2.3.

Both optimizers and learning rate schedulers introduce multiple hyperparameters as a trade-off. A *hyperparameter* is defined as a parameter of the optimization process which is not adjusted by the optimization process. For example, those are parameters like the learning rate, $\beta_1$ and $\beta_2$ for the Adam optimizer, and the length of the cooling-down period for cosine annealing with warm restarts. While they might seem mundane or even unimportant at first, research shows that proper tuning of those hyperparameters is an extremely significant contributor to the trained models' performance [23]. In the model architecture, every connection that exists or does not exist can unto itself be considered a hyperparameter. Hyperparameters define a space wherein each point is a possible solution; some solutions are good, and some are not. As those parameters are not differentiable, and as they do not form a convex problem, therefore classic search algorithms are used. Every potentially good solution needs to be evaluated, which in practice means that a model needs to be trained with those hyperparameters, making the search slow and expensive. Bayesian search uses the Bayesian theorem to create and update a model of the hyperparameter solution space and propose which hyperparameters are worth evaluating [24]. Grid search [25] evaluates equidistant solutions in a subspace of the hyperparameter

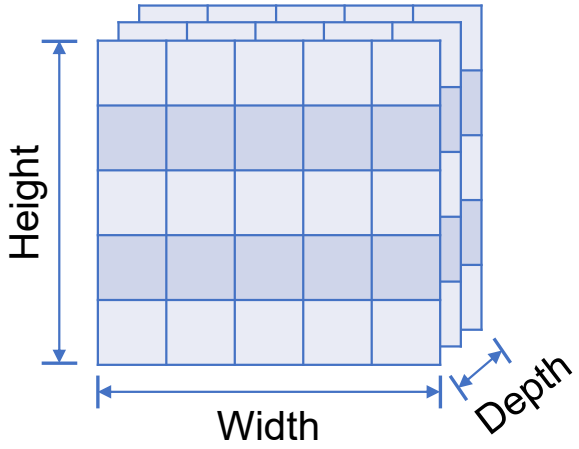


**Figure 2.1:** A representation of a 3D tensor. A 3D tensor can be imagined as a stack of 2D images of the same width and height. The number of images in this stack is the depth of the tensor, which can also be described as *the number of channels*. In computer vision, a tensor of this shape is sometimes called a *feature map*.

space, effectively spanning an N-dimensional grid in hopes of finding a good solution. The approach relies on a good subspace selection and a high sample count. Random search [26], as its name suggests, evaluates random points, which at first sight does not seem like a good strategy, but gives strong statistical guarantees and works well in practice. A benefit of the grid and random search compared to Bayesian search is that all evaluations are fully independent, allowing for strong parallelization of the search. This thesis makes strong use of hyperparameter optimization, and a detailed analysis of the proposed search solution is shown in Chapter 5, Section 5.2.

Fully-connected networks, as have been proposed by Rosenblatt [8, 9] are theoretically universal approximators, but as shown by LeCun [16], a minimization of free parameters is in practice good for generalization. The search and model space can be significantly reduced by introducing a set of inductive biases, which is achieved by introducing different types of layers. The structure of images is well known, as are some useful properties. These can and have been used to construct a neural network layer that is well suited for the task of image analysis. Images are 3D structures (width, height, and color channel), and related pieces of information are usually close to each other. In less abstract terms, in a picture of a dog, all the pixels related to the dog will be close together - they show the dog "object." This mathematical structure is also called a tensor. A visualization and explanation of a tensor and its dimensions is shown in Figure 2.1.

The neural network thus does not need to look all over the image to extract relevant information about; it can thus focus on a pixel's immediate neighborhood. The convolutional layer has been constructed to exploit those properties. A convolutional layer is defined by an N by N matrix of learnable parameters called a kernel, which processes its input (a tensor) in a sliding-

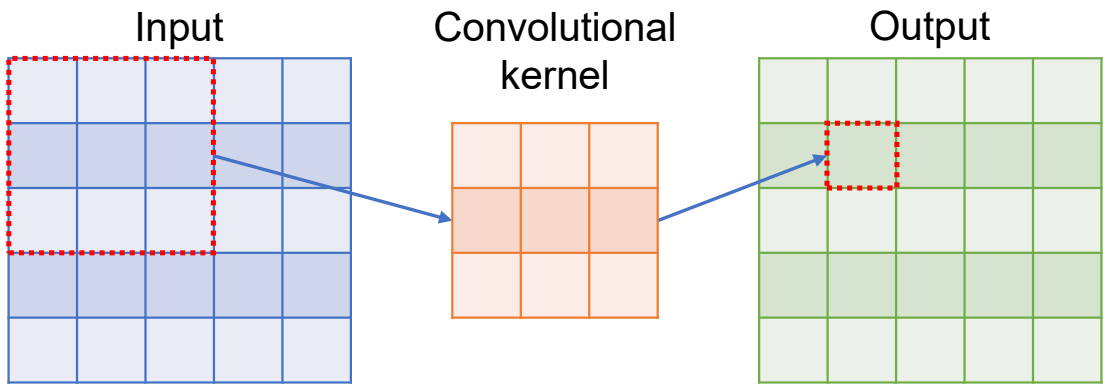


**Figure 2.2:** Visualization of how a tensor is processed by convolution. The convolution layer processes the input tensor with a sliding window, calculating the weighted sum as defined by its kernel for each position within it, thereby producing the output tensor.

window fashion. $N$ is a hyperparameter that defines the size of the kernel, which defines the size of the sliding window. A visualization of a convolutional layer is shown in Figure 2.2. Convolutional layers can, but do not have to, change the tensor's shape, which is determined by its other hyperparameters. Those hyperparameters include the *stride*, the step or distance between two sliding windows, and the *padding*, the filler values added to all sides of the spatial dimensions. A more direct approach to the manipulation of the tensor shape is realized in the form of pooling layers. Pooling layers work similarly to convolutional layers, and they are defined by their kernel size, stride, and padding, but they do not have a kernel of learnable parameters. Instead, they apply a specific function to their sliding windows, most often the *maximum*, *minimum*, or *average* function. By default, their stride is equal to the kernel size, resulting in no overlapping windows and a halving of the spatial dimensions (but no change to the number of channels). Still, convolutional layers do not universally solve all image analysis problems, as there are countless ways to connect those layers into a computer vision neural network architecture. Luckily, years of research have produced universally successful architectures, with the most notable being DenseNet201 [27], InceptionResNetV2 [28], ResNet50 [29], VGG16, VGG19 [30] and Xception [31]. Performance varies between architectures, partly due to their capacity, architecture, the domains they are applied to, the amount and type of data available, and many other known and unknown unknowns.

Data availability is a general problem for almost every researcher in the field of deep learning-based methods. The problems with data can be broad, be it for specific biases in the collected data [32], or from an overall difficulty in acquiring data. An interesting property of images is that images, even when slightly modified, still represent the same object. In other words, an image of a dog is an image of a dog, even if it is mirrored, flipped upside-down, resized, or blurred. With the introduction of this kind of noise, additional samples can be generated, effec-

tively expanding the dataset. In general, by adding noise to the data, the network should be able to pick up on some underlying general patterns and features which help it to generalize better [33]. The generation of virtual samples by introducing noise is called *image augmentation*. Image augmentation has its limits. The modifications applied to the image cannot entirely distort the image. What is a suitable modification varies from task to task, as the same modification might improve the training outcome for one task and fully destroy the necessary information for another. The choice and magnitude of the applied modifications are additional hyperparameters that must be tuned appropriately. Another approach to solving issues stemming from a limited dataset is the usage of pretrained model weights [34]. The same convolutional neural network architecture can be trained for different tasks. Laboratories and companies with huge resources train deep learning-based models for general image analysis tasks on enormous datasets of 10s of millions of images, which they make publicly available. General image analysis tasks vary but are most often image classification models with hundreds or even thousands of classes. Such trained models and their weights can be used as a starting point for the training of other tasks, in their entirety or some parts. The core idea behind this approach is that, in the early layers of the network, some generally functional features are learned, which can successfully be applied to other tasks. For tasks in the natural image domain, more of the pretrained network is usable, while for domains like medical image analysis, only lower-level features are helpful.

One major drawback of deep learning-based neural networks is their black-box nature. Convolutional neural networks consist of tens of millions of learnable parameters used to transform data non-linearly, making it impossible to deduce how they function. Moreover, while this is a problem in general, it is especially a problem in medical image analysis, as medical decisions cannot be made without logical backing. Interpretability methods are intended to bridge this gap in understanding. The most basic technique is the occlusion method. A sliding window approach is used to cover up part of the input image, which the model then processes. This process repeats until the model has processed all possible image variations generated by the sliding window. The changes in the model prediction indicate which areas are important to the model, with drops in probability indicating that the covered area contains salient information for the target class. While intuitively understandable, this approach has two substantial problems. One problem is the inability to handle interpretability cases where the model prediction is based on multiple areas of the image. The other is the high computational requirement, as all variants of the covered-up image must be generated and subsequently processed by the model. Gradient-based methods solve those issues, and GradCAM [35] has been shown to be one of the most reliable and most widely used. GradCAM constructs a Class Activation Maps by calculating the gradient between the output neuron of the target class and the output of the last convolutional layer. Those gradients are averaged per channel and used as the weights for the weighted sum of activations of the last convolutional layer to calculate the Class Activation Map. A visualization

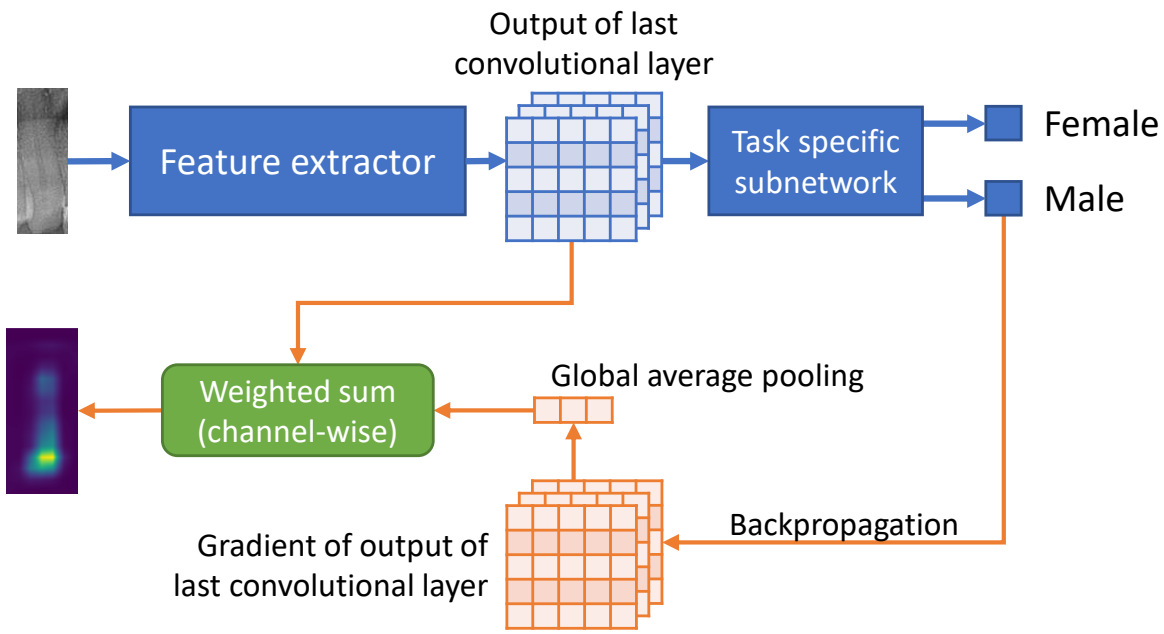


**Figure 2.3:** GradCAM uses the gradient of the output of the feature extractor to calculate the saliency map for interpretability. Global average pooling is used to calculate the gradient-based weight vector. The resulting saliency map is the weighted sum (channel-wise) of the last convolution layer's feature map and the gradient-based weight vector.

of this approach is shown in Figure 2.3.

The attention mechanism attempts to directly resolve the issue of complex relations between learned features [6]. It started in the field of Natural Language Processing (NLP) as a means to handle relations between words. For example, translation from one language to another is a task tackled by NLP. The attention mechanism would explicitly model the relation between words in the original sentence and their contribution to the predicted words in the translation. Over time the attention mechanism was adapted to other deep learning domains, including computer vision. For convolutional neural networks, attention can be realized as the Convolutional Block Attention Module (CBAM) [36]. It can be calculated from any intermediate convolutional layer's output tensor and multiplied with that output tensor to achieve adaptive feature refinement. The Convolutional Block Attention Module calculates the attention map by using both the inter-channel and the inter-spatial relationship of features. The channels of a feature map (output tensor of a convolutional layer) are considered feature detectors [37], therefore the inter-channel attention focuses on the "what" is important in the target tensor. The spatial attention, on the other hand, focuses on the "where" question, giving attention to the informative parts of the target tensor while suppressing noise from the uninformative parts.

Ultimately, all the theoretical achievements have led to the construction of deep learning-based models and training algorithms which show success in medical applications. As early as 2015, convolutional neural networks have been used to classify lung nodules of thoracic

Computed Tomography screenings [38], and for assessing myofiber orientation in high resolution, phase-contrast CT images [39]. With the development of more complex architectures and hardware capable of performing the required computations in a reasonable time, convolutional neural networks have been used for hemorrhage detection in color fundus images [40], for Alzheimer's disease diagnostics from MRI scans [41], mitosis detection in breast cancer histology images [42] and in radiographic images [43], automated pulmonary nodule detection in CT screenings [44], and in many more, as shown in survey and review studies [4, 5]. Even recently, with the terrifyingly quick spread of the SARS-CoV-2 virus and the COVID-19 illness, deep learning-based methods were used for diagnostics in radiographic images [45], and even detection of the illness by a recording of the cough of a potentially infected person [46].

## 2.2 Forensic odontology

### 2.2.1 A brief overview of the evolution of forensic odontology

Forensic odontology focuses on the analysis of dental remains with the goal of identification of a person and determination of their demographic information. This demographic information is often enough to identify a person in some scenarios. Nevertheless, even if the basic demographic information, specifically age and sex, is insufficient to identify a person, they are still vitally important and one of the first aspects examined during the forensic process. The usefulness of teeth and dental indicators for forensic research and age determination was discovered back in 1837 [47]. Likewise, early research suggests that the sex of an adult person can be determined with 100% accuracy if the entire skeleton is present [48]. However, having the entire skeleton available for forensic analysis when a forensic analysis is needed is a rare luxury. Different external factors (blunt force, fire, bacterial decomposition) can lead to irrecoverable damage to remains. Dental remains, teeth, and the jaw are different in that regard. They can withstand higher blunt force impacts, they are more resistant to fire, and less susceptible to bacterial decomposition [1, 2, 3]. Still, to perform the forensic analysis, early methods necessitated the destruction of the dental remains [49]. Radiographic imaging [50] improved over time, leading radiographic, non-destructive approaches to viability [51, 52]. More recent research has validated the initial viability, which has lead non-destructive methods to match and eventually outperform their destructive counterparts [53, 54, 55, 56, 57]. One major limitation of current age estimation and sex assessment methods is the requirement for healthy, unaltered, undamaged teeth [58].

Six methods of radiographic imaging are dominating dentistry and forensic odontology [59]. Bitewing x-ray images are high-resolution images focusing on smaller areas. Periapical radiographs target a wider area, imaging the entire tooth and part of the jaw bone. Occlusal x-ray

images target the mouth's roof or floor, allowing for better insight into tooth development and placement. Panoramic dental x-ray images are a complete projection of the mouth onto two dimensions. They show the entire skeletal and dental structure and give detailed insight into the state of the dental remains or patient. Cephalometric projection is a side-view x-ray of the entire skull, or what remains of it. It extends beyond just the jaw and teeth and gives insight into the entire state of the skull. Cone beam computer tomography is the most advanced and costly approach, constructing a 3D representation of the jaw and teeth. The used indicators, ratios of tooth structure dimensions, can easily be disturbed and destroyed, be it naturally through tooth decay, blunt external force, or therapeutic intervention (for example, dental fillings). While incorporating deep learning-based methods into medical image analysis is relatively new, some studies have explored the possibilities for dentistry and forensic odontology. Some examples include the estimation of dental parameters for orthodontic assessment [60] and attempts of full identification of a person by their panoramic dental x-ray image [61], and a recent review study [62] shows promising studies in a wide breadth of possible tasks. Those tasks include detecting and identifying diseases, from common tooth decay to cancerous lesions, and even prosthesis fabrication. Most research is aimed at dentistry, with a minority of studies tackling forensic tasks.

### 2.2.2 Age estimation

Early forensic researchers in 1837 realized and demonstrated that teeth are useful for the estimation of age [47]. Age estimation by means of the skeleton was more relevant in the beginnings [63, 64], but further research indicated that teeth might be a better and more reliable source for age estimation [65]. Early age estimation methods were focused on children. One factor contributing to this focus was practical in nature, as child labor protection laws were introduced, which brought the need for a reliable method to determine if a child was of working age. The other factor stems from the fact that dental development strongly correlates to a child's age. In fact, early [66, 67], and contemporary [68] studies have shown that tooth development is on a strictly defined schedule and that the age of a child can be determined within an error measured in months. After dental maturation, those indicators could no longer be used, rendering the age estimation methods ineffective. It was later discovered that after development, tooth calcification and decay become the main driving forces of change in the tooth [69]. Tooth development has been separated into different stages, either 8 [70] or 10 [71]. Linear models have also been constructed that link orthodontic measurements with the age of a child [72]. While samples used for age estimation need to be healthy and without illnesses or alterations, contemporary efforts are trying to establish development atlases for children with systemic diseases [73]. While estimating the age of an adult is still a challenge, the difference between juvenile and adult teeth was examined, and a method to differentiate between them was developed [74].

Three foundational studies heavily influence modern age estimation of adults in forensic odontology. The earliest study, Kvaal et al. [75], constructs a linear model for age estimation based on the measurements of the dental pulp cavity. The discovery that enabled this is that secondary dentine deposits slowly shrink the dental pulp cavity with age. A standard error in the range between 8.6 and 11.5 years is achieved depending on the exact configurations of teeth used. Using the coronal pulp cavity, Drusini et al. [76] incorporate the tooth-coronal index [77] for premolars and molars into their age estimation method and achieve a standard error between 5.88 and 6.66 years. The last of the foundational studies is Cameriere et al. [78], which constructs a linear model for age estimation based on the measurements of the single-rooted maxillary right canine. Those methods have been verified over and over again by being applied to different populations [79, 80, 81, 82, 83, 84, 85]. Interestingly, while the methods are valid and perform well, the parameters used by those models vary between populations, highlighting the need to determine population-specific parameters to achieve the most precise estimation. The importance of those foundational studies is seen all throughout forensic odontology research for age estimation, as modern methods can be classified into distinct categories depending on which study they build upon [86].

Some studies have attempted to combine deep learning-based image analysis approaches with forensic odontology age estimation. Some are based on development stages, automatically classifying child teeth into the Demirjian stages [87]. Other studies forgo the image analysis aspect and train deep learning-based models with manual orthodontic measurements [88]. As classical studies have shown, child samples' age and development stage can be accurately estimated, and deep learning-based approaches are no different. Direct automated estimation of age from panoramic dental x-ray images has been attempted with a unified model for both children and adults, albeit with a majority being child samples [89], achieving an error between 2.84 and 6.59 years. Other methods approach the age estimation problem as a classification problem, with some classifying panoramic dental x-ray images of younger individuals into discrete 1-year age groups for the ages between 5 and 24 [90], and others classifying them into either three or five discrete, hand-selected age groups [91], both achieving an accuracy of 90% or higher. These three studies, their approach, and their results are examined and compared in detail in the discussion of the thesis results in Chapter 6. Not all methods deep learning-based computer vision methods for forensic age estimation are structured around panoramic dental x-ray images. Cone beam computer tomography scanning approaches are the second-most studied, achieving currently inadequate but promising results [92, 93].

### 2.2.3 Sex assessment

As with age estimation, the first attempts of sex assessment were done by the analysis of the entire skeleton [48], which yielded a 100% when the entire skeleton was present. Sex can be

assessed as sexual dimorphism is expressed throughout the body as it develops. Studies show that this is no different for the maxilla and mandibula (the jaw bones), an integral part of the dental system [94]. In fact, a wide range of different morphometric and nonmetrical parameters are great indicators of sex. Those indicators, while plenty, too become unreliable in samples with illnesses and alterations [58]. Additionally, while sexual dimorphism develops as a person matures, external environmental factors, socioeconomic status, and nutrition can significantly influence its expression, further complicating the task of sex assessment [95]. Nevertheless, sex assessment is, alongside age estimation, one of the first steps of a forensic investigation [96]. The variety in applicable mandibular parameters can be seen in early studies, each analyzing a different set of parameters and achieving satisfactory results [97, 98, 99, 100]. The studies claim accuracies between 70% and 99%, and their sample sizes vary from 40 to over 400 samples. All studies focus on adult samples, with the age of a sample being between 19 and 86 years. By combination of different mandibular parameters, older studies acheive an accuracy of 85% [101] and 81.5% [102]. Studies based on a geometric morphometric approach of 10 mandibular parameters achieve an accuracy of 95% on their dataset of 40 individuals [103]. Sinus dimensions have been used too, achieving a relatively low accuracy of 69.2% to 69.4% [104]. The largest study [58] uses mandibular ramus flexure as a morphologic indicator of sexual dimorphism with a dataset of 419 samples and achieves a precision of 94.6% for female samples and 47.6% for male samples. Numerous combinations of metrical indicators have been analyzed across a multitude of populations in the last decade. All those newer studies achieve a similar accuracy ranging from 71% to 87% [96, 105, 106, 107, 108, 109, 110, 111, 112]. Nonmetrical approaches also successfully assess the sex of a person by examining indicators like the shape of the chin, divergence of the gonial angle, the profile of the chin, the contour of the base of the mandible, and shape of the ramus of the manbile, albeit with a less reliable accuracy between 75% and 95.2% [113]. Sex assessment is also performed when only a single tooth is available. Diagonal parameters, especially of canines, have been shown to be highly discriminative in regard to sex assessment, with a model based on those measurements achieving an accuracy of 83.3% [114], shown on a dataset of 60 samples. A newer study found that the mesiodistal widths of upper and lower canines, premolars, and molars are significantly different between females and males, devising a linear model with an accuracy of 75% on a dataset of 168 samples. Interestingly, research has shown that assessment of sex from only a single tooth cannot be considered reliable and should only be used in conjunction with other sources of information [115, 116]. The findings [115] suggest that sex assessment methods based on a single tooth cannot achieve an accuracy higher than 80%. Ultimately, the systematic review [117] shows that sex assessment methods follow the same principle by constructing a linear model from different combinations of manually measured dental parameters.

Two studies have examined the applicability of deep learning-based methods for sex as-

sessment of panoramic dental x-ray images. An approach that uses a custom neural network architecture trained on a dataset of 3400 samples achieves an accuracy of 84% [118]. Their architecture consists of two feed-forward branches of six convolutional layers, with a skip connection every two layers between those two branches. The deep learning-based study with the largest dataset of panoramic dental x-ray images achieves an accuracy of $94.6\% \pm 0.58\%$ with a sample count of 19 976 images of Chinese patients [119]. Their neural network architecture is a combination of ResNet [29] and DenseNet [27], using skip connections and a multi-scale approach. Interestingly, deep learning-based computer vision models for sex assessment from individual tooth x-ray images have not been demonstrated in current literature. As with age estimation, the cited studies will be examined in more detail and compared to the thesis results in Chapter 7.

### 2.2.4 Segmentation of dental images

Tooth segmentation is not performed in classical forensic odontology in the same sense as in computer vision. Forensic odontology experts analyze the image and visually note and differentiate between teeth, tooth alterations, and other structures visible in a panoramic dental x-ray image. This kind of analysis is formalized as segmentation in computer vision. The success of a segmentation model can be measured with accuracy and with the Sørensen–Dice coefficient and their definitions and trade-offs are examined in Section 5.4. Early research did not use deep learning-based approaches to segment images. Fuzzy clustering of both panoramic and periapical dental x-ray images [120], and color-based segmentation of dental fillings of tooth photographs [121] achieved medicore but promising results. A segmentation pipeline consisting of CLAHE [122], Otsu's method [123], and Connected Component Analysis achieved an accuracy of 84% for the segmentation of periapical images. Segmentation without neural networks is not just limited to 2D images, with an active contour approach segmenting CBCT scans with a Volume Overlap Error (VOE) of 10.65%. The newest study, which does not use neural networks, segments the root canal of a tooth instead of the tooth [124]. They achieve an accuracy between 82.0% and 88.2% using an improved level set algorithm constrained by a new regularization function developed for this study.

As with most of computer vision, neural networks significantly outperform any classical image processing methods, and segmentation is no different. Most neural network-based approaches, except for the earliest, achieve a Sørensen–Dice coefficient of over 90% on all varieties of dental imaging. A deep learning model that is a combination of ResNet101 [29] and the Feature Pyramid Network (FPN) [125] was trained on panoramic dental x-ray images, and achieved a Sørensen–Dice coefficient of 88% [126]. TSASNet, a custom-designed two-stage attention-based neural network model for tooth segmentation, is trained with 1500 panoramic dental x-ray images and achieves a Sørensen–Dice coefficient of 92.72% [127]. Another tooth

segmentation network, trained on 100 CBCT scans, uses a symmetric, fully convolutional residual network with a Dense Conditional Random Field. This approach achieves a Sørensen–Dice coefficient of 91.66%. UDS-Net trained on 120 CBCT scans achieves a Sørensen–Dice coefficient of 91.8% [128]. This networks is a variant of UNet [129], a popular segmentation architecture explained in depth in Section 5.4. In contrast to the previously shown studies, a thorough evaluation of Mask R-CNN [130], PANet [131], HTC [132], and ResNeSt [133] yielded a PANet model trained on 1500 panoramic dental x-ray images which achieves a Sørensen–Dice coefficient of 91.65%. Another study that examined multiple models, this time variants of UNet [134] found a configuration that achieves a Sørensen–Dice coefficient of 92.89%. A combination of Mask R-CNN [130] and FPN [125] yielded a Sørensen–Dice coefficient of 94% for a dataset of 433 CBCT scans. And lastly for tooth segmentation, MSLPNet is a multi-scale custom architecture built for tooth segmentation. It is trained on a dataset of 1500 panoramic dental x-ray images and achieves a Sørensen–Dice coefficient of 92.72%. Overall, while many neural network architectures have been evaluated over a big batch of studies, the results are close together without much variance. This holds true for other studies which a similar approach, too [135]. One study differs from the rest as it does not segment teeth or tooth structures like the root canal. It instead focuses on the segmentation of carious lesion - tooth decay. CariesNet [136] is trained on 3127 regions containing tooth decay cropped from 1159 panoramic dental x-ray images, and the specially designed attention-based model achieves a Sørensen–Dice coefficient of 93.64%.

### 2.2.5 Tooth type determination and detection

Tooth detection is one more task that is done by forensic experts but which requires no special studies from the human-expert perspective. Humans are extremely good at distinguishing objects, and distinguishing teeth from each other and non-teeth tissue in panoramic dental x-ray images is a trivial task for a person. Tooth type determination is a topic taught early in dentistry education through dental morphology understanding [137, 138]. Age estimation and sex assessment methods make use of different teeth and distinguish them by their type. Some methods develop a set of linear models depending on the tooth type, and applying those models to the measurements of the wrong tooth type can lead to wrong assessments. Other methods develop models based on a set of teeth, and their measurements are not interchangeable within those models. To complicate matters further, different studies use altogether different classification systems for tooth types. The internationally accepted standard is the FDI dual notation system [139], which assigns a two-digit label to a tooth depending on its location within the dental system. However, this 32-class system is not universally used in forensic studies. A deep dive into the intricacies of tooth type classification systems is given in Section 5.3.4. Given all this, three overall approaches dominate in the literature - tooth type determination from individual tooth x-

ray images, tooth detection without tooth type determination, and simultaneous tooth detection with tooth type determination. One of the older studies tackled simultaneous tooth detection and type determination by using a sliding windows approach, and a deep learning-based model [140]. Each image generated by the sliding window approach is analyzed by the AlexNet model trained to distinguish teeth using a 3-class tooth type classification approach. Another study in the same year [141] forgoes detection and neural networks entirely and uses projection profile analysis on individual tooth x-ray images to determine the tooth type using a 4-class approach. Studies for simultaneous tooth detection with tooth type determination also tend to use heuristics to correct their raw neural network results in post-processing steps [142], where a neural network determines the type of all teeth on a periapical x-ray image using a 32-class system. The heuristic assumes that the Faster R-CNN [143] neural network is mostly correct and uses the classifications of a tooth's neighborhood to determine if it is correct and, if not, corrects it to fit into the chain. This approach significantly improves the results, with the major drawback being that teeth still need to be embedded into the jaw. Another Faster R-CNN neural network avoided tooth type determination and detected teeth only, but on panoramic dental x-ray images of children [144], achieving promising results but demonstrating the complexity of automated tooth type detection in children, where the primary and permanent teeth are simultaneously visible and often overlap in the panoramic imaging approach.

# Chapter 3

# Data: Acquisition, labeling, and preparation of dental x-ray images

## 3.1 Acquisition and labeling of the data

The research was conducted using panoramic dental x-ray images as the primary data source. Panoramic dental x-ray images, also called orthopantomographic images or "OPGs," show the entire mouth, upper and lower jaw, all teeth, as well as the surrounding structures. Those images are created with panoramic dental x-ray machines - a medical device that rotates around a person's head while the head's position is fixed in place. The procedure is not invasive, entirely painless, and takes less than five minutes from start to finish. The imaging device uses a minimal amount of ionizing radiation, which is safe and leads to no short- or long-term consequences. A sample of a panoramic dental x-ray image can be seen in Figure 3.1. As can be seen, the skeletal and dental structure is clearly visible, as are any alterations to the teeth like fillings or decay. This allows for a detailed analysis of the dental tissue for medical or forensic purposes.

The samples in the dataset belong to the collection of the Department of Dental Anthropology School of Dental Medicine University of Zagreb, and they have been created using multiple panoramic dental x-ray machines over a period of 5 years. The use of this collection for research purposes has been approved by the ethics committee School of Dental Medicine University of Zagreb. Likewise, the ethics committee of the Faculty of Electrical Engineering University of Zagreb has approved this research.

All personal information has been anonymized. Each image was originally recorded with information about the sex of the imaged person, their age on the day of imaging, and an "identity hash." The identity hash is a one-way coding used to tell if two images are from the same person, but it cannot be used to reconstruct a person's personal information. The age is stored as a floating point number with two decimal places, and the sex is stored as a binary variable.

Experts of the Department of Dental Anthropology School of Dental Medicine University

**Figure 3.1:** Sample of a panoramic dental x-ray image. In those x-ray images, the entire mouth, including the surrounding tissue and all teeth, are visible.

of Zagreb have annotated a subset of the dataset with the tooth status information of individual teeth. The annotation contains precise information about the location of each individual tooth, given in the form of their bounding box. This allowed for the extraction of individual tooth images. In addition to the position, tooth interventions, as well as the tooth type, were annotated. Tooth type is annotated according to the FDI dual notation system (ISO 3950) [139], which gives every tooth a designation based on its position in the oral cavity. A tooth status annotation also contains information about any alterations, namely tooth decay, fillings, root canal fillings, crowns, bridges, tooth germs, leftover roots, dental implants, missing teeth, and crowns. If a tooth is missing, a bounding box is created in the place where the tooth would have been had it not been removed. These annotations were created manually using the "Computer Vision Annotation Tool (CVAT)" tool.

As mentioned, those interventions to the tooth are visible on the panoramic dental x-ray image. Another subset of the dataset is annotated with detailed per-pixel segmentation maps by dentistry experts of the Department of Dental Anthropology School of Dental Medicine University of Zagreb. Each tooth is annotated individually. The annotation consists of as many segmentation maps as there are alterations on the tooth and a separate segmentation map for the tooth itself. The per-pixel segmentation maps are created using the "GNU Image Manipulation Program (GIMP)" software.

## 3.2 Preparing datasets for different forensic odontology tasks

Panoramic dental x-ray machines record raw readings in proprietary formats, which are then converted to 8-bit JPEG images. Images vary in size as the samples are taken with different

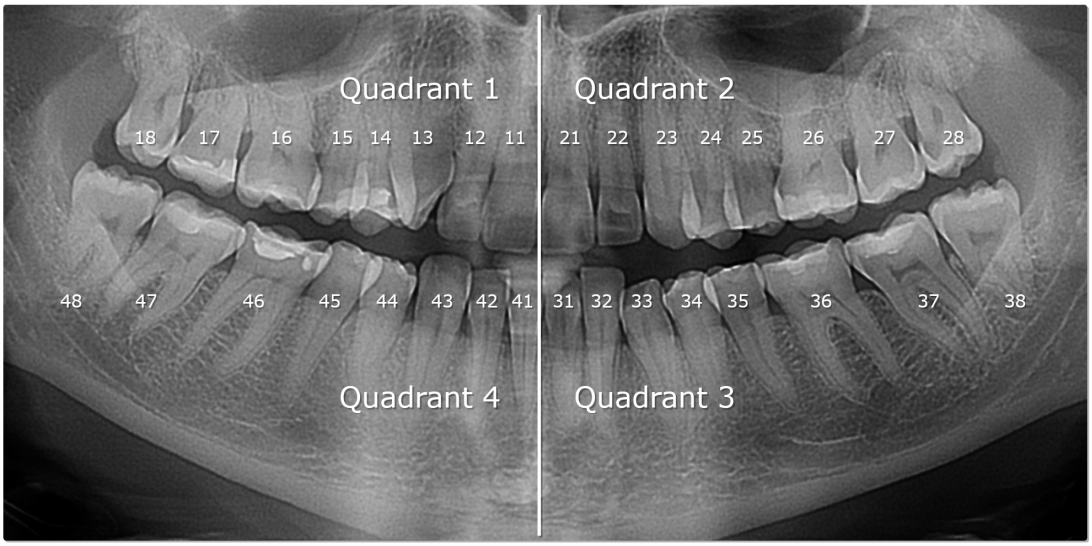


**Figure 3.2:** The FDI dual notation system. As defined in the ISO 3950 standard [139], each tooth is given a label based on its position in the oral cavity. The first number defines the in which quadrant the tooth resides, and the second number defines the position within this quadrant.

dental x-ray machines. Specifically, the converted images have a resolution of 1127 px to 3260 px in width and 553 px to 1536 px in height. To be usable with convolutional neural networks, the images are resized on a case-by-case basis. The specifics for each approach will be described in Chapter 5.

While multiple forensic tasks have been tackled, the gathered data can be formalized as just three different datasets. The first dataset consists of panoramic dental x-ray images. These images are accompanied by the age and sex of the imaged person, as well as the precise tooth positions on the image. This dataset is used to develop full panoramic dental x-ray image models for age estimation, sex assessment, tooth detection, and tooth detection with tooth type determination. The second dataset comprises of individual tooth x-ray images. This set is generated by clipping out individual teeth by their precise bounding-box position. Every image is annotated with the age and sex of the imaged person and the tooth type. Additionally, a subset of those teeth are annotated with tooth status information. The third and final dataset is the dataset of pixel-level segmentation masks of teeth and any alterations they might contain. Every tooth is annotated with a separate segmentation mask, as are any alterations of those teeth. With such a high degree of separation, variants of the dataset are generated as required - be it segmentation masks of individual teeth and their alterations in any combination or the generation of full panoramic dental x-ray segmentation masks for teeth or any alternation. The following sections will give a detailed overview of each of the three mentioned datasets.

For proper training, validation, and testing, each dataset is separated into three subsets. The training subset consists of 80% of all samples. As the dataset can contain multiple images of the same person, to prevent data leakage and any unaccounted biases in the results, all images that share the same identity hash are placed into the train set. In other words, if a person's images

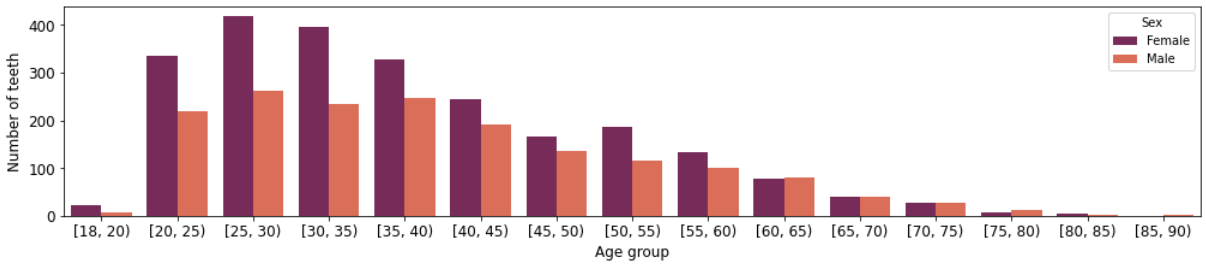


**Figure 3.3:** The distribution of samples in the dataset per age and sex. As can be seen, a slight bias towards female samples is present, as well as a trend of decreased sample count with increased age.

appear multiple times in the dataset, they can only be in the training subset. The validation and test subset share the remaining 20% of data. For most approaches, an even 50:50 split between validation and test sets is used, except for the case of individual tooth x-ray images. As only some images have annotated status information, all those are placed into the test set, allowing for an in-depth analysis of the impact of tooth alterations on the performance of the models. The validation set is used during research to analyze the generalization performance of trained models, and the results of the validation set inform research decisions. On the other hand, the test set is a hold-out set that is evaluated only once, once the approach has been fully developed and the model has been fully trained. Those results are used to measure the approach's success, and those results are ultimately presented in research papers and this thesis. All approaches use the same dataset split - an image used for testing will be used for testing in all developed approaches, making the results directly and easily comparable.

## 3.3 Properties of the datasets

### 3.3.1 Panoramic dental x-ray images

The dataset of panoramic dental x-ray images consists of 4035 panoramic dental x-ray images, the age, sex, and identity hash of the imaged person, and the location of each tooth in the form of a bounding box. A sample of a panoramic dental x-ray image can be seen in Figure 3.1. A person's age is represented as a floating-point number, and it is calculated as the difference between the day of birth and the day of imaging expressed in years. Ages range from 19 to 90 years. Sex is represented as a binary variable, with the value being either female or male, with a ratio of female to male being 58.7%:41.3%. This totals 2346 female and 1647 male samples belonging to 3994 individuals. For a subset of 2899 images, tooth position is given as a bounding box around every individual tooth, as well as the tooth type in FDI dual notation. The ratio of female to male samples in this subset is the same as for the entire dataset.

An overview of the distribution of samples across age groups can be seen in Figure 3.3. A detailed count is given in Table 3.1. A bias towards younger samples, as well as towards

**Table 3.1:** Sample count per age group and sex of panoramic dental x-ray images in the dataset, with the age groups spanning five years.

| Age group | Female sample count | Male sample count |
|---|---|---|
| **[18, 20)** | 21 | 8 |
| **[20, 25)** | 337 | 221 |
| **[25, 30)** | 419 | 264 |
| **[30, 35)** | 399 | 234 |
| **[35, 40)** | 325 | 247 |
| **[40, 45)** | 246 | 188 |
| **[45, 50)** | 167 | 137 |
| **[50, 55)** | 185 | 118 |
| **[55, 60)** | 132 | 98 |
| **[60, 65)** | 75 | 79 |
| **[65, 70)** | 30 | 35 |
| **[70, 75)** | 24 | 24 |
| **[75, 80)** | 6 | 11 |
| **[80, 85)** | 2 | 3 |
| **[85, 90)** | 0 | 1 |
| **Subtotal** | **2368** | **1667** |
| **Total** | **4035** | |

female samples, can be observed. While the age is biased towards younger samples, no sample is younger than 19, and the average age is 38.17 years. This dataset, therefore, comprises adult samples - images of dental systems that have finished their development. Samples are sparse for the age range of 75+, which is taken into account during result analysis.

Samples in the dataset are not filtered by any tooth quality criteria and represent real-world data. In other words, the samples can contain various pathologies, like the loss of mandibular molars, anomalous molars and teeth, and others. Most studies in the field of forensic odontology are not conducted with such data [58]. Alterations and pathologies are deformations of the tooth and changes to the structure of the dental tissue. This leads to either unmeasurable dental parameters due to the lack of measurable tissue or unreliable measurements, which are thus unsuitable for analysis and study.

**Figure 3.4:** Samples of individual teeth from the dataset. Every x-ray image of a tooth originates from a panoramic dental x-ray image, using expert's bounding box annotations.

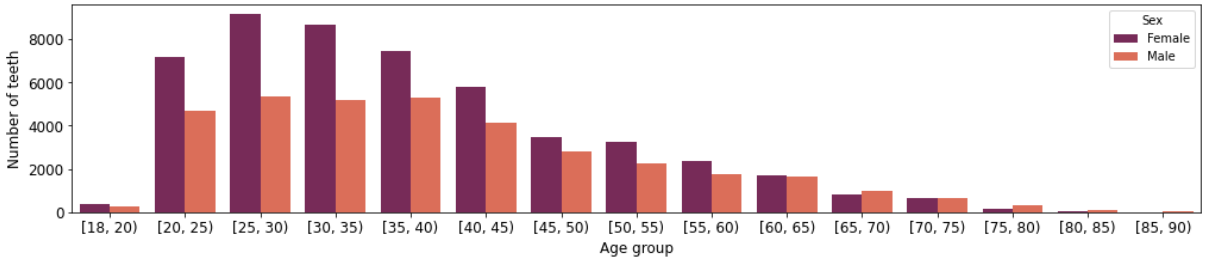


**Figure 3.5:** The distribution of samples in the dataset of individual tooth x-ray images, per age and sex. The slight bias towards female samples is preserved. The trend of decreased samples with age is exacerbated in this dataset, as the loss of teeth is more common with higher age.

### 3.3.2 Individual dental x-ray images

The dataset of individual dental x-ray images consists of 86495 individual tooth images, Those images are obtained by clipping out individual teeth from the 2899 panoramic dental x-ray images, which have bounding box annotations. The ratio of female to male samples in this dataset is 59.03% to 40.97%, and the ages range from 19 to 86 years. As with the panoramic dental x-ray images, this dataset is biased toward younger samples too, but the average age in the dataset is 38.41 years. Examples of individual tooth images and their possible alterations can be seen in Figure 3.4. Each image of a tooth is accompanied by information about the type of the tooth and the imaged person's age and sex. The distribution of samples across age and sex can be seen in Figure 3.5.

A subset of 7630 images is additionally annotated with information about tooth status. Tooth status annotation contains information about any alterations to the tooth. Specifically, tooth status contains information about tooth decay, fillings, root canal fillings, crowns, bridges, tooth germs, leftover roots, dental implants, missing teeth, and crowns. Not all alterations appear with the same consistency. The most numerous alterations are missing teeth, dental fillings, root canal fillings, and tooth decay. The other alterations do not appear as often, and due to the small sample size of the annotated teeth with status information, tooth germs, leftover roots, dental implants, bridges, and crowns are grouped into one category, which will be referred to as "Other" in this thesis. Tooth alterations are not mutually exclusive. One tooth can have no alterations, one alteration, or multiple alterations. In this dataset, 66.37% have no alterations, 27.93% of all teeth have one alteration, 5.10% have two alterations, 0.59% have three, and

0.01% have four alterations. A detailed overview of sample count per sex, alteration, and age group can be seen in Table 3.2.

**Table 3.2:** Sample count for the dataset of individual tooth x-ray images per age and sex. The distribution of alterations is also shown, but only for the subset of 7630 teeth with those annotations available.

| 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** |

Teeth sizes vary by type, with incisors and canines being narrow and long and premolars and molars being almost similar in width and height. Still, no tooth exceeds a size of 528 px in either the width or height dimension. To avoid including neighboring teeth, individual teeth are clipped out as defined by the bounding-box expert annotations. However, as images need to be equally sized for efficient training on the GPU, the resulting images are padded to a size of 528 x 528 px. Preliminary experiments show that the choice of padding does not impact the results of the models. Therefore zero-padding is chosen as it is the simplest solution.

### 3.3.3 Segmentation masks of tooth status

The dataset of segmentation masks consists of 29698 pixel-precise annotation maps. These annotations of individual teeth and their alterations originate from 813 panoramic dental x-ray images. The average age of a sample is 38.62 years, and the ratio of female to male samples is 56.86% to 43.14%. As with status annotations for individual teeth x-ray images, this dataset covers tooth decay, fillings, root canal fillings, crowns, bridges, tooth germs, leftover roots,

dental implants, missing teeth, crowns, and the teeth themselves. However, as was the case in the dataset of individual teeth x-ray images, some alterations are too rare, and the dataset currently does not have enough samples for those alterations. For that purpose, this thesis focuses on the segmentation of teeth, crowns, bridges, root canal fillings, dental fillings, and tooth decay. The distribution of alterations across age groups can be seen in Figure 3.6, and a detailed overview can be seen in Table 3.3.

**Table 3.3:** Sample count for the dataset of panoramic dental x-ray images with annotated segmentation maps for teeth and dental alterations. Alterations are not equally represented, with fillings being the most common and crowns being the least common.

| Age group | Teeth | Crowns | Bridges | Root canal fillings | Fillings | Teeth with decay |
|---|---|---|---|---|---|---|
| [18, 20) | 180 | 0 | 0 | 1 | 24 | 5 |
| [20, 25) | 3345 | 5 | 50 | 109 | 546 | 22 |
| [25, 30) | 4381 | 2 | 4 | 134 | 759 | 51 |
| [30, 35) | 3549 | 16 | 13 | 147 | 822 | 41 |
| [35, 40) | 2467 | 13 | 26 | 143 | 572 | 32 |
| [40, 45) | 2218 | 21 | 33 | 134 | 610 | 36 |
| [45, 50) | 1855 | 16 | 74 | 109 | 502 | 37 |
| [50, 55) | 1789 | 23 | 83 | 124 | 486 | 43 |
| [55, 60) | 1701 | 27 | 152 | 133 | 400 | 65 |
| [60, 65) | 860 | 8 | 62 | 48 | 197 | 26 |
| [65, 70) | 149 | 3 | 1 | 5 | 40 | 7 |
| [70, 75) | 104 | 3 | 1 | 3 | 19 | 0 |
| **Total** | **22598** | **137** | **499** | **1090** | **4977** | **365** |

The task of dental segmentation can be done on two levels - on the level of the entire panoramic dental x-ray image and the level of an individual tooth. An example of an entire panoramic dental x-ray image segmentation map can be seen in Figure 3.8, where the segmentation maps for teeth and dental fillings are shown. Individual teeth x-ray images are, again, clipped out of the panoramic dental x-ray image. The bounding box is calculated from the segmentation of the tooth by determining the coordinates of the left-most and right-most points. Figure 3.7 shows examples of each alteration and their segmentation maps. As for the sizes of those images, the same properties hold true as for the dataset of individual teeth. Therefore no image in this dataset exceeds 528 px in any dimension.

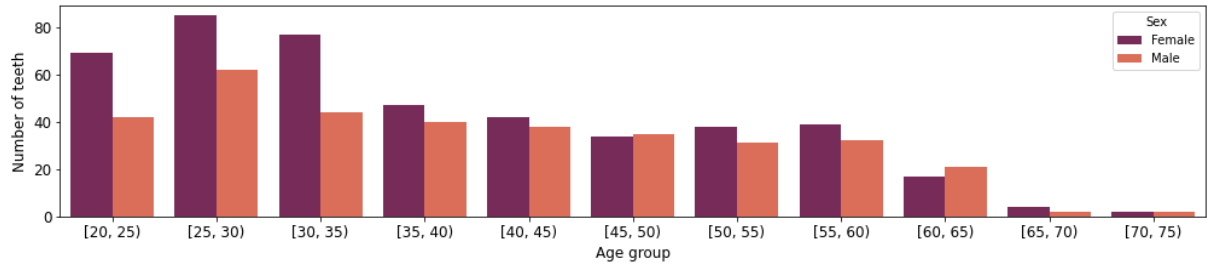


**Figure 3.6:** The distribution of samples in the dataset of panoramic dental x-ray images with annotated segmentation maps for teeth and dental alterations, per age and sex. Both the slight bias towards females and the trend of decreasing samples with age are present in this dataset.

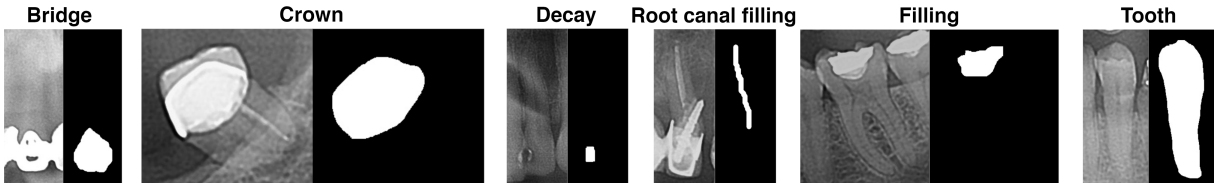


**Figure 3.7:** Examples of individual tooth x-ray images with dental alterations and segmentation maps. Some alterations, like crowns, cover more of the tooth, while alterations like tooth decay cover only a small area.

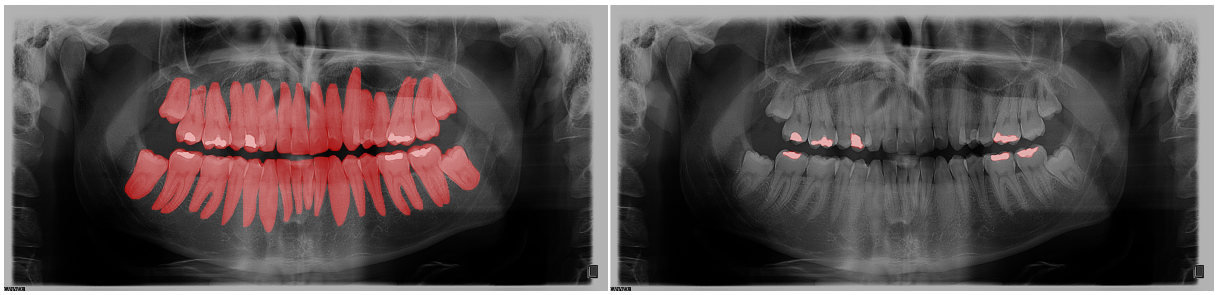

**Figure 3.8:** Example of a panoramic dental x-ray image and its segmentation maps for teeth and dental fillings. The left image shows the segmentation maps for all teeth, and the right image shows the segmentation map for dental fillings. Every annotation is stored separately, allowing for the generation of segmentation masks of the entire panoramic dental x-ray image and segmentation masks for individual teeth for any tooth and its alterations, as shown in Figure 3.7.

# Chapter 4

# Model discovery and design for the analysis of dental x-ray images

## 4.1 State-of-the-art feature extractors with added attention

### 4.1.1 Motivation and building blocks

According to the Universal approximation theorem, neural networks are universal approximators. However, figuring out the optimal parameters of a neural network is challenging. The neural network architecture, the specific connections, and the types of operations are used to introduce an inductive bias that allows for the successful optimization of its parameters.

While many ad-hoc architectures are used in literature and demonstrate good performance in specific cases, they usually generalize poorly. However, over the years, certain neural network architectures have been shown to work well across a wide variety of tasks. Those state-of-the-art architectures offer a good framework. Using their pretrained weights and transfer learning, they can easily be adapted and specialized for many image analysis tasks. They are usually pretrained on ImageNet, a generalist real-world image dataset. While it might seem counter-intuitive that a model trained to differentiate airplanes, boats, cats, dogs, people, and many other objects could be successful in medical applications, a totally different modality of images, it has empirically been shown to work well. When such a network is fine-tuned, its lower layers, those closer to the input image, do not change much. It has been shown that those layers learn general image analysis features, some fundamental "image decomposition" that works well across vision-based tasks. This generalization property is exactly what the first model in this thesis uses. The convolutional layers of state-of-the-art networks are copied and used as the feature extractor. The feature extractor processes an image to some representation that can later be used to extract the desired information from the image, like the person's age or sex. So instead of using a few ad-hoc convolutional layers, this model is built around the feature

extractor.

Recently, the attention mechanism has gained traction in machine learning and image analysis literature [6]. While initially originating in the field of natural language processing, variants of this approach have brought improvements across the entirety of neural network research. Attention allows a neural network to ignore or highlight specific parts of the processed data. However, neural networks could already do that - that is the purpose of "weights," the parameters of a neural network. However, this data weighting is defined and unchanging once training has finished. The attention mechanism allows for the weighting to be a learned function in itself. Thus, the weights can "change" depending on the input, allowing the network greater flexibility in its processing. The model in this thesis uses the Spatial Attention Module of Convolutional Block Attention to implement attention [36]. This approach calculates a feature map that is used to "scale" the output of the feature extractor. The sigmoid activation function is used, so a feature can be either fully suppressed ($\sigma(x) = 0$), fully let trough ($\sigma(x) = 1$), or scaled to any in-between value. The sigmoid function is defined as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{4.1}$$

A detailed diagram of the attention mechanism, and the entire model, can be seen in Figure 4.1.

### 4.1.2 Architecture overview

This architecture consists of four major parts. The central part is the feature extractor based on state-of-the-art architectures. While any architecture can be used for feature extraction, this thesis focuses on six architectures that have proven themselves in literature. Those architectures are: DenseNet201 ([27]), InceptionResNetV2 ([28]), ResNet50 ([29]), VGG16, VGG19 ([145]) and Xception ([31]).

The output of those feature extractors is a feature map with a very high channel count. This channel count is high as those models are made to differentiate between a vast number of classes. However, for forensic odontology tasks, the number of possible classes never exceeds 100. Therefore, such a high channel count can lead to overfitting, as the network has the capacity to memorize the training samples outright. A 1x1 convolutional layer is used to adjust the number of channels in the final feature map to prevent overfitting and memorization of the samples. The resulting feature map has the same width and height and only differs in the number of channels. This thesis explores models with a feature map depth between 5 and 1000.

The next step is the attention mechanism. As already mentioned, the Spatial Attention Module of Convolutional Block Attention is used [36]. This module first applies 2x2 maximum pooling and 2x2 average pooling and constructs a tensor by concatenating those two. A 1x1 convolution is applied to that tensor, and the sigmoid function is used as the activation function,

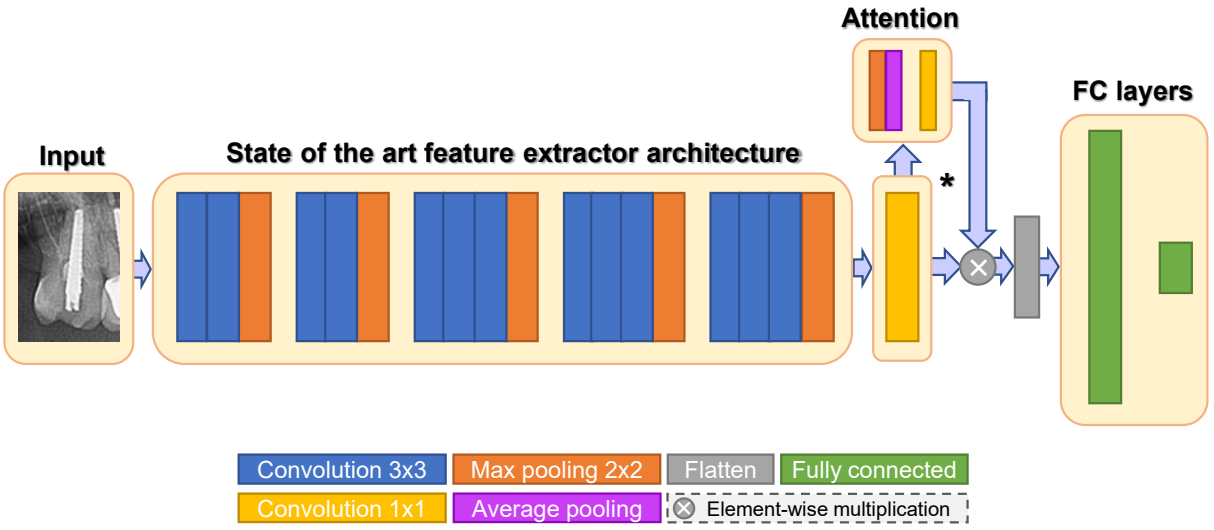


**Figure 4.1:** The model architecture is based on state-of-the-art feature extractors and the attention mechanism. While VGG16 is shown as the feature extractor, any feature extractor can be used.

which results in a feature map with one channel. This map is then multiplied element-wise with the output of the 1x1 feature map depth rescaling layer. Once the image features are fully calculated, the resulting tensor is flattened into a 1D vector. Two fully-connected layers process this vector, the latter of which is the size of the forensic task target that is estimated (for example, size 1 for age regression, size 2 for sex estimation, and size 16 for the 16-type tooth type problem).

The activation function used across all layers except the attention module is ReLU. Each layer is regularized by the usage of batch normalization and dropout. Batch normalization is used before the activation function, and dropout is used after it.

This architecture has four hyperparameters. The first and most important hyperparameter is the choice of the feature extraction architecture. This defines the overall capacity of the model. The second hyperparameter is the desired depth of the feature extraction feature map. State-of-the-art feature extractors have a very high channel count in the final feature map, as they have to handle thousands of different classes. This hyperparameter allows for the configuration of the bottleneck in such a way as to achieve the best possible performance. The attention mechanism follows this. To determine if the attention module helps and to measure the performance impact it has on the performance, the attention mechanism in itself is optional. In other words, the hyperparameter answers the question if attention is used. The last hyperparameter is the size of the first fully-connected layer. This allows for the scaling of the capacity of the fully-connected subnetwork. In this thesis, the explored range for this hyperparameter is between 1 and 2048. Models following this architecture have between 15,000,000 to 150,000,000 learnable parameters, with the most successful models for the forensic odontology tasks of age estimation, sex assessment, and tooth type determination having between 18,000,000 and 30,000,000 learnable

**Table 4.1:** The hyperparameters and their value range for the model based on a state-of-the-art feature extractor and attention. Every adjustment can strongly influence the model performance, but their optimal values change from task to task.

| Hyperparameter | Search space |
|---|---|
| Pretrained feature extractor | DenseNet201 ([27]),<br>InceptionResNetV2 ([28]),<br>ResNet50 ([29]),<br>VGG16 ([145]),<br>VGG19 ([145]),<br>Xception ([31]) |
| Number of channels in the final feature map | Between 5 to 1000 |
| Presence of attention mechanism | Present or not present |
| Size of intermediate fully connected layer | Between 1 and 2048 |

parameters. An overview of hyperparameters and their ranges can be seen in Table 4.1, and a visualization of the architecture can be seen in Figure 4.1.

## 4.2 A minimized neural network for forensic odontology

### 4.2.1 Motivation and building blocks

General purpose state-of-the-art architectures tend to work well enough in most cases, but they also tend to have very high resource requirements. These networks also rarely explore and exploit the possible interactions and, therefore, valuable data that can be processed with a multi-scale approach. Modern deep learning and computer vision has recently shifted from development and the scientific exploration of interesting algorithmic approaches to the engineering challenge of scaling compute. While this is achieving marginal improvements on benchmarks, the models themselves are becoming less and less usable by the average end-user due to the massive compute requirements. Likewise, the average computer vision laboratory is hard-pressed to use or even verify the newest work due to the required infrastructure and massive, often web-scraped, and not publicly shared datasets. Neural networks with a minimized number of free parameters tend to generalize better, as has been proven even in early neural network research [15].

To that end, a small and efficient model is constructed that uses multi-scale analysis to achieve performance parity with state-of-the-art based architectures for forensic tasks on dental

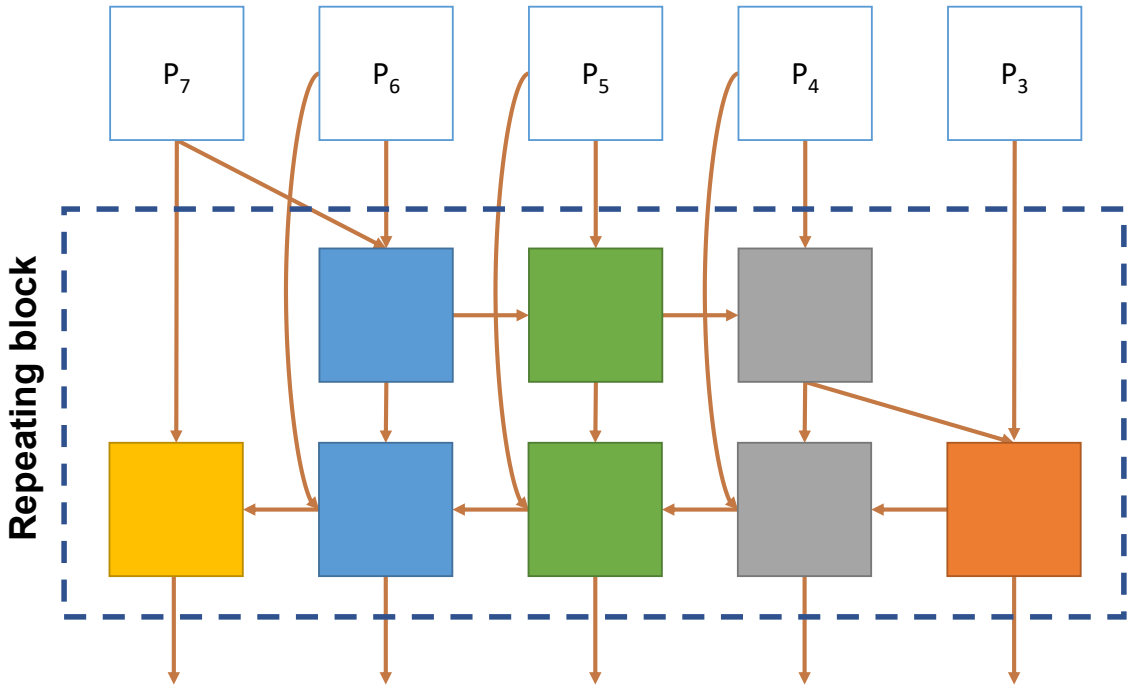


**Figure 4.2:** Diagram of the Weighted Bi-directional Feature Pyramid Network. The colors represent different feature map scales.

x-ray images while having a much smaller carbon footprint. This model consists of well known elements, such as convolutional layers and skip connections [29], but also a new element called "Weighted Bi-directional Feature Pyramid Network" (BiFPN) [146]. A diagram of this module can be seen in Figure 4.2. This module takes as input five convolutional tensors of descending scales ($P_7$ to $P_3$), combines them as shown, and outputs five tensors of descending scales. This module is repeatable, too, allowing for additional flexibility.

Instead of classical convolutional layers, the BiFPN module uses depthwise separable convolutions to improve efficiency further. This concept of depthwise separable convolutions was introduced with the Xception architecture [31]. Depthwise separable convolutions slightly increase the computation time in exchange for a decrease in the number of parameters. To get an intuitive understanding of kernel separation, let us separate a well-known image processing kernel, the Sobel operator [147].

$$\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \times \begin{bmatrix} -1 & 0 & 1 \end{bmatrix} \tag{4.2}$$

As can be seen, the 3x3 matrix can be separated into two matrices, one 3x1 and one 1x3. No information was lost, yet the representation is smaller, which for deep learning models

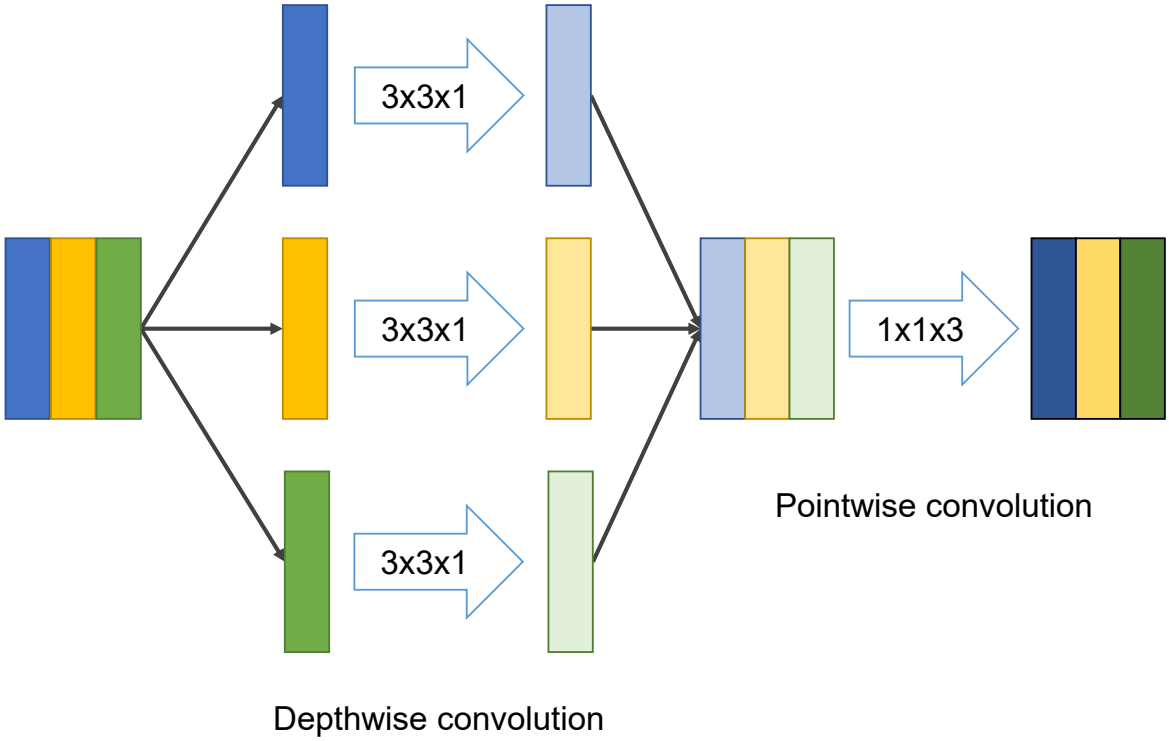


**Figure 4.3:** Overview of a 3x3 depthwise separated convolution on a 3-channel tensor. The convolution is split into two operations that apply two smaller kernels, which results in a net decrease in parameters. The trade-off is a slight increase in computation time.

translates to a smaller number of learnable parameters that express the same operation as a "normal" convolution. While this might intuitively seem like a, pardon the pun, more convoluted and thus more challenging to optimize approach, in practice, depthwise separable convolutions perform the same or better than their "normal" counterparts. Depthwise separable convolutions operate on the same principle, just in more dimensions. An overview of a depthwise separated convolution applied to a 3-channel tensor can be seen in Figure 4.3.

In deep learning literature, 3x3 convolutions have become the norm. However, research has shown that, for a fixed amount of parameters, deeper neural networks achieve better generalization [148]. Therefore, replacing the 3x3 convolution with multiple convolutions that produce a tensor of the same shape as a 3x3 convolution can be used as a drop-in replacement. For that purpose, a 3x3 convolution was replaced by two 2x2 convolutions. The first convolution has a kernel size of 2, a stride of 1, padding of 0, and no bias, whereas the second convolution has the same parameters except padding, which is set to 1. This achieves a reduction of 11% in parameter count and floating-point operations while achieving marginally better results in the preliminary experiments.

Finally, the matter of activation function needs to be defined. ReLU has proven itself to be a good default choice, but this architecture struggles with ReLU due to its gradient for $x < 0$. This

property poses an issue for the training of the BiFPN module due to the gradients converging to zero, halting the optimization process. The original paper in which the BiFPN module is introduced uses the Sigmoid Linear Unit (SiLU) function. This solved the training issues and is thus used as the activation function across the entire architecture. The SiLU function was introduced as a drop-in replacement for ReLU. SiLU showed promise in its initially study [149], and further experimentation showed a significant improvement in some applications [150, 151].

### 4.2.2 Architecture overview

This architecture is built around the BiFPN module and consists of four parts, and a general overview can be seen in Figure 4.4. The first part is a series of convolutional layers that act as the feature extractor prior to BiFPN. It consists of 8 convolutional blocks, where each block can consist of one or multiple 3x3 (or equivalent) convolutions. Specifically for this model, the first three blocks consist of four convolutional layers, which mimic two 3x3 convolutional layers as described in the previous section. Each convolution is followed by batch normalization, the activation function, and dropout. Every convolutional block ends with a 2x2 maximum pooling layer. The last three convolutional blocks consist of two 2x2 convolutional layers.

The output of the last five convolutional blocks is then passed to the BiFPN module, which produces five output tensors of the same shape. Due to the network depth, vanishing gradients and a very spiky loss landscape significantly hinder model training. To address this issue, skip connections between the first three convolutional blocks of the feature extractor and the largest three outputs of the BiFPN module are added. This allows gradients to, as the name suggests, skip some of the architecture, thereby smoothing the loss surface [152] and preventing the vanishing gradient problem [29]. The skip connection is realized as a concatenation between the two tensors, which are then processed by a 1x1 convolution.

As continuing to work with all five branches independently would be compute and memory intensive, the largest two and the smallest three branches are merged. Each branch is scaled down in width and height to the size of its smallest member. The resulting concatenated tensor is then processed by a 1x1 convolution that can be interpreted as a "feature compressor," as it reduces the size of the resulting feature maps for each branch, akin to the feature map depth rescaling layer used in the other, state-of-the-art based architecture described in this thesis. The layers are separate for each branch but produce tensors of the same depth. Finally, those two branches are concatenated and flattened, which makes them useable by the fully-connected subnetwork. Like in the previously described architecture, the first fully-connected layer is of variable size, and the forensic task target defines the second layer size.

At first sight, this architecture has far too many hyperparameters, as the depth of the feature map of every convolutional layer has to be defined. However, in current state-of-the-art architectures in literature, the feature map depths follow one of three strategies. One strategy is to

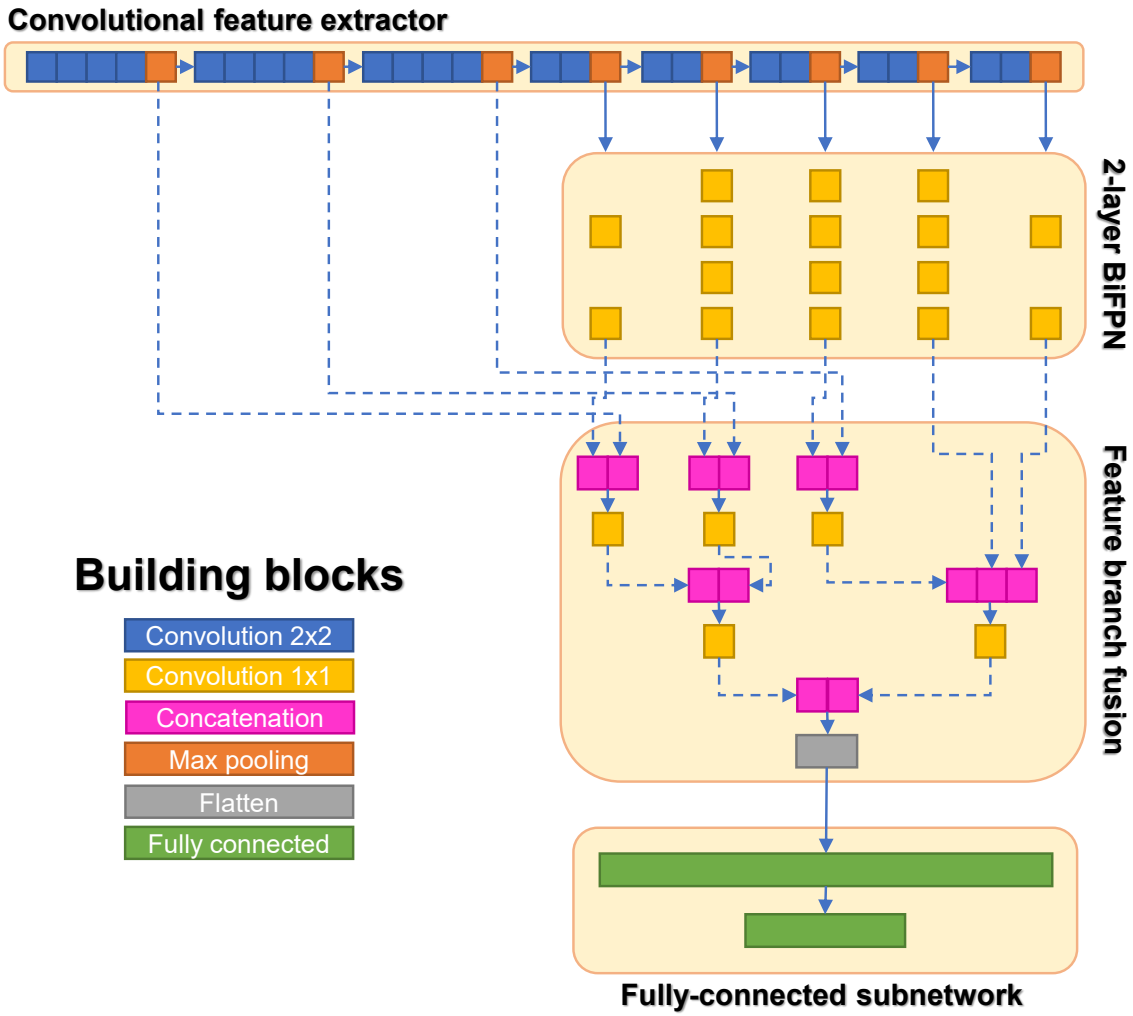


**Figure 4.4:** The architecture of the BiFPN-based model. The connections in the BiFPN module are left out for the sake of readability, but the overview of the connections can be seen in Figure 4.2.

start with just a few channels, and scale them up exponentially, like VGG16, VGG19 [145], or ResNet50 [29]. A similar approach, seen in the Inception family of models [28, 31], is to scale the depth exponentially but to define a value past which no scaling occurs. The newest architectures, like the EfficientNet family [153], scale the depths linearly. Preliminary experiments have shown that linear scaling performs best for the architecture presented in this thesis. Applying this scaling approach to this architecture, the hyperparameter count for the convolutional feature extractor decreases to just two - the starting depth and the depth at the end.

The remaining hyperparameters are much more straightforward to define. The BiFPN module has two hyperparameters. One hyperparameter defines how often the BiFPN block (also called BiFPN layer) gets repeated, and the other defines the feature size. The feature size, in this case, defines the feature map depth of the resulting tensors. As the outputs of the BiFPN module are merged into more manageable chunks using "feature compressor" convolution layers, the depth of that resulting feature map is a hyperparameter too. And last but not least, the size of the first fully-connected layer is a hyperparameter. A nice side-effect of this architecture

is that the network scales its parameters by itself depending on the image size, reducing the time required to find a good set of hyperparameters for different image sizes. Models following this architecture can have as low as 43,000 learnable parameters, with the most successful models in this thesis having between 50,000 to 350,000 parameters, which is a reduction in parameters of 98% to 99% over the state-of-the-art architectures while achieving equal or marginally better results. An overview of the hyperparameters and their search space can be seen in Table 4.2.

**Table 4.2:** The hyperparameters and their value range for the BiFPN-based model. As with the model based on a state-of-the-art feature extractor and attention, every adjustment can strongly influence the model performance, but their optimal values change from task to task.

| Hyperparameter | Search space |
|---|---|
| Starting depth of the feature extractor | Between 8 and 48 |
| Final depth of the feature extractor | Between 48 and 512 |
| Number of BiPFN layers | Between 1 and 16 |
| BiFPN feature size | Between 1 and 64 |
| Feature compressor feature map depth | Between 1 and 8 |
| Fully-connected layer size | Between 1 and 512 |

# Chapter 5

# Optimization, training, and analysis of forensic neural network models

## 5.1 Overview of the approach

Optimizing a neural network model is usually seen through the lens of "training." As described, training is the process of determining the optimal adjustable parameters of a neural network with the goal of minimizing the loss function. However, the training process itself has parameters that significantly influence the performance of the resulting model but that cannot be optimized by gradient descent. Those parameters are called hyperparameters. To optimize those parameters, non-gradient-based optimization and search algorithms need to be employed. The possible hyperparameters are numerous, but some of the most common ones are which optimization algorithm is used for training of the neural network, any decisions about the data preprocessing, representation of the target, the loss function, and even the neural network architecture itself and every part of it can be considered a hyperparameter. Many approaches to this problem exist, but this thesis explores the usage of grid search [25] and random search [26]. The most viable model candidates are determined and constructed, the hyperparameters are enumerated, and the search space is then explored by training the model "just enough" to determine their relative performance. The best-performing models are then fine-tuned - trained under different conditions requiring much more compute time but ultimately resulting in the near-best performance the model can achieve.

This chapter will detail the overall approach for model training and hyperparameter optimization, explain the differences between the employed search strategies, and highlight all specificities in the training and optimization procedure for each forensic task.

## 5.2 Hyperparameter optimization

### 5.2.1 Hyperparameter optimization as a search problem

Neural networks are designed in such a way that the gradient of the loss towards each parameter can be calculated, which allows for the usage of gradient-based optimization algorithms. However, those optimization algorithms have (hyper)parameters too, which significantly influence the resulting model's performance. Research decisions, too, can be considered a hyperparameter - what data and features are used, how the data is processed, what model is being used, and many more. The most apparent and naive approach is to simply evaluate all possibilities. That would, however, require an enormous (potentially infinite) amount of resources, most of which would be wasted on obviously non-functional models. The search space can be restricted, but evaluating all possibilities is still not a viable solution. Many search algorithms exist, but as the loss function is non-convex, most are not applicable for this task.

Genetic algorithms have proven themselves to be successful at general search problems. This type of algorithm starts from a random sample in the restricted search space and gradually, through a loop of evaluation, recombination, and randomization, searches for an optimum. In other words, a handful of randomly selected models are trained to convergence, and their performance is evaluated, after which they are recombined into new samples. This recombination is a function that, given two samples, creates a new sample that is "related" to the original two samples. How exactly they are "related" is up to the researcher and the implementation, which is again a factor that can significantly influence the resulting model's performance but cannot be optimized or deduced by itself. The randomization, also called "mutation," is a random change in the newly generated samples. The purpose of this randomization is to decrease the chance of getting stuck in local minima. However, the rate, magnitude, and target of this randomization are up to the researcher, which is again a factor that significantly influences the results while not being optimizable itself.

Bayesian search [24] is an informed search algorithm, which is, as the name implies, based on Bayes' theorem. Bayes' theorem says:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{5.1}$$

In the context of hyperparameter optimization, the theorem can be stated as:

$$p(\text{performance}|\text{hyperparameters}) = \frac{p(\text{hyperparameters}|\text{performance})p(\text{performance})}{p(\text{hyperparameters})} \tag{5.2}$$

This theorem describes a probabilistic model that links a model's performance probability and its hyperparameters. This effectively estimates the objective function with a surrogate function

that can be efficiently sampled. As this is an informed search algorithm, each step is dependent on its predecessors. While this might lower the number of iterations of the search algorithm, it tends to require more total time due to its consecutive nature.

Grid search [25] is a search algorithm that works in a restricted search space, and its steps are independent of each other. Instead of estimating the objective function, this approach samples a number of equidistant points in the search space, effectively spanning a grid. This approach can be turned into an informed search algorithm by repeating this process on more and more restricted search spaces, which would be informed by the previous sampling of the grid. Due to the independent nature of each point in the grid, this approach can leverage the parallel processing capability of modern high-performance compute (HPC) clusters. An example of grid search can be seen in Figure 5.1.

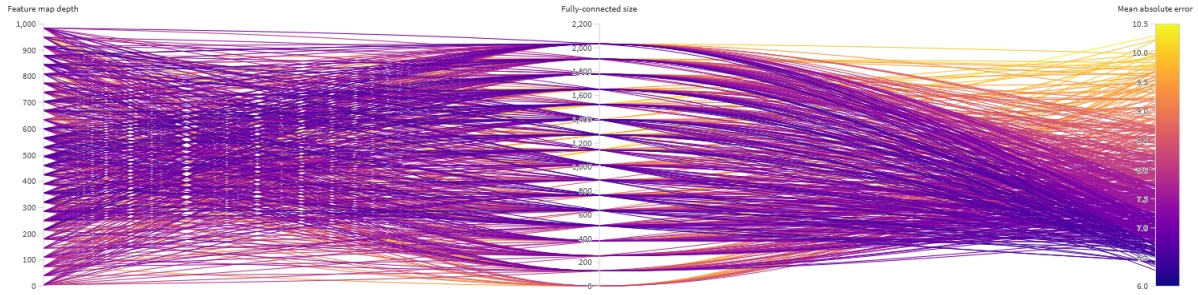


**Figure 5.1:** Parallel coordinate plot of hyperparameter optimization for age estimation. Each line represents one experiment, and each vertical bar represents the values of a hyperparameter. Each line is colored by its performance, shown in the rightmost colored column.

Random search also works in a restricted search space, and all of its steps are independent. Points are sampled randomly from the search space and then evaluated. While this approach seems wasteful and counterintuitive to the nature of searching, probabilistic analysis shows that random search can perform very well, which has proven itself true in practice.

A major problem in hyperparameter optimization is the so-called curse of dimensionality. As more hyperparameters are taken into consideration, the more potential samples there are to be evaluated. This increase in volume is exceedingly high. Each added hyperparameter adds exponentially more samples, inflating the search space and making it much harder to find the desired solution. In the domain of search problems, this problem is combated by careful analysis and selection of hyperparameters.

While all approaches are viable, this thesis focuses on an approach that combines grid and random search, thereby leveraging the parallel computing capabilities of modern HPC clusters while minimizing waste of resources. The following section will analyze these methods and explain how they have been combined and used.

### 5.2.2 Grid and random search

As already mentioned, grid and random search samples are independent, which allows for the proper exploitation of the parallel computing capabilities of modern HPC clusters. The search method in this thesis combines grid and random search into a semi-informed search algorithm. Each iteration of this method evaluates the sampled points independently, but each iteration is informed by its predecessor. The information obtained in each iteration is applied as further restrictions on the search space.

Random search [26] seems like an unintuitive approach for searching. The obvious question is, how can *random* search at the same time be random and search for something? While it is very unlikely that random search will find the global optimal minimum, the probability changes drastically if we introduce tolerance into the equation. Tolerance defines a distance range, expressed as a percentage, within which a point can be sampled and still be considered within the optimum. To get a better understanding of this idea, let us imagine a rectangle. The probability of randomly selecting any one specific point is infinitesimally small. However, if we define a range around that point, the probability that a point is sampled within that range is the ratio of the target area and the area of the entire rectangle. To translate this to the problem of hyperparameter optimization, the rectangle is the restricted search space, and the target area is the tolerance range around the optimum. For this to be a useful property, tiny changes in hyperparameter values must result in tiny changes in performance, which for neural networks holds true [23].

Formally, for a tolerance range $p_1$, a success probability $p_2$ of being within that tolerance range, and for $n$ samples, the following holds:

$$1 - (1 - p_1)^n > p_2 \tag{5.3}$$

For a tolerance range of $p_1 = 0.05$, and a probability of $p_2 = 0.95$, the number of samples needs to be $n \geq 60$. Depending on the tolerance and desired probability of success, the value of $n$ can be adapted accordingly, effectively translating into an increase in evaluated samples for an increased success probability and reduced tolerance. A visualization of success probabilities for a range of tolerances $p_1$ and number of experiments $n$ can be seen in Figure 5.2.

Let us examine this approach on a simple two-parameter model. As already mentioned, the loss function is non-convex and has multiple optima, some of which generalize better than others. The same holds true for the loss function in regard to hyperparameters. After the hyperparameters are enumerated, the potentially infinite search space must be restricted. Using grid search, equidistant points are sampled and evaluated. At the same time, randomly sampled points are evaluated around the best performing samples obtained by the grid search. This can be repeated until the tolerance and success probability are within an acceptable range. A

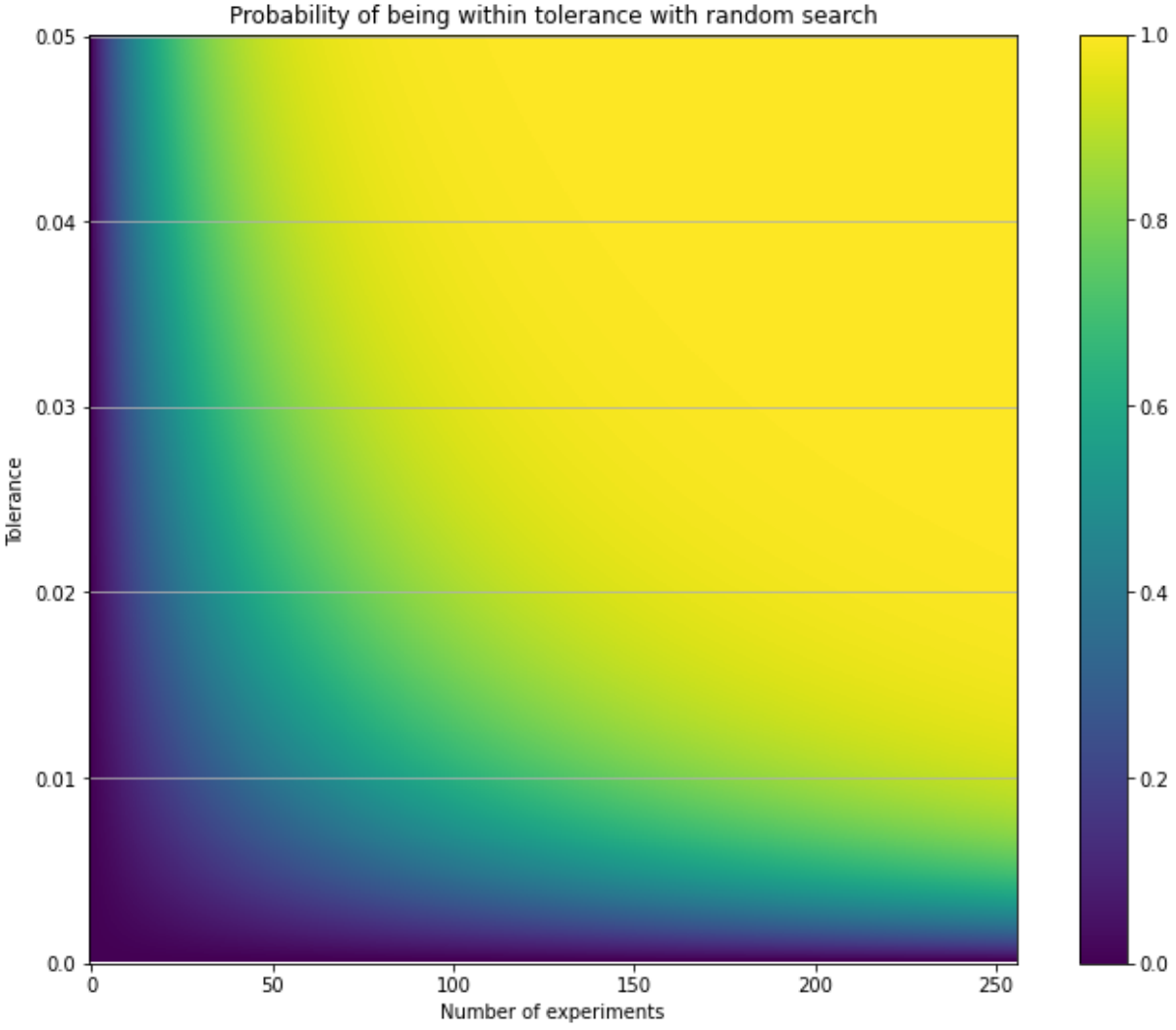


**Figure 5.2:** The probability map of a random search sample being within a tolerance per number of experiments. The x-axis represents the number of random samples (or the number of experiments for hyperparameter optimization), and the y-axis represents the tolerance from the global minimum in the restricted search space.

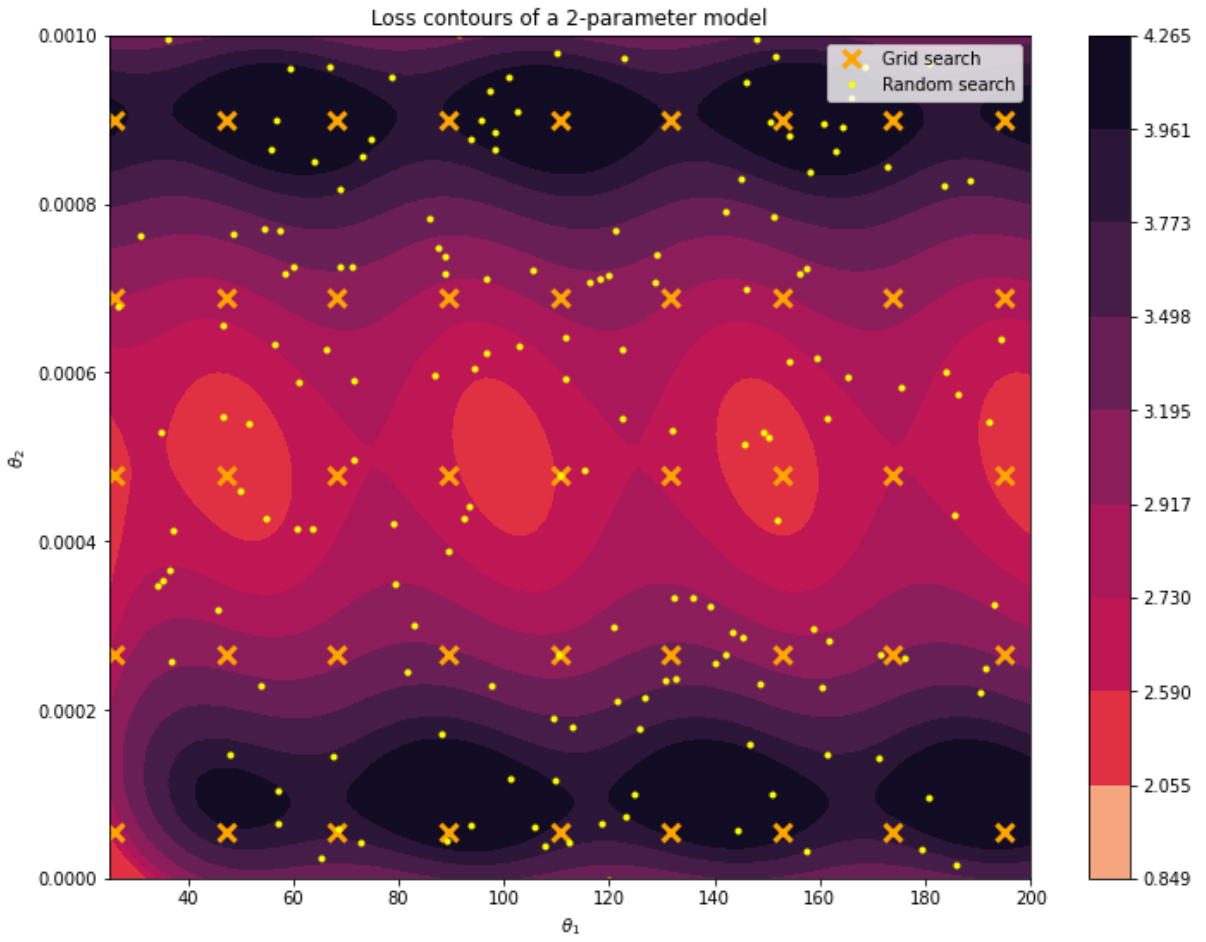


**Figure 5.3:** Visualization of grid and random search samples on a contour plot of a loss function for a 2-parameter model. The orange X shows points sampled by grid search, while the yellow dots show points sampled by random search. As can be seen, both methods manage to sample points close to the minima.

visualization of this example can be seen in Figure 5.3.

### 5.2.3 Training, fast and slow

Training a model is time-consuming. While performing hyperparameter optimization, every evaluated point is a fully trained model. This produces the most reliable results, but it is a huge waste of resources as all but one of the models will be discarded. To reduce the resources required while still maintaining good search results, training is different during hyperparameter optimization and during fine-tune training.

During hyperparameter optimization, models are trained using the Adam optimizer [19]. This is an adaptive optimization algorithm that tends to converge significantly faster than standard Stochastic Gradient Descent (SGD). Research has shown that adaptive methods tend to produce models that generalize worse [20]. However, this is not an issue during hyperparameter optimization, as the relative ranking of models is important, not the top performance. Ad-

ditionally, no learning rate schedule is used. The number of epochs the model trains for is also reduced, as, again, only the relative ranking between models is important. The values used in this thesis have been empirically determined with preliminary experiments. The Adam optimizer is using a learning rate of $3.24 \cdot 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, $\lambda = 0$, and training is limited to 100 epochs.

Once the best hyperparameters have been determined, models are fine-tuned accordingly. Fine-tuning is standard model training, but training hyperparameters are adjusted to produce the best performing models, which take a lot more time to train. The most obvious change is to the number of epochs trained. Instead of training for just 100 epochs, models are trained for 2048 epochs. Preliminary experiments have shown that models tend to converge around 1600 epochs, but this has been extended by 30% to ensure that the model truly has time to converge. Another change is the optimization algorithm used. Instead of Adam, conventional Stochastic Gradient Descent (SGD) is used. Models trained with adaptive methods tend to generalize worse [20], so SGD is chosen to produce the best possible final model. The combination of plain SGD and a small learning rate can lead to low-performing models due to getting stuck in local minima. Increasing the learning rate helps skip over narrow local minima but can lead to the inability to converge. A learning rate schedule is therefore used to help the optimizer. Specifically, cosine annealing with warm restarts is used [22]. Formally, this schedule can be defined as:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + cos\left(\frac{T_{cur}}{T_i}\pi\right)\right) \tag{5.4}$$

Here, $\eta_t$ is the learning rate in epoch $t$, $\eta_{min}$ and $\eta_{max}$ 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. Following this schedule, the learning rate starts at $\eta_{max}$ and decreases to $\eta_{min}$ in $T_i$ epochs. On epoch $T_i + 1$, the learning rate shoots back up to $\eta_{max}$, hence causing a "warm restart". Experiments have shown that the most successful parameters for the forensic odontology tasks tackled in this thesis are $\eta_{min} = 10^{-7}$, $\eta_{max} = 10^{-3}$, and $T_i = 50$. A batch size of 32 is used for the models based on state-of-the-art feature extractors and attention, and a batch size of 128 is used for the BiFPN-based model. The learning rate schedule for 250 epochs following those parameters can be seen in Figure 5.4.

The learning rate can be interpreted as the resolution at which the loss function is sampled. In each optimization step, the gradients at the current point of the loss function are calculated, and the learnable parameters are modified to move towards a minimum as defined by the optimization algorithm. The loss function is non-convex and has many local minima, many of which do not result in a high-performance model. By increasing the learning rate, only wider basins of attraction will affect our optimization step. This holds true in neural networks, where early research shows that smoothing approaches improve neural network training performance [154]. Intuitively, we only let larger basins of attractions affect us, and by gradually decreasing

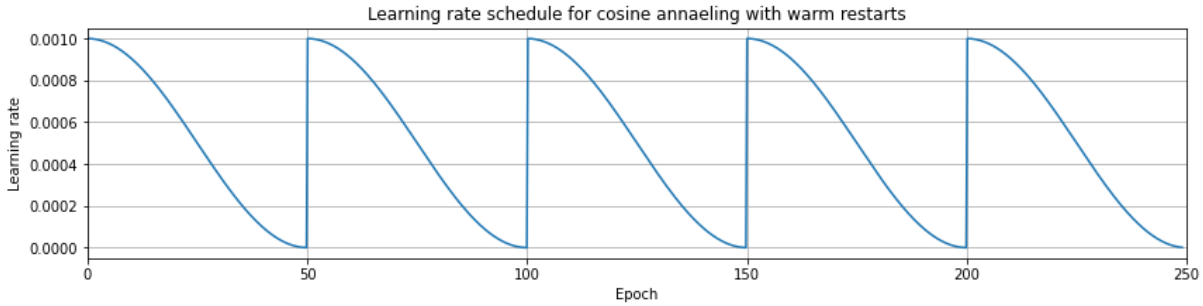


**Figure 5.4:** The learning rate following the cosine annealing with warm restarts schedule. Shown are 250 epochs with the schedule parameters being $\eta_{min} = 10^{-7}$, $\eta_{max} = 10^{-3}$, and $T_i = 50$.

the learning rate, the model is allowed to converge into its nearest local minima. With the sudden warm restart, if the basin of attraction of the local minima is too small (which is a property of poorly performing minima), the optimization step will move us out of it. However, if the basin of attraction of the minima is larger, the model will temporarily perform worse, but it will re-converge to the well-performing local minimum.

This dual approach to training allows for a significant saving in computing resources while simultaneously achieving the search performance of more demanding search methods. The fast approach generates a ranking between hyperparameters, and then the best performing models can be trained to their full capacity. Experiments have been done for the selection of the top models. To avoid letting minor variances in training performance lead to worse model selection, the best five models were fine-tuned instead of just the very-best model per forensic odontology task. However, this precaution turned out to be unnecessary, as the ranking between the top-5 models did not change with fine-tuning.

## 5.3 Model training for classification and regression tasks

### 5.3.1 Age estimation with direct regression

While the overall approach of model design, hyperparameter optimization, and fine-tuning is applied to all forensic odontology tasks researched, every task has specific changes in the data processing and training approach. Age estimation with direct regression is used for full panoramic dental x-ray images and individual tooth x-ray images.

The models used are described in Chapter 4. Models that process panoramic dental x-ray images use transfer learning [34], and models that process individual dental x-ray images do not. Preliminary experiments have shown that models consistently converge faster and perform better when using transfer learning, while it does not make a difference to individual dental x-ray images. This is due to the sample size, as 86495 individual tooth x-ray images are enough to train a state-of-the-art feature extractor model from scratch, but 4035 panoramic dental x-ray

**Table 5.1:** List of augmentations and their hyperparameters used during training for age estimation.

| Augmentation | Hyperparameters |
|---|---|
| Left-right flip | p: 50% |
| Coarse dropout | Image rescale factor: 2% to 5% |
| | Amount of dropped pixels: 0% to 1% |
| Average blur | Square kernel size in px.: 0 to 2 |
| Gamma contrast | gamma: 0.85 to 1.15 |
| Rescaling | x: 1 to 1.4 |
| | y: 1 to 1.2 |

images are not.

The loss used is the mean squared error function, which is formally defined as:

$$\mathscr{L} = (y_{\text{true}} - y_{\text{pred}})^2 \tag{5.5}$$

There $y_{\text{true}}$ is the real age of the sample, and $y_{\text{pred}}$ is the age estimated by the model. The ground truth data is not modified and is represented as a floating-point number of the imaged person's age in years. In literature, regression outputs are often scaled to the interval between 0 and 1. Preliminary experiments have shown that this choice did not influence the results; therefore, no transformation was done on the ground truth data for age.

The images are resized to 512 px by 512 px, as models using this image size provided the best results. Image augmentation is evaluated, too, as research shows that it can improve model performance [33]. Augmentation is the process of randomly modifying the image during training. Those augmentations do not significantly change the image, thus leaving the target information in the image intact. This addition of noise effectively increases the number of samples in the dataset. An example of image augmentation can be seen in Figure 5.5. The list of augmentations used and the values of their hyperparameters are shown in Table 5.1.

For panoramic dental x-ray images, three different data variants were evaluated. The first data variant is the baseline - the model processes the full panoramic dental x-ray image and estimates the age. The second data variant uses image augmentation to increase the effective size of the dataset. This variant is used to show if the usage of image augmentation is justified and how much it impacts the model performance. The third data variant is stratified by age. As described in Chapter 3 (Data), the dataset for panoramic dental x-ray images is biased towards younger samples. This bias will be adopted by the model, which in turn will lead the model to estimate younger ages when uncertain, as it statistically would be the better choice. Stratifica-

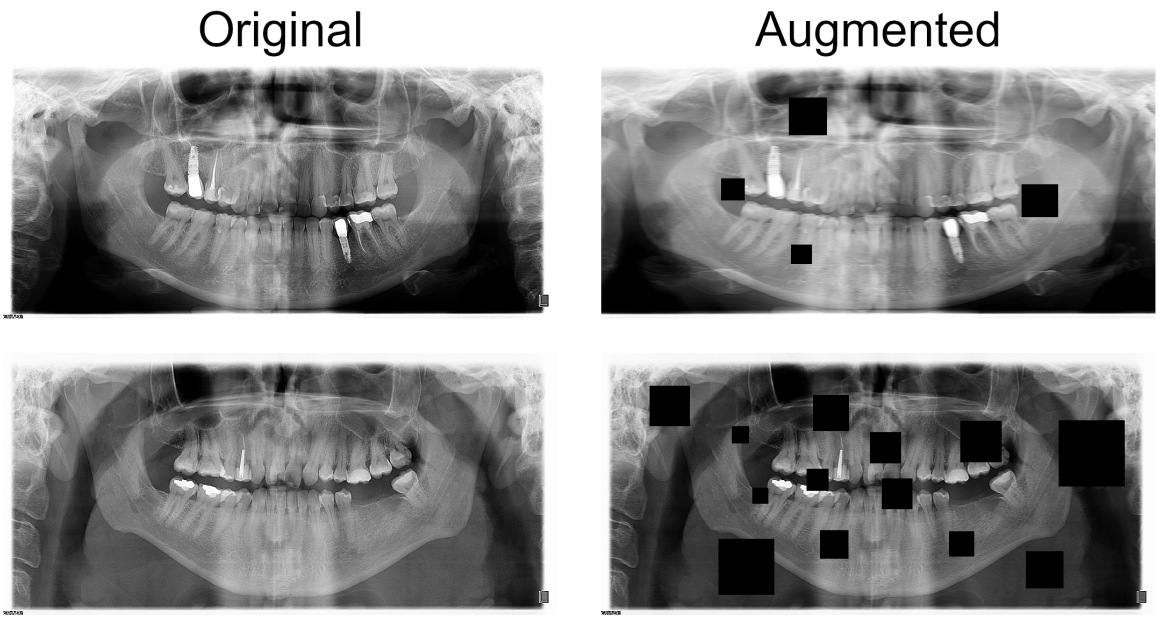


**Figure 5.5:** Two examples of panoramic dental x-ray images with applied augmentation. The first row shows the effects of gamma contrast, blurring, rescaling in the y dimension, and random coarse dropouts. The second row shows gamma contrast change opposite of the first row, a lower degree of blurring, more random coarse dropouts, and no affine transformations.

**Figure 5.6:** Examples of the three approaches used to determine the most information-rich area for age estimation. The same panoramic dental x-ray image is shown three times, (left) with the teeth roughly covered, (middle) with only the teeth visible, and (right) with the teeth finely covered.

tion is applied by splitting the dataset into 5-year-long age groups and equalizing the number of samples in each with duplication. This approach also uses augmentation, as the noise will counteract some of the adverse effects of sample duplication.

In addition to these data variants, this thesis explores how much teeth and the surrounding structures independently contribute to a correct age estimation. The best performing model, which includes the augmentation data variant, is retrained on three additional data variants. In the first variant, only teeth are visible. In the second variant, the teeth are covered while the rest of the tissue is visible. The third variant also shows only the surrounding skeletal structure. However, the teeth are precisely covered by their bounding box, leaving as much of the surrounding skeletal structure visible as possible. An example of those image variants can be seen in Figure 5.6. A detailed overview of the results for direct regression age estimation for panoramic dental x-ray image analysis models is given in Section 6.1.1.

Age estimation from individual teeth images is a more challenging problem, as less infor-

mation is available. The approach is generally the same, with slight differences in the data preprocessing and models. As described in Section 3.4, no individual tooth x-ray image exceeds 528 px in either dimension. To avoid any loss of information or introduction of some artifacts due to resizing, the models are trained on images of size 528 px by 528 px. Sizes differ from image to image, but to effectively use them with neural networks and GPU acceleration, their size needs to be the same. Therefore, all individual teeth images are placed within a 528 px by 528 px image, where all values are initially set to 0. For training, the image is placed randomly while ensuring that no part of the image is cropped out. During validation and testing, all images are placed centrally to ensure consistency between experiments.

Two metrics are used to evaluate the performance of direct regression age estimation models. Those are mean absolute error, which shows the statistically expected value of the error, and median absolute error, which shows the model performance without the influence of extreme outliers. The mean absolute error is calculated as follows:

$$\mathrm{MSE}(y_{\mathrm{true}}, y_{\mathrm{pred}}) = |y_{\mathrm{true}} - y_{\mathrm{pred}}| \tag{5.6}$$

The median absolute error is determined by calculating the absolute error for all samples, sorting that list, and taking the middle element. In case when the length of the list is even, and there is no middle element, the middle two elements are averaged to obtain the median absolute value. An analysis per age group, (when applicable) tooth type, and tooth alterations is performed to determine the model's trends, impacts, and potential biases. A detailed overview of the results for direct regression age estimation for individual tooth x-ray image analysis models is given in Section 6.1.3.

### 5.3.2 Estimating age as a probability distribution

Age estimation in adults is a complex problem. For younger samples, age can be estimated by determining the stage of development of their teeth, as tooth development is strictly defined by genetics and can therefore be used to estimate the age with an error measured in months. Inherently, estimating age carries some uncertainty. Likewise, the same error value is not equally detrimental across the entire age spectrum. Mistaking an 18-year-old for a 27-year-old is a huge error, but mistaking a 70-year-old with a 79-year-old is still a mistake, but a far less egregious error. To address the inherent uncertainty and allow for the error to be weighted differently across the age spectrum, this approach does not estimate age directly. Instead, it estimates the probability of each discrete age category [155].

The estimation is realized as a discrete distribution of probabilities for every age between 0 and 100 years. While the entire probability distribution is the estimation, the expected value of this distribution is taken as the numerical value of the estimation. However, as the estimation is a

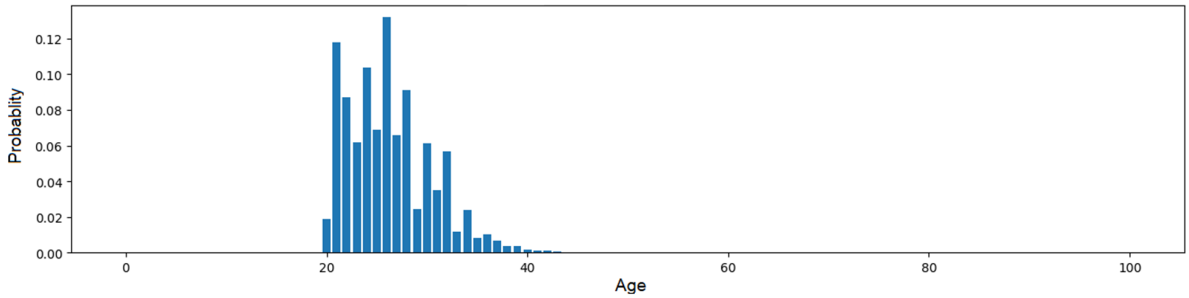


**Figure 5.7:** An example of an estimated probability distribution. The model assigns a probability for every possible age, and the estimated age is calculated as the expected value of the predicted distribution. A narrow distribution can be interpreted as a more confident prediction. In this sample, the true age is 27.49 years, and the model estimated an age of 26.28 years.

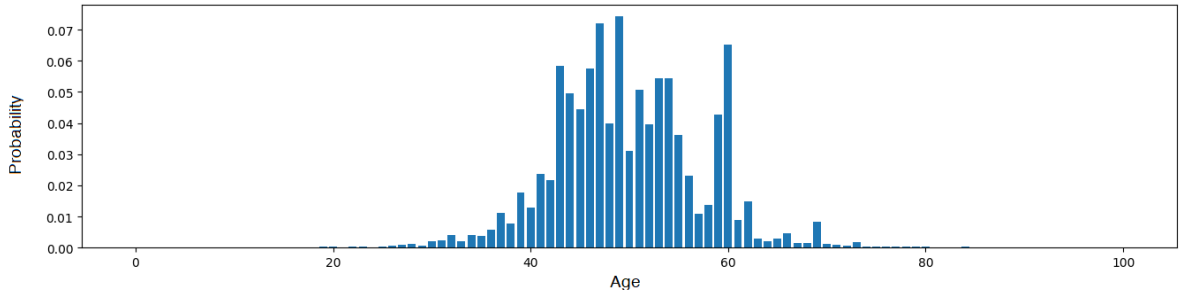


**Figure 5.8:** An example of a wider estimated probability distribution. The variance of the estimated distribution is higher, which can be interpreted as a lower model confidence. In this sample, the true age is 50.18 years, and the model estimated an age of 49.98 years.

probability distribution, the variance can also be calculated. The variance varies from sample to sample, as it describes the uncertainty of the model for that sample. A higher variance indicates a sample that is harder for the model to analyze. It is important to note that this uncertainty does not necessarily correlate with the estimation error, as the model can be uncertain and still correct. In essence, the uncertainty is correlated with some sample difficulty inherent to the trained model. To better understand the problem formulation of this approach, a visualization of the model output is shown in Figure 5.7.

This approach has been tested with the models described in Section 4.1. As described in Chapter 3 (Data), the ages range from 19 to 90. Despite that, the models are designed to predict from age 0 to 100 to give the model the room to express its uncertainty. An example of a more uncertain, and thus wider, prediction can be seen in Figure 5.8.

The image data is prepared the same way as described in Section 5.3.1. Both models are processing the same images with the same goal, the only difference being the modality of the output. The ground truth however needs to be transformed into a different representation. Instead of age being represented as a floating-point number of years, the age is transformed into a one-hot encoded vector, where the category is the rounded value of the age.

The loss combines the mean squared error and the categorical cross-entropy loss, which

is regularized by the variance of the estimated probability distribution. To properly tune the model, this three-component loss introduces two hyperparameters. The loss is formally defined as:

$$\mathscr{L} = \text{CCE}(y_{\text{true; onehot}}, y_{\text{pred}}) + \lambda_1 MSE(y_{\text{pred; expected value}}, y_{\text{true}}) + \lambda_2 \text{Var}(y_{\text{pred}}) \tag{5.7}$$

CCE is the categorical cross-entropy, MSE is the mean squared error, Var is the variance of the distribution, and $\lambda_1$ and $\lambda_2$ are the hyperparameters. The expected value and variance are calculated as:

$$\mu = y_{\text{pred; expected value}} = \sum_{j=0}^{N} j p_j \tag{5.8}$$

$$\text{Var} = \sum_{j=0}^{N} p_j (j-\mu)^2 \tag{5.9}$$

In this formulation, $p_j$ is the probability of age $j$; $p_j$ is $y_{\text{pred}}$ in the loss formulation. With this loss, the penalization due to a bad estimation can be lessened if the model is uncertain in its estimation. This allows the model to improve its performance, as optimization will not degrade learned features that work well on most samples for the sake of improving estimation on an outlier sample.

The evaluation approach is similar to Section 5.3.1. The same trends per age group are analyzed, and additional analysis of the models' uncertainty for low and high error samples, as well as the general uncertainty trends with age.

### 5.3.3 Sex assessment

Sex assessment is another routine task of forensic odontology. Sexual dimorphism, the structural differences in the bodies of different sexes, can be observed throughout the entire human body, and the dental system is no exception. While it is stronger expressed in skeletal structures than in teeth, sex can still be assessed with reasonable accuracy from individual teeth. This task can be formulated as a classification problem with two classes. The model discovery and hyperparameter optimization approach for sex assessment are the same as described in Section 5.1.

As with age estimation by direct regression, sex assessment uses both models described in Chapter 4. The models are modified to include two outputs (one for the probability of each class), like age estimation by direct regression has just one output, and age estimation with a probability distribution has 100 outputs. The loss function used is categorical cross-entropy

(CCE). Formally, CCE is defined as:

$$\mathrm{CCE} = \frac{1}{N} - \sum_{i=0}^{N} y_{\mathrm{true;\,i}} \cdot \log(y_{\mathrm{pred;\,i}}) \tag{5.10}$$

Where $y_{\mathrm{true;\,i}}$ is the one-hot encoded category of sample $i$, and $y_{\mathrm{pred;\,i}}$ is the model prediction for that sample, and $N$ is the number of samples.

For panoramic dental x-ray images, the models based on state-of-the-art feature extractors and attention are resized to 512 px by 512 px, as preliminary experiments have shown that this size produces the best-performing models. For the BiFPN-based model, the best performing size was 256 px by 256px, so images are resized accordingly for that model. In both cases, augmentation is used.

For individual tooth x-ray images, images are of size 528 px by 528 px. As described in Section 3.3.2, all teeth can fit within a 528 px by 528 px image in their original form; therefore, no resizing is done to individual tooth x-ray images.

As this is a classification problem, and as the classes are mostly balanced, accuracy is used as the performance metric. Accuracy is defined as:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5.11}$$

Acc is the model's accuracy, TP is the number of true positive samples, TN is the number of true negative samples, FP is the number of false positive samples, and FN is the number of false negative samples. An analysis per age group and (when applicable) tooth type and dental alterations is performed to determine the trends, impacts, and potential biases of the model.

### 5.3.4 Tooth type determination of individual tooth x-ray images

The tooth type is an important factor in classical forensic odontology approaches. Tooth measurements are compared to reference tables per tooth type, therefore, knowing the tooth type is a prerequisite for sex assessment and age estimation in classical forensic odontology. This is a classification problem, and therefore CCE loss is used, which is defined in Section 5.3.3. Only individual tooth x-ray images are used; hence, as described in Section 3.3.2, all images are of the size 528 px by 528 px. Image augmentation is not used. Both models described in Chapter 4 are used.

Tooth type classification presented another challenge. The standard approach to tooth type classification is defined by ISO-3950 standard [139], also commonly known as the FDI dual notation system. The FDI dual notation system assigns two numbers to every tooth based on its location; one for the quadrant it resides in and one for its location within the quadrant. This totals 32 classes, one for each tooth. However, in literature, different tooth type classification

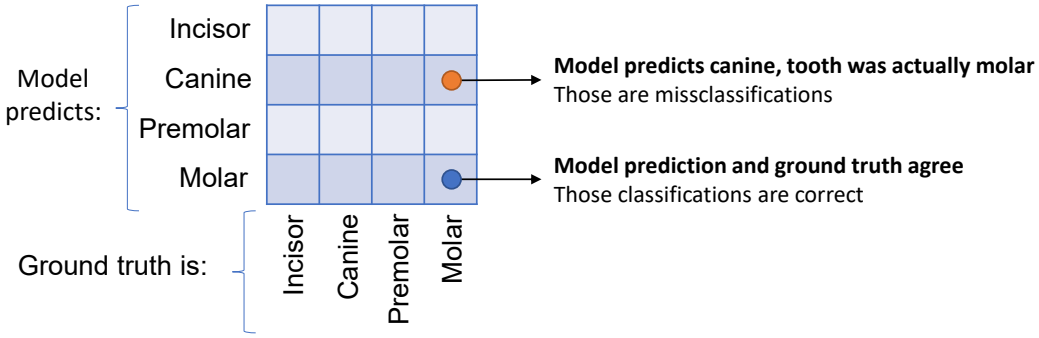


**Figure 5.9:** An example of a confusion matrix for a 4-class tooth type determination model. The rows represent the model predictions, and the columns represent the ground truth values. The figure shows an example of a correct classification (molar-molar) and a misclassification (canine-molar).

systems are used. Some classical age estimation methods consider teeth to be left-right symmetrical, thus merging quadrants 1 and 2, 3 and 4, resulting in 16 classes, two teeth per class. Some research drops the distinction between mandibular and maxillary teeth, resulting in an 8-class system. Finally, some research differentiates teeth into the four basic types by function: incisors, canines, premolars, and molars. For the sake of completeness and general usability, this thesis examines and designs models for all four classification approaches.

Tooth type determination, like sex assessment, is a classification problem. As the classes are mostly balanced, accuracy is used as the evaluation metric. In addition to the analysis of trends per age group and the impact of alterations, an analysis of misclassifications is performed to determine which tooth relations cause the most issues for the model. This analysis is done using the confusion matrix. A confusion matrix is an N by N matrix, whereby N is the number of classes. Each row represents predicted classes, and each column represents the true classes, with the number representing either the number or percentage of samples that belong to a specific prediction-ground truth pair. Consequently, the elements on the diagonal represent the correct classification, and the neighboring values represent how the model misclassified samples. An example of a confusion matrix for a 4-class tooth type determination system is shown in Figure 5.9.

### 5.3.5 Multi-task models

Given all the segmentation and regression tasks, and given that the input data is the same, it might be natural to assume that a multi-task model would perform equally or better, while at the same time requiring fewer resources. As all images are ultimately teeth or the entire jaw, and the demographic data is linked to some morphology seen on those images, it is fair to assume that there are shared features between multiple demographic information estimation tasks. As the model is estimating multiple demographic facts, there might be some correlation between features and different demographic information that could be exploited to achieve better results.

To that end, this thesis explored the usage of the models described in Section 4.1 for this multi-task approach. The loss of such a model is the weighted sum of losses for every subtask. Formally, for age estimation and sex assessment, the loss is:

$$\mathscr{L} = \mathrm{CCE}(x,y) + \lambda \mathrm{MSE}(x,y) \tag{5.12}$$

CCE is the categorical cross entropy (for sex assessment), and MSE is the mean squared error (for age estimation). In this loss, $\lambda$ is the tradeoff hyperparameter.

During hyperparameter optimization, a total of 457 experiments have been performed, and no successful model has been found. No model could perform both sex assessment and age estimation successfully. Some models reached performance close to single-task models, but the other task would perform significantly worse than its single-task counterpart. Figure 5.10 shows the performed experiments. As this approach was unsuccessful, Chapter 6 (Results) does not contain a detailed analysis of the results.

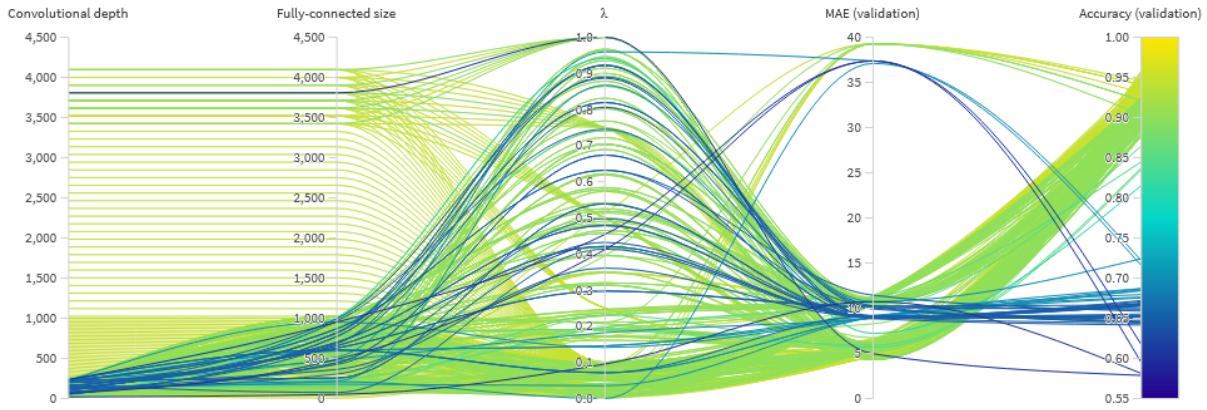


**Figure 5.10:** Parallel coordinates plot of grid and random search for multi-task models that simultaneously estimate age and assess sex. Grid and random search initially used a range for $\lambda$ between 0 and 1. On the second iteration of hyperparameter optimization, the range was reduced between 0 and 0.1, as that interval performed best.

## 5.4 Segmentation of tooth alteration in panoramic and individual dental x-ray images

Segmentation is a different problem from age estimation and sex assessment, but it is nonetheless very important for forensic odontology. It is a form of classification, but instead of doing one general classification, segmentation classifies every pixel of the input image individually. Teeth can be afflicted by different alterations, and those alterations prevent the classical forensic odontology method from analyzing those teeth. Additionally, a segmentation of the entire tooth

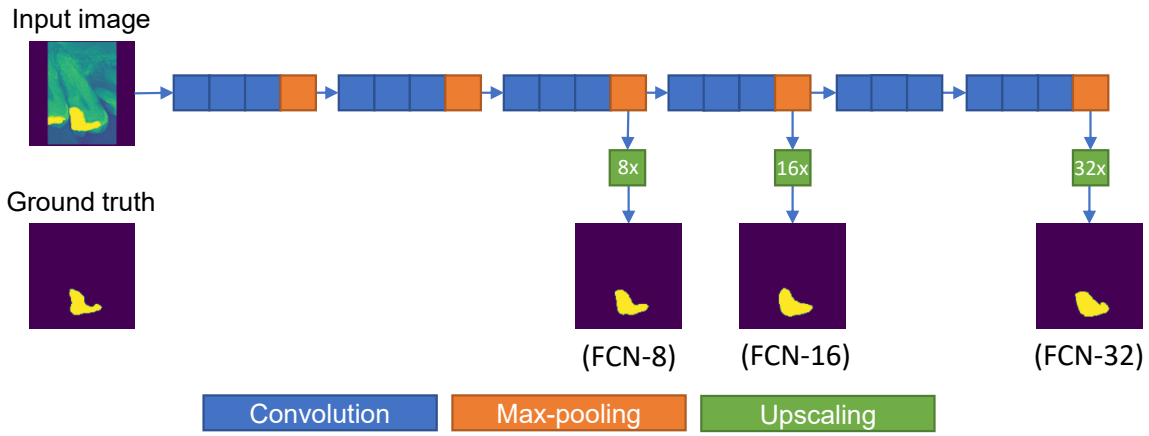


**Figure 5.11:** The architecture of the FCN-8, FCN-16, and FCN-32 models. The models use a similar convolutional backbone and differ in the upscaling factor used to produce the segmentation mask.

allows for the removal of all surrounding tissue and thus excess information, making measurements easier and enabling automated methods to dedicate their capacity to estimation, not noise removal.

In this thesis, segmentation is performed on panoramic dental x-ray images, as well as individual dental x-ray images. A data variant of individual tooth x-ray images is also evaluated, where the surrounding structures are removed. As every tooth has a segmentation map for the tooth itself, the surrounding structure can easily be removed. The dataset is slightly different from age estimation, sex assessment, and tooth type determination due to the complex and time-consuming nature needed for segmenting images by hand. An overview of the 813-image dataset is given in Section 3.3.3. For the experiments performed in this thesis, panoramic dental x-ray images are resized to 768 px by 768 px, and individual images are resized to 128 px by 128 px.

State-of-the-art segmentation models are well established in the literature. The three most used and successful models in the literature are Fully Convolutional Networks [156], DeepLab v3 [157], and UNet [129]. Fully convolutional networks (FCNs), as their name suggests, only use convolutional layers. This allows them to accept images of any size. There are three variants of FCN - FCN-32, FC-16, and FC-8. Those numbers refer to the upsampling rate used after feature extraction by the network. In other words, for FCN-32, once the network processes the image, the final feature map is upscaled by a factor of 32. This, in turn, causes a rougher segmentation map, as more spatial information was lost due to the depth of the network. Likewise, FCN-16 upscales its final feature map by a factor of 16, and FCN-8 upscales it by a factor of 8. The FCN architecture can be seen in Figure 5.11.

DeepLab v3 is a deep-learning architecture for segmentation that leverages atrous convolutions and Atrous Spatial Pyramid Pooling (ASPP). Atrous convolutions, also called dilated convolutions, are convolutions that introduce a spacing between the values of their kernel. This

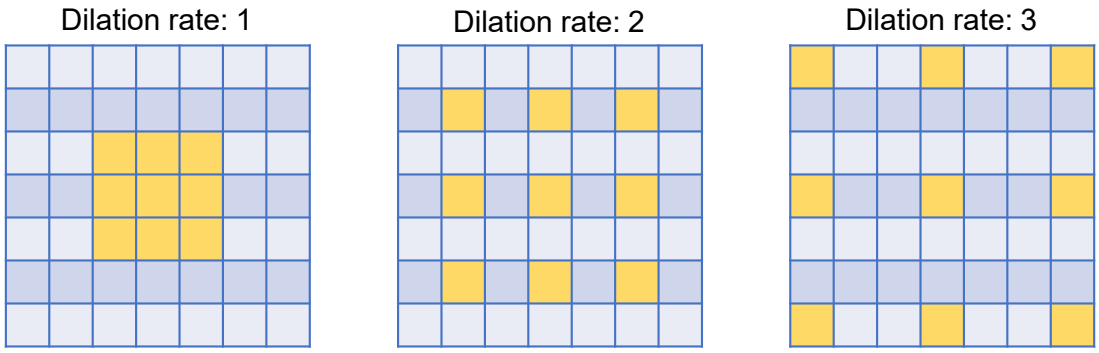


**Figure 5.12:** Visualization of the processing of atrous (dilated) convolutions. The dilation rate increases the distance between neighboring values, leaving a gap between them.

spacing is called the dilation rate. A 3x3 convolution with a dilation rate of 2 has an effective field-of-view of size 5x5. Equivalently, a 3x3 convolution with a dilation rate of 1 is just a conventional convolution. This way, a wider area can be processed with an overall lower parameter count. Feature maps of conventional convolutions shrink layer by layer, summarizing the image and estimating the target information from that representation. However, segmentation requires spatial information, making it harder to construct a segmentation map for a "fully summarized" image. Atrous convolutions can maintain the same spatial dimension of a feature map and simultaneously allow for the processing of features on different scales, as the dilation rate defines the field of view that the convolution processes. DeepLab v3 uses atrous convolutions with a dilation rate of 6, 12, and 18. A visualization of a atrous convolution can be seen in Figure 5.12.

The Atrous Spatial Pyramid Pooling module of DeepLab v3 is used for multi-scale processing of the feature extractors' final feature map. Atrous convolutions of different dilation factors process the feature map at effectively different scales due to their field-of-view, but all resulting feature maps retain the same spatial dimensions. The output of all those atrous convolutions is concatenated and processed by a 1x1 convolution to reduce the number of channels. A diagram of the Atrous Spatial Pyramid Pooling module can be seen in Figure 5.13.

The DeepLab v3 architecture combines those two elements with any feature extractor to form the full model. The most used feature extractors, and those analyzed in this thesis, are MobileNet [158], HRNet V2 (W32 and W48) [159], and ResNet (50 and 101) [29].. The rest of the architecture is the connective tissue between those elements. The full architecture consists of the feature extractor, the ASPP module, two 3x3 convolutions that reduce the number of channels to the number of classes the model is segmenting, and finally, a bilinear upscaling by a factor of 8. The architecture can be seen in Figure 5.14.

UNet is a deep-learning image segmentation architecture that has a characteristic U shape. It uses only the most basic elements of convolutional neural networks: 3x3 and 1x1 convolu-

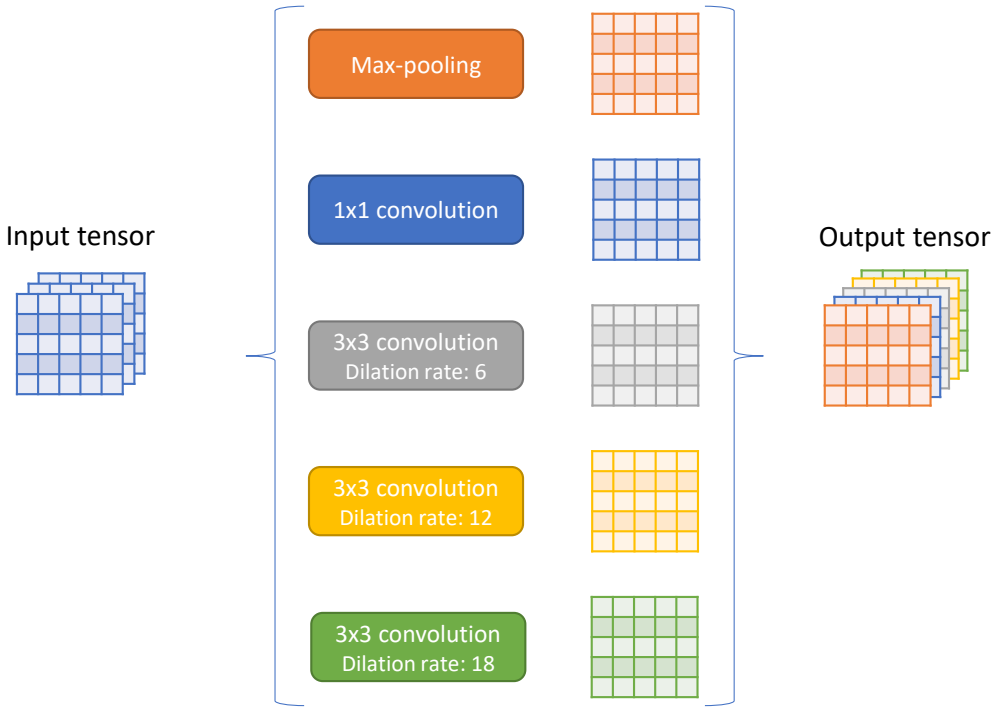


**Figure 5.13:** Visualization of Atrous Spatial Pyramid Pooling. The same feature map is processed at different scales using atrous convolutions. This produces feature maps of the same size, but due to the dilation factor, the processing is effectively on different scales.

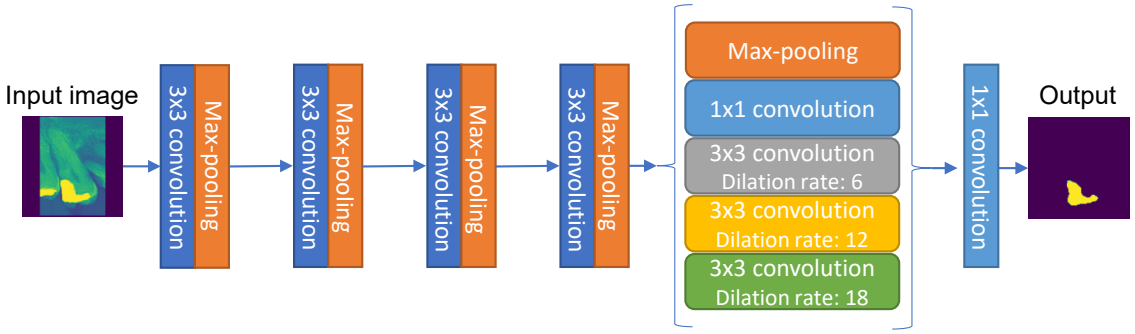


**Figure 5.14:** The DeepLab v3 architecture. A convolutional backbone processes images and Atrous Spatial Pyramid Pooling processes the resulting feature map. Next, the outputs of the ASPP module are concatenated and finally processed by a 1x1 convolution that produces the predicted segmentation mask.

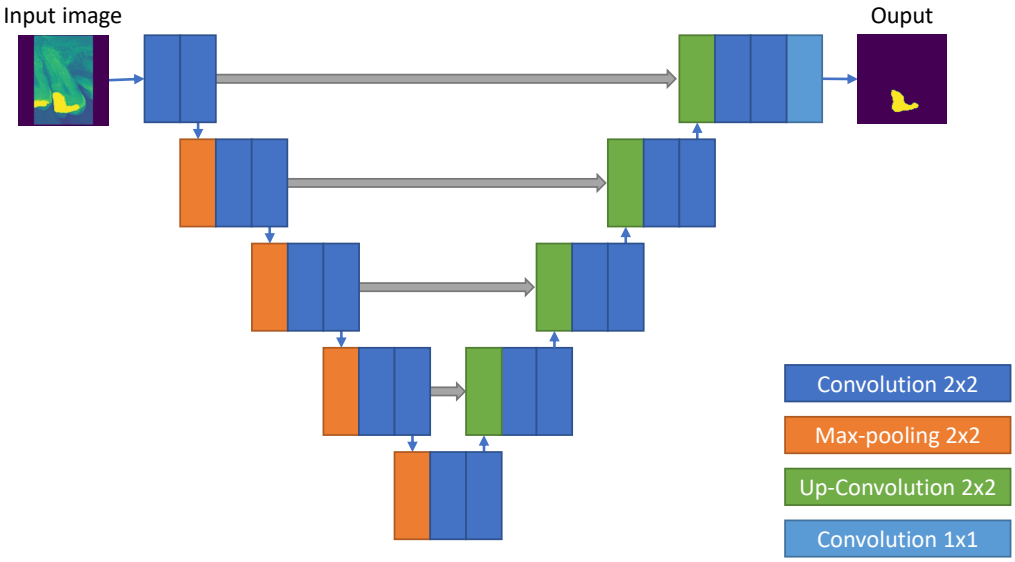


**Figure 5.15:** The baseline UNet architecture. An image is processed by a sequence of convolutions and max-pooling, after which it is upscaled using up-convolutions. Skip connections are added between the same scale in the "down" and "up" parts of the network. The network capacity is adjusted by changing the number of channels across all convolutional layers.

tions, 2x2 max-pooling, and 2x2 up convolutions (also called transposed convolutions). The architecture follows the encoder-decoder schema, where one half of the network (the "down network") reduces the spatial dimension and increases the depth of the feature map as the network progresses, and the other half (the "up network") increases the spatial dimension while it reduces the depth of the feature map, eventually producing an output of the same size as the input image, and with a number of channels equal to the number of segmented classes. Skip connections across the subnetworks are used between layers that produce tensors of the same spatial dimension. An overview of the architecture can be seen in Figure 5.15. In addition to baseline UNet, this thesis proposes three additional variants that change the model's capacity. All variants have the same architecture, but the number of output channels of each convolution is scaled up or down to change the model's capacity. The smallest variant, UNet Micro, has 1/16th of the output channels for every convolution compared to the baseline UNet. The second smallest variant, UNet Mini, has 1/4th, and the biggest variant, UNet Big, has 2x the number of output channels in every convolutional layer compared to baseline UNet. An overview of all networks and their number of parameters can be seen in Table 5.2.

Training segmentation networks is slightly different from classification and regression models. While classification and regression models achieved better results when trained with SGD and cosine annealing with warm restarts as the learning rate schedule, segmentation model performance was better when trained with the Adam optimizer. This holds true not only for segmentation of the entire tooth but for every tooth alteration, and both for segmentation of individual tooth x-ray images and for the entire panoramic dental x-ray images.

**Table 5.2:** The number of parameters of all evaluated segmentation architectures and their variants. UNet has the largest variance, while the differences for FCN variants are minor. DeepLab v3 variants are defined mainly by their feature extractor, as the segmentation head itself is not overly large relative to the feature extractor.

| Base model | Variant | Parameter count |
|---|---|---|
| Unet | Micro | 68 266 |
| | Mini | 1 081 762 |
| | Baseline | 17 266 306 |
| | Big | 69 037 314 |
| FCN | FCN32 | 134 271 430 |
| | FCN16 | 134 273 924 |
| | FCN8 | 134 285 122 |
| DeepLab v3 | MobileNet | 5 108 994 |
| | HRNet V2 W32 | 33 790 658 |
| | ResNet50 | 39 633 986 |
| | ResNet101 | 58 626 114 |
| | HRNet V2 W48 | 71 592 002 |

Three metrics are most often used for segmentation: per-pixel accuracy, intersection over union (Jaccard index), and the Sørensen–Dice coefficient. Per-pixel accuracy treats every pixel as a classification sample, which can either be classified as a true positive, true negative, false positive, or false negative. This metric is a useful indicator when the segmentation maps are not sparse. For sparse segmentation maps, most pixels are correctly classified as true negatives, which inflates the accuracy score. Segmentation of dental alterations and teeth is a task with sparse segmentation maps, which makes accuracy an unsuitable metric. For example, tooth decay is often not widespread in panoramic dental x-ray images, as decay is quickly fixed when found by dentists. Teeth in a panoramic dental x-ray image, if they even have decay, have just one or two instances of it. This leads to a sparse segmentation map. In this case, even if the model always outputs an empty segmentation map, the accuracy is still over 95%, as it is technically correct that most of the image does not contain tooth decay.

The Jaccard index, also called intersection over union, solves this problem by calculating the percentage of overlap between the ground truth and predicted region. The bigger the overlap, the higher the value of the resulting metric. This avoids the issue of the true negative bias that

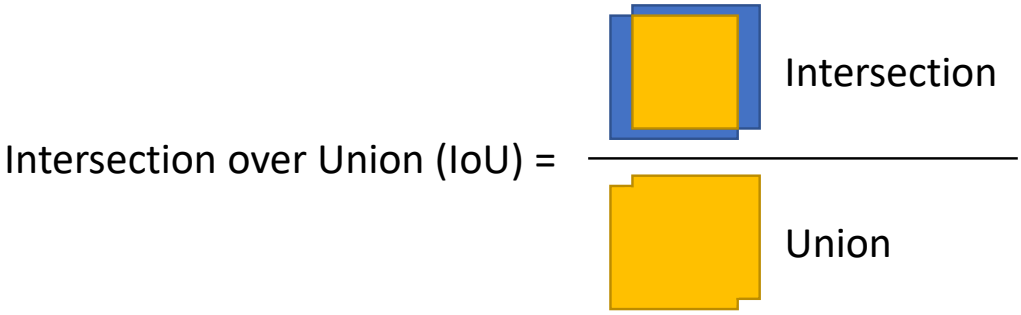


**Figure 5.16:** The Jaccard index, also called the Intersection over Union. The metric describes the percentage of overlap between the ground truth and predicted regions, thus avoiding the true negative bias of the per-pixel accuracy.

per-pixel accuracy has. Formally, the Jaccard index is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{5.13}$$

In this equation, A is the ground truth region, and B is the predicted region. This equation can be visualized as shown in Figure 5.16.

The Sørensen–Dice coefficient is somewhat similar to the Jaccard index, as it too calculates the overlap between the ground truth area and the estimated area. Moreover, the Sørensen–Dice coefficient and the Jaccard index are always positively correlated. However, those metrics are not functionally equivalent. Similarly to L2 and L1 errors, the Jaccard index penalizes instances of wrong classification significantly more than the Sørensen–Dice coefficient. In aggregate, the Sørensen–Dice coefficient represents the performance closer to average, while the Jaccard index represents the performance closer to the worst-case scenario. Formally, the Sørensen–Dice coefficient is defined as:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2TP}{2TP + FP + FN} \tag{5.14}$$

In this equation, X and Y are the ground truth and estimated regions, respectively. The Sørensen–Dice coefficient can also be used for boolean data. The second part of the equation defines the DSC in that scenario. TP is the number of true positives, FP is the number of false positives, and FN is the number of false negative classifications (per-pixel). Segmentation studies in medical applications, and especially forensic odontology, report results using the Sørensen–Dice coefficient.

## 5.5 Detection and type determination of teeth in panoramic dental x-ray images

In the context of forensic odontology, detection is used to locate individual teeth in panoramic dental x-ray images. While teeth are in a somewhat similar position in relation to each other, their absolute position and size on the image change due to differences between people. In forensic cases, further complications are possible due to fractures of the jaw and the skull, as well as if teeth are not affixed to a jaw. There are many object detection approaches in the literature, from the traditional approaches like the Viola-Jones detector [7], HoG detector, and DPM, to current deep learning-based approaches. In deep learning, there are two branches of object detection networks - one- (like the YOLO model family, SSD, RetinaNet) and two-stage (like the R-CNN model family, SPNet, FPN) detection approaches. They differ by either having a separate region-proposal stage or by directly predicting the bounding boxes. Two-stage detection approaches are better at detecting irregularly shaped objects, while one-stage detectors prioritize inference speed and parameter count. As it is very rare for teeth to have outliers in shape, single-stage detectors are the right choice for forensic odontology.

This thesis focuses on the newest iteration of the YOLO architecture (YOLO v5 [160]) and its variants. YOLO v5 has four variants in regards to capacity, which are called "small," "medium," "large," and "extra large." The scaling is controlled by two parameters, the width and depth multiple. These parameters define how the compound scaling is applied to the network. The compound scaling method used was proposed by the EfficientDet paper [146], which in detail explains the heuristic used. In addition to compound scaling, a series of so-called "P6" models were evaluated. The baseline variant of YOLO v5 outputs three layers (called P3, P4, and P5) from which the final detection inference is made. The "P6" variant has an additional output layer, P6, which is also taken into consideration. The baseline architecture uses Spatial Pyramid Pooling [161] and Cross Stage Partial Network modules [162] to reduce computational time and parameter count while maintaining the same detection performance. The main feature of Spatial Pyramid Pooling is the removal of the fixed-size constraint of neural networks, which enables a convolutional neural network not to require an input image of a specific size. The Cross Stage Partial Network module employs a split-and-merge strategy to allow for more gradient flow through the network. The baseline architecture of YOLO v5 can be seen in Figure 5.17.

The models are trained using Stochastic Gradient Descent (SGD), with the learning rate being regulated by the One Cycle Learning Rate schedule [163]. The loss is a compound loss consisting of three components. The first component is the bounding box regression loss, which is realized as the mean squared error. This component corrects the output of the raw bounding box prediction. The second component is the objectness loss, which is realized with binary

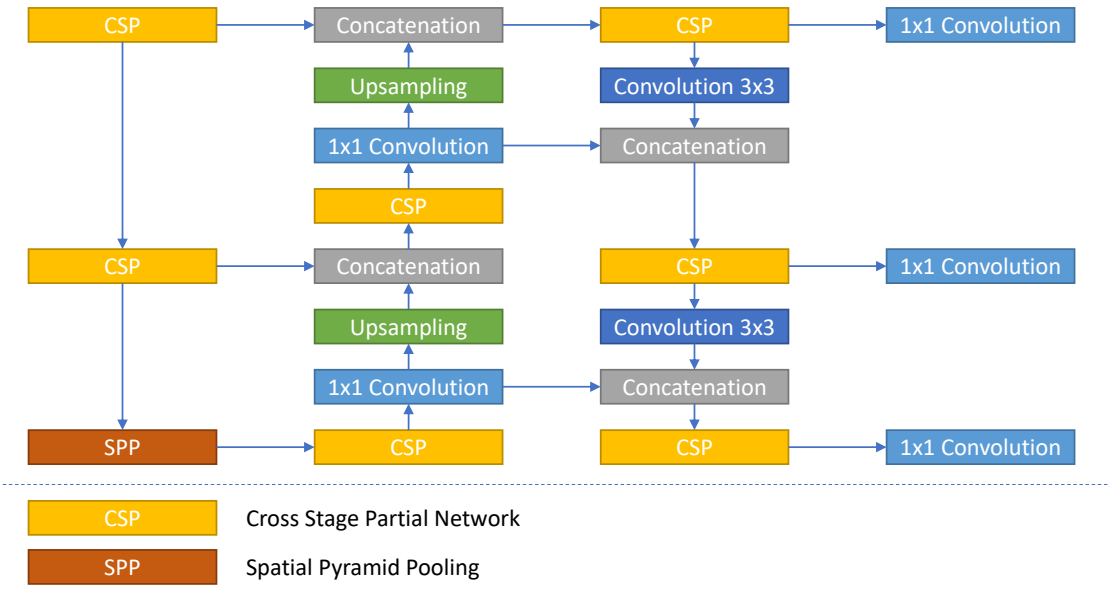


**Figure 5.17:** The architecture of the YOLO v5 detection model. The architecture uses multiple CSP modules and a high degree of skip connections and concatenations to allow for easier gradient calculations.

cross-entropy, with the goal of determining if the image within the predicted bounding box is an object or just the background. The third component is the classification loss, which is realized as categorical cross-entropy. Classification loss is used for the final classification; in this case, the determination of the type of a detected tooth.

Multiple variants of the data have been evaluated to determine the best hyperparameters for detection. In addition to the different model configurations, models were trained and evaluated on four input image sizes: 256, 640, 1280, and 1820. Detection models can detect and classify objects. To that end, five classification approaches have been tested. The baseline approach is a pure detection of the tooth with no additional classes. The additional variants classify the type of the tooth during detection. All four classification approaches are tested, with teeth being classified into either 4, 8, 16, or 32 classes as described in Section 5.3.4. Compared to the method proposed in Section 5.3.4, the difference in this approach is the presence of more context. An image of an individual tooth contains only the tooth and some surrounding tissue, while the classification during object detection can incorporate a wider context in its prediction.

To properly evaluate object detection systems, four metrics are used. Those metrics are intersection over union, precision, recall, and the mean average precision (mAP). Similar to intersection over union (IoU) described in Section 5.4 and shown in Figure 5.16, IoU for object detection describes the overlap between the ground truth and the predicted region, with those regions being bounding boxes. The value of IoU is between 0 (no overlap) and 1 (total overlap), and higher values describe better detection results.

Precision and recall are metrics often used in classification. Precision is the ratio of correctly positively classified samples (true positives) and all samples that have been positively classified

(true and false positives). It describes the correctness of positive classifications - if the model positively classifies a sample, how likely is it to be correct? On the other hand, recall describes how many samples are correctly positively classified. The recall is the ratio of positively classified samples (true positives) and all samples that should be classified as positive (true positives and false negatives). Formally, precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{5.15}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{5.16}$$

While the difference seems subtle, the metrics describe very different properties. In the context of object detection, recall tells us how many objects were found out of all objects that should have been found, and precision tells us how likely it is that a detection is actually an object.

The mean average precision (mAP) is the most commonly used metric for object detection, as it best encapsulates the performance of a detection model [164]. It is calculated as the mean AP for all classes for a certain IoU threshold. Average precision is calculated as the area under the precision-recall curve. As precision and recall are in opposition (the more prediction are correct, the more predictions will be missed, and vice-versa), every precision value will have a corresponding recall value. The precision and recall are usually calculated at a certain IoU threshold, and the precision-recall pairs are calculated cumulatively sorted by model confidence. This combination of cumulative (also sometimes called "progressive" in literature) precision-recall calculation and sorting by model confidence effectively represents the calculation of precision and recall at all confidence thresholds.

# Chapter 6

# Results

## 6.1 Age estimation

### 6.1.1 Direct regression for panoramic dental x-ray images

Estimation of age with direct regression for panoramic dental x-ray images is the baseline approach to age estimation. As described in Chapter 5, Section 5.2, and Section 5.3.1, the results presented here are from the fine-tuned model discovered with a combination of grid and random search. Mean and median absolute errors are used as the evaluation metrics, and the error is measured in years. The dataset and train/test split is described in Chapter 3 , Section 3.3.1. The results are analyzed per age group, anatomical region, data sampling and handling, and by overall performance.

The hyperparameter search produced models with mean absolute errors ranging from 5.67 years to 20 years. Surprisingly, models with the attention mechanism have consistently underperformed compared to their non-attention counterparts. The best-performing model has the following hyperparameters: VGG16 as the feature extractor, with 40 channels in the final convolutional layers, 128 units in the second to last fully-connected layer, and no attention mechanism. A visualization of the predictions of the best model can be seen in Figure 6.1.

The training data was prepared in three different ways. The baseline approach, with no modifications to the images and no influence on the sampling, achieves an overall mean absolute error of 4.06 years and a median error of 3.11 years. The best performance is achieved in the age group of 20 to 25 years, with a matching mean and median error of 2.82 years. The worst performance is for the age group of 65 to 70 years, when the mean error is 11.78 years and the median error is 13.75 years. To account for the bias towards younger samples, stratification per age was evaluated. The results improve on average, achieving a mean error of 4.02 years, but getting slightly worse in the median case, with the median error increasing from 3.11 years to 3.17 years. In this case, the best performance shifted to the age group of 30 to 35 years, with a

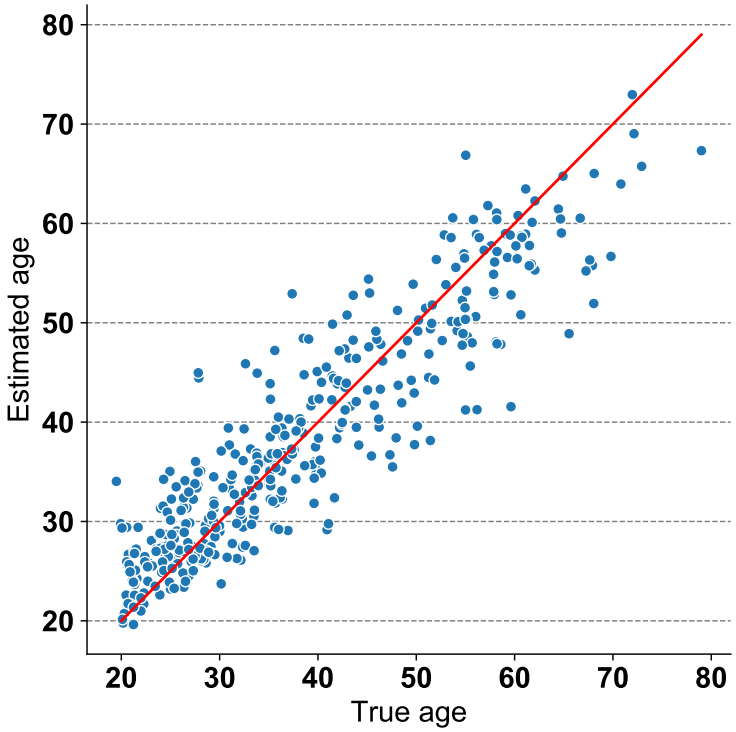


**Figure 6.1:** Model age estimations and their true values. Each dot is a sample from the dataset, the x-axis represents the true age of a sample, and the y-axis represents the age the model estimated for that sample. The diagonal line is the line of perfect estimations. As can be seen, younger samples are closer to the diagonal line, while older samples have a higher degree of variance. It can also be seen that the model does not systematically over- or underestimate age.

mean error of 3.29 years and a median error of 2.61 years. The best median performance stays in the age group of 20 to 25 years, with an error of 2.32 years. When augmentation is added, the overall mean absolute error decreases from 4.06 to 3.96 years, and the median absolute error decreases to 2.95 years. The best performing age group is again between the age of 25 and 30, with a mean absolute error of 2.97 years and a median absolute error of 2.77 years. The best median performance is for the age group between 25 and 30 years, where the median absolute error reached just 2.12 years. Overall, the best-performing model uses augmentation and no stratification. An overview of the results in terms of absolute mean and median error for every age group for all three approaches can be seen in Table 6.1. A visualization of those results can be seen in Figure 6.2.

In addition to the baseline models, the architecture described in Section 4.2 was used to construct a model for age estimation. The search resulted in a model of 367 129 learnable parameters, which is just 2.3% of the learnable parameters of the model based on VGG16, which had 16 238 185 learnable parameters. The hyperparameters for this model are: 2 BiFPN layers, a BiFPN feature size of 16, 128 units in the second-to-last fully-connected layer, an image size of 256 by 256 px, a feature compressor depth of 1, and a convolutional depth starting

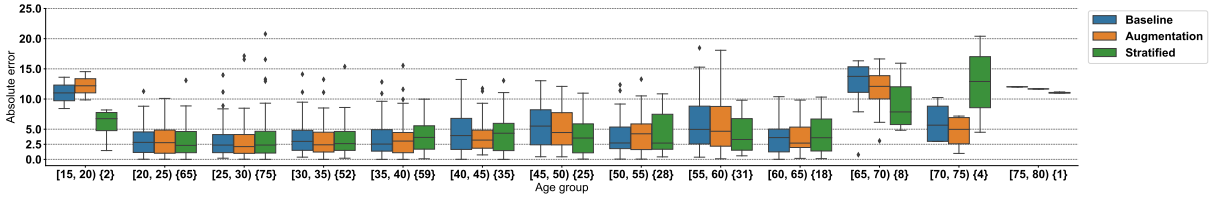


**Figure 6.2:** Overall results of age estimation per age group and data-variant approach for the model based on a state-of-the-art feature extractor and attention. As can be seen, younger samples have a lower overall median absolute error and lower variance. The first age group is a statistical anomaly due to the low sample count.

**Table 6.1:** The results of the age estimation per data-variant approach and age group for the model based on a state-of-the-art feature extractor and attention. Two errors are shown, the mean absolute error as $\mu$ and the median absolute error as $\hat{y}$.

| Age group | Baseline | | Augmentation | | Stratified sampling | |
|---|---|---|---|---|---|---|
| | $\mu$ | $\hat{y}$ | $\mu$ | $\hat{y}$ | $\mu$ | $\hat{y}$ |
| [15, 20) | 10.37 | 11.01 | 11.80 | 12.19 | 6.45 | 6.74 |
| [20, 25) | 2.82 | 2.82 | 2.97 | 2.77 | 3.58 | 2.32 |
| [25, 30) | 2.97 | 2.37 | 3.02 | 2.12 | 3.50 | 2.38 |
| [30, 35) | 3.72 | 2.99 | 3.27 | 2.42 | 3.29 | 2.61 |
| [35, 40) | 3.67 | 2.53 | 3.63 | 3.03 | 3.80 | 3.64 |
| [40, 45) | 4.65 | 3.94 | 4.47 | 3.20 | 4.45 | 4.33 |
| [45, 50) | 5.62 | 5.52 | 5.54 | 4.45 | 3.76 | 3.54 |
| [50, 55) | 4.07 | 2.72 | 4.25 | 4.24 | 4.24 | 2.70 |
| [55, 60) | 6.03 | 4.97 | 5.76 | 4.66 | 4.16 | 3.30 |
| [60, 65) | 3.84 | 3.61 | 3.29 | 2.70 | 4.20 | 3.59 |
| [65, 70) | 11.78 | 13.75 | 11.25 | 12.10 | 9.08 | 7.86 |
| [70, 75) | 5.01 | 5.67 | 3.90 | 4.98 | 12.68 | 12.90 |
| [75, 80) | 11.85 | 12.02 | 12.97 | 11.68 | 11.03 | 11.01 |
| All ages | 4.06 | 3.11 | **3.96** | **2.95** | 4.02 | 3.17 |

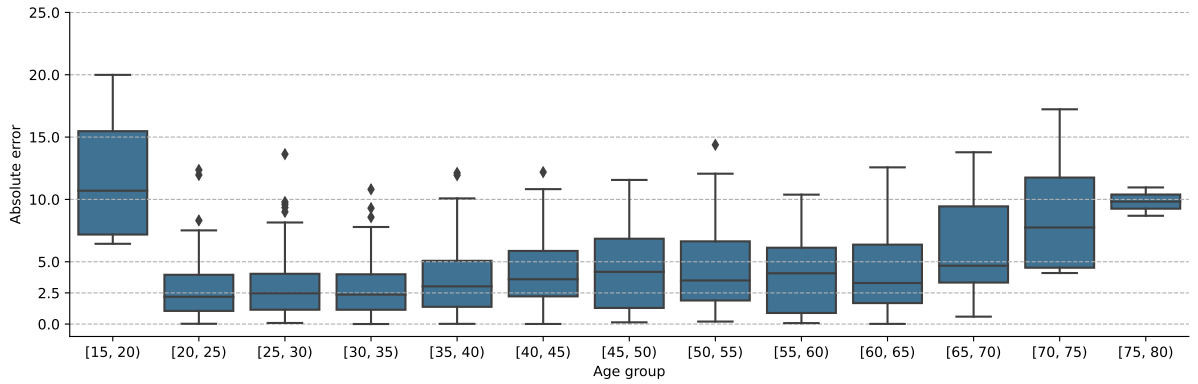


**Figure 6.3:** Overall results of age estimation per age group for the BiFPN-based model. Similarly to the previous model, younger samples perform better with a lower degree of variance. Compared to the previous model, the variance is across the board lower, even for higher age groups. As before, the first age group is a statistical anomaly due to the low sample count.

**Table 6.2:** The results of the age estimation per data-variant approach and age group for the BiFPN-based model. The mean absolute error is shown as $\mu$ and the median absolute error as $\hat{y}$.

| Age group | [15, 20) | [20, 25) | [25, 30) | [30, 35) | [35, 40) | [40, 45) | [45, 50) | [50, 55) | [55, 60) | [60, 65) | [65, 70) | [70, 75) | [75, 80) | All ages |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mu$ | 11.96 | 2.95 | 3.03 | 2.91 | 3.37 | 3.96 | 4.30 | 4.60 | 3.87 | 3.91 | 6.14 | 8.73 | 9.82 | 3.69 |
| $\hat{y}$ | 10.70 | 2.19 | 2.47 | 2.37 | 3.03 | 3.59 | 4.19 | 3.50 | 4.08 | 3.29 | 4.68 | 7.75 | 9.82 | 3.11 |

at 16 and finishing with 128 channels. The overall mean absolute error for this model is 3.69 years, which is 0.27 years better than the model based on a state-of-the-art architecture. The overall median absolute error is 3.11 years, which is slightly worse than the model based on a state-of-the-art architecture, the difference being 0.16 years. The best performing age group is between the years of 30 and 35, with a mean absolute error of 2.91 years and a median absolute error of 2.37 years. A full overview of the this models results per age group can be seen in Table 6.2, and a visualization of the results per age group can be found in Figure 6.3.

To determine which regions of interest, from an anatomical point of view, contribute the most to a correct age estimation, three additional variants are trained as explained in Section 5.3.1. The models are on images of only the tooth region, of only the surrounding tissue, and of only the surrounding tissue, but with teeth being covered individually instead of blocking the entire tooth region. For models trained on only the tooth region, an overall mean absolute error of 4.65 years is achieved, which is a 0.7 years difference from the best full OPG model, while the overall median absolute error is 3.62 years. The best performing age group is between the ages of 25 to 30, where a mean absolute error of 3.04 years and a median absolute error of 2.19 years is achieved. Models trained on the surrounding tissue achieve an overall mean absolute error of 5.06 years and an overall median absolute error of 4.12 years. This is a 1.1-year increase in mean absolute error compared to the best full OPG model. The best performing age group is slightly older for this case, with a mean absolute error of 3.26 years being achieved for samples between the age of 35 and 40. If teeth are precisely covered by their bounding boxes instead

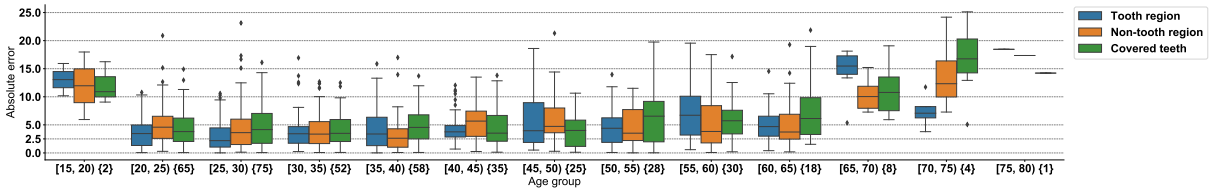


**Figure 6.4:** Overall results of age estimation per age group and anatomically important regions of interest for the model based on a state-of-the-art feature extractor and attention. The performance trends for age are similar to before, just with higher variance due to a decrease in available information for the model. As can be seen, the region with teeth present performs best.

of a general cover of the tooth region, the overall mean absolute error increases even higher, reaching the value of 5.42 years - a full 1.46 years increase compared to the best performing full OPG model. The best performing age group is the same as for the model with the covered tooth region, with the samples between the ages of 30 and 35 achieving the smallest error, which is still a significantly higher 4.19 years. The median error for the best-performing age group is 3.49 years. The overview of all results in regards to the region of interest models per age group and image variant can be seen in Table 6.3, and a visualization of the results can be seen in Figure 6.4.

Regarding computational performance, all models based on state-of-the-art architectures and attention described in Section 4.1 perform inference of 100 images in 1.65 seconds, regardless of sampling approach, augmentation, or region of interest take the same time. On the other hand, models trained on the minimized architecture described in Section 4.2 perform inference of 100 images in 0.12 seconds, an order of magnitude faster. This performance is measured on an Nvidia 2080 Ti GPU. Age estimation with current manual measurement methods takes 20 to 30 minutes per image.

Attention has been shown to overall improve deep learning model performance. However, extensive testing has shown that the models equivalent (up to the attention module) perform worse for this specific use case. Attention helps with filtering of unnecessary information from the input. Learned features can be shared across different classes, which can be detrimental to the differentiation between some class pairs or sets. With attention, parts of an image are effectively blocked out, and thus the negative influence of the distracting signal is diminished. Panoramic dental x-ray images are aligned, as a person's head is fixated during scanning. Thus, images do not have the distracting variance that attention combats. Interestingly, when models with attention worked well, their attention module converged to a state to always return a feature map with all values equal to or very close to 1, effectively disabling the attention module.

The results clearly indicate a trends of decreased performance with increased age, as can be observed in Figure 6.2, Figure 6.4, Table 6.1, and Table 6.3. This stems mainly from two factors. The first factor, as shown in Chapter 3, is that the available data is slightly biased towards younger samples. This inevitably leads to the model inheriting that bias, as the optimization

**Table 6.3:** The results of the age estimation per age group and anatomically important regions of interest for the model based on a state-of-the-art feature extractor and attention. Two errors are shown, the mean absolute error as $\mu$ and the median absolute error as $\hat{y}$.

| Age groups | Teeth region | | Surrounding tissue | | Precisely covered teeth | |
|---|---|---|---|---|---|---|
| | $\mu$ | $\hat{y}$ | $\mu$ | $\hat{y}$ | $\mu$ | $\hat{y}$ |
| [15, 20) | 12.04 | 13.06 | 11.80 | 11.96 | 11.31 | 10.91 |
| [20, 25) | 3.67 | 3.45 | 5.17 | 4.59 | 3.56 | 3.80 |
| [25, 30) | 3.04 | 2.19 | 4.61 | 3.61 | 5.22 | 4.15 |
| [30, 35) | 4.20 | 3.41 | 4.44 | 3.35 | 4.19 | 3.49 |
| [35, 40) | 4.43 | 3.20 | 3.26 | 2.73 | 5.14 | 4.55 |
| [40, 45) | 4.84 | 3.76 | 5.39 | 5.67 | 5.30 | 3.52 |
| [45, 50) | 5.72 | 3.96 | 6.26 | 4.73 | 4.54 | 4.01 |
| [50, 55) | 4.68 | 4.40 | 4.58 | 3.52 | 6.30 | 6.55 |
| [55, 60) | 7.17 | 6.81 | 5.68 | 3.83 | 5.92 | 5.73 |
| [60, 65) | 4.52 | 4.69 | 5.51 | 3.73 | 7.56 | 6.13 |
| [65, 70) | 13.57 | 15.48 | 10.02 | 10.03 | 10.61 | 10.78 |
| [70, 75) | 6.80 | 7.07 | 14.12 | 12.33 | 15.38 | 16.76 |
| [75, 80) | 20.18 | 18.45 | 16.90 | 17.35 | 18.05 | 19.86 |
| All ages | **4.65** | **3.62** | 5.06 | 4.12 | 5.42 | 4.40 |

process will produce a model that will underestimate the age in case of uncertainty, as this will produce overall better results. The stratified sampling approach fixes that issue. Stratified sampling produced a model that overall achieves worse performance, but the performance per age group is more stable. Additionally, with more data being available in younger age groups, better discriminative features can be found, and thus better results can be achieved in younger age groups. The other factors are tooth changes, be they natural or external. With age, teeth stop their development, and calcification and decay become the driving force behind dental structure changes [69]. This naturally leads to less discriminative features and information being contained in a tooth, as there is less difference between more mature age groups. As people age, teeth get damaged, and the damage and alterations to fix that damage accumulate more and more throughout a lifetime. As those alterations remove part of the tooth and replace it with artificial material, any information that the original tissue contained is gone. Therefore, older samples, which usually have more alterations, consequently contain less age-related information. The latter's effect can be seen in the model based on stratified training, as older age groups still have a decreased performance trend with rising age, albeit a less pronounced trend than in the baseline and augmentation cases.

Classical age estimation methods are primarily based on tooth parameters. As the deep learning model processes the entire panoramic dental x-ray image, it is necessary to determine what part of the image contributes to the estimation. From the results shown in Table 6.3 and Figure 6.4, it can be seen that the tooth region produces the best performing model, while the surrounding tissue and precisely covered teeth perform worse. The performance difference between the roughly and precisely covered teeth can be explained by outliers. While the tooth region covers the region where teeth are visible, not all samples have well-aligned teeth of the same size. In other words, some images have edges of teeth or roots visible, which the model can exploit to improve its estimation. The overall results give insight into the importance of the different anatomical regions of interest. The models that include teeth perform better, which indicates that teeth contain either easier accessible or more information about age. The surrounding tissue can also be used for age estimation, albeit with worse performance. More importantly, given that models that combine both dental and skeletal features perform better, it can be concluded that the skeletal and dental tissue both contain information expressed in different, independent features. Another interesting observation that can be seen by comparing Table 6.1 and Table 6.3 is that by removing teeth the best performing age group increases to the years of 35 to 40. This indicates that age-related changes are more pronounced in the surrounding skeletal tissue than in individual teeth.

A problem with stratification by duplication is the potential for the model to overfit to the few samples, be it in just the age bracket or in general. In this case, the model produces overall worse results, but the performance in older age groups was better. As the train and test set

are fully separated, and the test set is never used for any research decisions, this clearly indicates that the model did not overfit to the train samples but managed to find better features for older samples. The larger sample count in younger age groups impacts the training process by suppressing the features that help identify older samples, as their presence lowers overall performance. Training time augmentation tackles this problem differently. By modifying images during training time, the model is never trained on the two identical samples more than once. These virtual samples increase the "effective sample count" for every age group, thereby diminishing the effect of the low sample count in higher age groups. This naturally is not a perfect substitute, as the lack of diversity in the dataset cannot be artificially added without real data. Thus the diminishing performance with increasing age is still present, but the overall results are better than the baseline case.

The Unnamed architecture described in Section 4.2 was designed with speed and efficiency in mind. Current state-of-the-art models compete on the ImageNet benchmark [165], which has 1000 classes. These models are designed to be high-capacity, with speed and efficiency being secondary objectives. ImageNet is also a generalist dataset, with natural images of a wide variety of objects. In contrast, forensic odontology estimates comparatively very few values, and those estimations are done on x-ray images of the human dental system. The input domain is significantly more limited, which leads to less capacity being needed to process the image. This is ultimately reflected in the results. The age estimation models produce a better mean absolute result by a margin of 0.27 years while only needing 2.3% of the number of learnable parameters. While the average case was improved, the median absolute error rose by 0.16 years. This difference in performance between the architectures indicates that the minimized model over- or underestimates less.

The age estimation methods currently in use are most often based one of three fundamental studies: Kvaal et al. [75], Drusini et al. [76], or Cameriere et al [78]. These studies use groups of tooth measurements to estimate a person's age. They work with individual teeth which need to be healthy, without any alterations. The developed models estimate a person's age from the entire panoramic dental x-ray image, and they do not need all teeth to be in perfect condition. Additionally, those studies have at least an order of magnitude fewer samples due to the difficulties of manual measurements.

Kvaal et al. [75] uses the teeth up-1, up-2, up-5, down-2, down-3, and down-4 to perform age estimation. The reported estimation errors are between 9.5 and 11.5 years, depending on which tooth is used for estimation. The data collection consists of 100 radiographs of dental patients, and only healthy and unaltered teeth are included in the study. Drusini et al. [76] approach age estimation in a similar way by measuring tooth parameters. The study analyzes the tooth-coronal index ([77]), which is used to estimate the age of an adult sample. The estimation error ranges from 5.88 years to 6.66 years on their dataset of 433 samples. Cameriere et al. [78]

focuses on the parameters of the maxillary canine, and they achieve a median absolute error of 3.7 years. This method constructs a linear model based on some parameters of the maxillary canine. The method has proven itself successful, but the parameters of the linear model need to be adjusted for different populations. This method opened the gate to the creation of many variants based on different populations [79, 80, 81, 82, 83, 84, 85], all of which perform on a similar level of accuracy. These three methods are the pillars of age estimation in modern forensic odontology and are the most commonly used classical methods. This thesis achieves an overall mean absolute error of 3.69 years, which outperforms all current state-of-the-art methods in forensic odontology while being fully automated and easily reproducible.

On the deep learning front, three studies stand out, Vila-Blanco et al. [89], Guo et al. [90], and Kim et al. [91]. These approaches use deep learning-based methods to construct a model that estimates a person's age from panoramic dental x-ray images. Vila-Blanco et al. [89] designed their own simple two-branch feed-forward convolutional deep learning architecture, which consists of 6 consecutive 3x3 convolution and batch normalization blocks separated by 2x2 max pooling. Each branch estimates either age or sex, and both estimations are done simultaneously. Some skip connections are added between the branches. They report a mean absolute error of 2.84 years. However, their dataset includes child samples, which are inherently easier to estimate due to the presence of developmental markers. In contrast, the youngest sample in their dataset is 4.5 years, whereas the youngest dataset used in this thesis is 19 years. The average age of their sample is between 17.1 years and 25.5 years, while the average of the dataset used in this thesis is 38.17 years. Their dataset has 2289 samples, compared to the 4035 used in this research. A direct comparison per age group is not possible, as they calculate the error from the youngest sample up to a certain age instead of calculating it for every age group independently.

All previous methods are formulated as regression problems. The other two studies for age estimation based on deep learning approach the problem as a classification problem. Guo et al. [90] are more focused on the estimation of age in younger samples. Their dataset consists of 10257 samples, ranging from the ages of 5 to 25 years, with a slight bias towards younger samples for males and a slight bias towards older samples for females. The overall sample count is slightly skewed towards female samples. Their model architecture is based on EfficientNet [153] and SE-ResNet [166]. They achieve an accuracy between 92.3% and 95.9% (depending on the age threshold used). Due to the significant difference in the sample age (child samples compared to our adult samples), a direct comparison is not possible. Lastly, [91] too tackles age estimation as a classification problem. Their dataset consists of panoramic dental x-ray images of 2025 patients of all ages. They test two approaches, where they divide their samples into either three or five categories. The three-category approach divides samples into the age groups of 0 to 19 years, 20 to 49 years, and over 50-year-old samples. The five-category approach

divides samples into the age groups of 0 to 19 years, 20 to 29 years, 30 to 39 years, 40 to 49 years, and over 50-year-old samples. Image patches containing the teeth "16", "26", "36", and "46" are extracted, and a ResNet-152 is trained. They report an accuracy of between 89.05% and 90.27% and an AUC between 0.94 and 0.98. Again, as the problem is tackled as a classification problem, where samples are grouped into broad age group classes, a direct comparison is not possible. However, our predictions can be converted to classification labels to facilitate a direct comparison. In that case, the 3-class variant achieves an accuracy of 92.62%, and the 5-class variant achieves an accuracy of 70.91%. This difference is based on two factors. The first factor is the difference in datasets. Kim et al. [91] have most samples with ages between 0 to 29 years (77% of their entire test set), which skews their results. The other factors are due to the training and model. The model proposed in the thesis is optimized in such a way that a difference between 49 and 50 years is penalized as a one-year error, while in the framework Kim et al. propose penalizes the difference between 49 and 50 years as a full miss, while the difference between 40 and 49 is not penalized at all.

### 6.1.2 Estimation as a probability distribution for panoramic dental x-ray images

Estimating the age as a probability distribution improves estimation performance and gives insight into which samples are harder for the model to estimate. The estimation is given as a discrete probability distribution, from which the numerical value of the estimation is calculated as the expected value, and the confidence is calculated as the variance. The data used does not differ from the data used for direct regression for age estimation of panoramic dental x-ray images, with the description of the data being given in Section 3.3.1. The best-performing model uses the combination of state-of-the-art feature extractors with optional attention architecture described in Section 4.1. The performance is evaluated using the mean and median absolute error of the expected value. This analysis examines the impact of the introduced trade-off hyperparameters on the estimated distributions and model performance, performance trends across age groups, as well as the behavior and properties of the variance.

The best performing model has the following hyperparameter values: VGG16 as the feature extractor, a depth of the final convolutional feature map of 40 channels, no attention, a fully-connected layer of size 128, and the loss trade-off parameters of value $\lambda_1 = 0.8889$ and $\lambda_2 = 0.04121$. The overall mean absolute error is 3.60 years, and the overall median absolute error is 2.76 years. The average standard deviation is 4.85 years, and the median standard deviation is 4.61 years. The best performance is achieved in the age group between the ages of 20 and 25, with a mean absolute error and standard deviation of 2.09 years and 2.96 years, respectively, as well as a median absolute error and median standard deviation of 1.71 years and 2.85 years

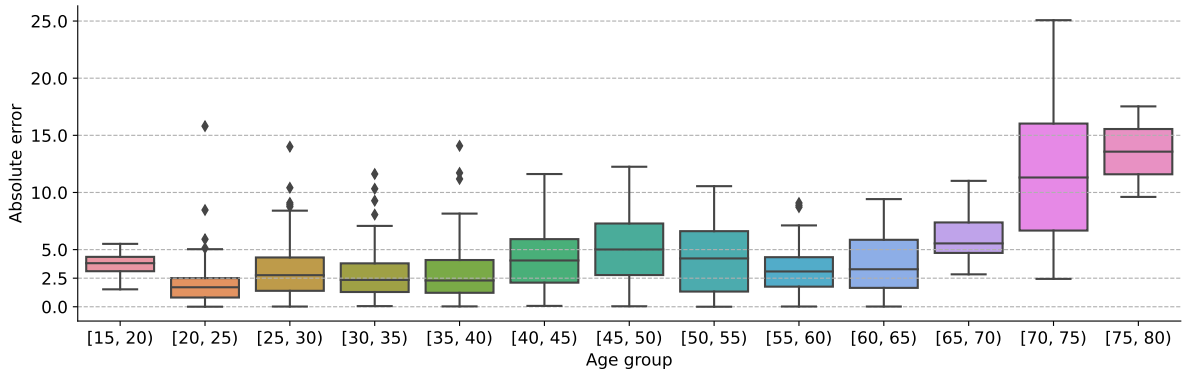


**Figure 6.5:** Overall results for age estimation with a probability distribution. As can be seen, the median absolute error across all age groups is lower compared to the direct regression approach, as is the variance of the error. Additionally, The absolute error does not consistently increase with age, as was the case with the direct regression approach.

respectively. A detailed overview of the estimation performance per age group can be seen in Table 6.4, and a visualization of the results is given in Figure 6.5.

The computational performance of the model is comparable to the direct regression approach. Both models use the model based on state-of-the-art feature extractors and attention, with VGG16 as their base. In regards to computational time, the VGG16 segment is the most significant. The model consists of 14M parameters and can estimate the age in the form of a probability distribution of 100 images in 1.2 seconds. This performance is measured on an Nvidia 2080 Ti GPU. To reiterate, age estimation with current manual measurement methods takes 20 to 30 minutes per image.

Similar to the phenomena observed in training age estimation models with direct regression, the attention variant of a model performed worse than its non-attention counterparts. The explanation for this is similar to the previous case. Panoramic dental x-ray images are very similar in their structure due to the imaging process. Attention improves performance by diminishing distracting signals from the input, but as the images are very similar, and as all information from the image is useful for age estimation, attention tends to worsen the results for this use case.

The estimation performance per age group follows a similar trend to the direct regression approach, albeit outperforming the direct regression approach in most younger age groups. Performance on younger samples is better, while the variance of the estimated probability distribution increases with age. The model performs worse in older age groups, similar to the direct regression model. It is important to note that this approach achieves better performance on younger samples than the direct regression approach but worse performance on older samples. For almost all age groups up to the age of 60, the performance is better. Age groups [25, 30) and [50, 55) have slightly worse performance (0.14 and 0.07 years, respectively). However, the estimation performance for samples up to 60 years is 1.41 years better. For the samples of more than 60 years, performance is decreased by 0.9925 years.

**Table 6.4:** The mean and median errors of the estimated age, as well as the mean and median standard deviation of the estimated distribution. Both are shown per age group.

| Age group | Absolute error | | Standard deviation | |
|---|---|---|---|---|
| | Mean | Median | Mean | Median |
| [15, 20) | 3.66 | 3.81 | 2.58 | 2.54 |
| [20, 25) | 2.09 | 1.71 | 2.96 | 2.85 |
| [25, 30) | 3.16 | 2.76 | 3.99 | 3.87 |
| [30, 35) | 2.86 | 2.35 | 4.43 | 4.26 |
| [35, 40) | 3.09 | 2.30 | 4.78 | 4.59 |
| [40, 45) | 4.27 | 4.05 | 5.39 | 5.30 |
| [45, 50) | 5.10 | 5.01 | 6.00 | 5.80 |
| [50, 55) | 4.32 | 4.23 | 6.46 | 6.12 |
| [55, 60) | 3.44 | 3.09 | 7.11 | 6.58 |
| [60, 65) | 3.85 | 3.28 | 6.49 | 6.25 |
| [65, 70) | 6.03 | 5.54 | 7.01 | 6.73 |
| [70, 75) | 11.93 | 11.31 | 5.77 | 6.14 |
| [75, 80) | 13.57 | 13.57 | 4.13 | 4.13 |
| **Overall** | **3.60** | **2.76** | **4.85** | **4.61** |

The trade-off hyperparameters influence not only the final performance of the deep learning-based model but also the shape of the estimated probability distribution. Specifically, mainly $\lambda_1$ influences the performance, while $\lambda_2$ restricts the spread of the probability distribution. As $\lambda_2$ increases, the estimation becomes more concentrated - in other words, the variance is reduced. However, if $\lambda_2$ is too large, the model performance decreases as the model does not have the flexibility to trade correctness for uncertainty. After extensive searching, results show that $\lambda_1$ needs to be an order of magnitude larger than $\lambda_2$ for best performance. A visualization of the performance difference for the trade-off hyperparameters $\lambda_1$ and $\lambda_2$ is shown in Figure 6.6. As can be seen in the upper figure, the mean absolute error decreases with the value of $\lambda_1$ up to the region around $\lambda_1 = 0.9$, increasing afterward. In the lower figure, the opposite can be seen, with the mean absolute error decreasing as $\lambda_2$ decreases but producing significantly worse results for $\lambda_2 = 0$.

With the estimated variance (and thus standard deviation), an interval of $\mu \pm \sigma$ can be constructed and examined in relation to the correct age value. For 71.43% of tested samples, the true age was within this interval. If this is extended to the variance in the form of $\mu \pm \sigma^2$, the ra-

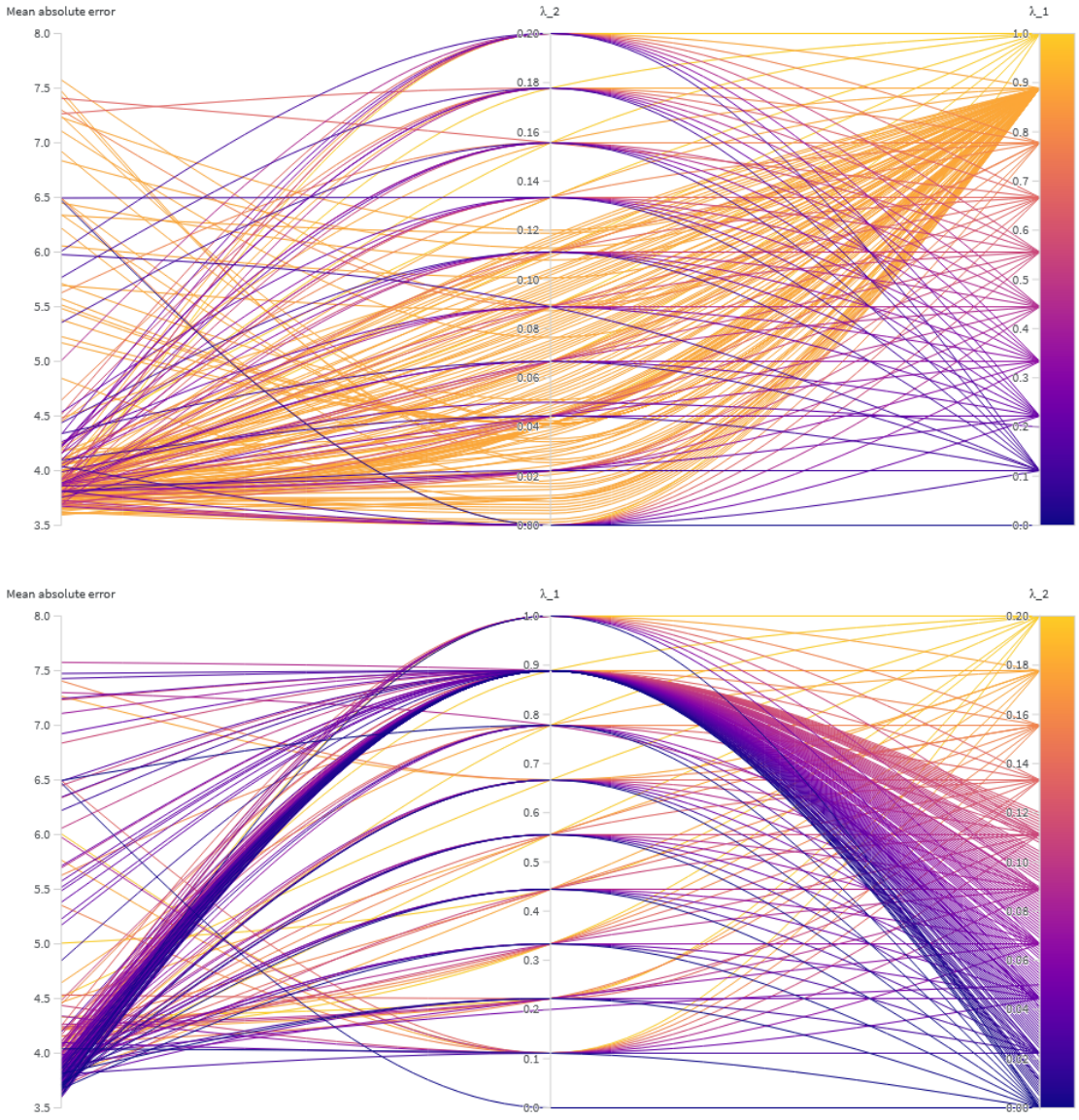


**Figure 6.6:** The effect of the $\lambda$ values on the mean absolute error. The mean absolute error is the leftmost axis. The upper figure is colored according to the value of $\lambda_1$, and the lower by the value of $\lambda_2$.

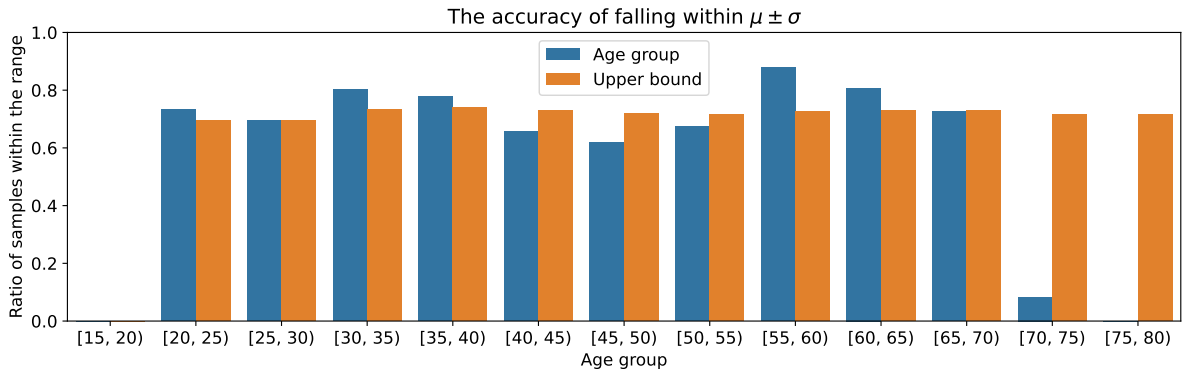


**Figure 6.7:** The accuracy of a sample being within $\mu \pm \sigma$ of the predicted probability distribution. The blue bars represent only samples within a specific age group, while the orange bars represent the cumulative results up to the upper bound of the age group.

tio of successfully estimated interval rises to 99.01%. This holds true across age groups. There are two approaches to this analysis. One approach looks at the success rate strictly within an age group, while the other approach is cumulative and defined only by an upper bound for age. The first approach allows insight into the change of the metric per age group, while the other approach shows the change of the trend with rising age. As shown in Figure 6.7, the accuracy is in both cases around the 71% mark, with the per-age-group approach having a slightly larger variance. Nonetheless, the results can be considered consistent across age groups.

While the estimated variance is a measure of uncertainty, it does not strongly correlate with the absolute error. The correlation between the variance and absolute error is 0.2612, and the correlation between the standard deviation and absolute error is 0.2851. The variance and standard deviation have the strongest correlation with the sample's age, which is 0.4983 and 0.6286, respectively. Most natural processes are either finished by a certain age or they slow down with age. Therefore, the number of changes and detectable differences diminish, leaving less and less discriminative information between age groups. For example, the difference between a 70 and a 75-year-old sample is minor compared to the differences in any previous 5-year period. Interestingly, when the variance is normalized by the age of the sample, it is near-constant across all age groups. This means that samples that are hard to estimate for the model are uniformly distributed across all age groups. Ultimately, the variance is a measure of model uncertainty, but it cannot be used to make meaningful conclusions about the accuracy of the estimation.

As the described results are for estimation of age from panoramic dental x-ray images, the comparable studies are the same to the direct regression approach [75, 76, 78, 89, 90, 91]. A detailed explanation each approach is given in Section 6.1.1. Classical methods range in sample size between 100 and 433, while this approach is evaluated on 4035 panoramic dental x-ray images. The estimation errors, depending on available measurements, varies between 5.88 to 11.5 years for Kvaal et al. [75] and Drusini et al. [76]. Cameriere et al. stand out with a median

absolute error of 3.7 years. The proposed deep learning-based approach achieves an overall median absolute error of 2.76 years, which outperforms the classical methods. In regard to other deep learning-based approaches, Vila-Blanco et al. [89] are the only non-classification-based estimation method. As mentioned previously, their dataset includes child samples, thereby reducing their overall mean absolute error. Other approaches classify samples into age groups. Guo et al. [90] focus on child and young adult samples, achieving an accuracy of 92.3% to 95.9%. A direct comparison is not viable due to the significant difference between target age groups. On the other hand, Kim et al. [91] classify samples into either three or five categories, either in the age groups of 0 to 19 years, 20 to 49 years, and over 50-year old samples, or into the age groups of 0 to 19 years, 20 to 29 years, 30 to 39 years, 40 to 49 years, and over 50-year old samples, reporting an accuracy between 89.05% and 90.27%. Converting the results of the proposed deep learning-based method into this system, it achieves an accuracy of 94.09% for the three-class approach and 73.07% for the five-class approach. As the proposed deep-learning approach was not explicitly trained for classification, and due to the difference in the distribution of ages in the mentioned study and this thesis, results differ slightly, performing worse on the five-class variant and performing better on the three-class variant.

### 6.1.3 Direct regression for individual dental x-ray images

Teeth hold information that can be used to estimate a person's age. When all teeth and the surrounding tissue are taken into account, an accurate estimation can be done, as shown in the previous Section. Estimating the age from a single, individual tooth is less precise and reliable. However, due to a variety of circumstances that can lead to the death of a subject, sometimes only individual teeth are available for analysis. The dataset used for these experiments contains 86495 images, as described in Chapter 3, Section 3.3.2, which includes the subset of samples annotated with tooth alterations. As with age estimation from panoramic dental x-ray images, the metrics to evaluate the age are the mean and median absolute errors. This section gives an overview of the overall results, as well as an analysis of the trends per age group, an analysis of performance per tooth type, the impact of alterations on the estimation, as well as an overview of the multi-task variant of the model, an explanation about the underperforming of attention, a comparison to current state-of-the-art methods, and a comparison between estimation models specialized per tooth type and a generalized model.

The best performing model has the following hyperparameters: the feature extractor is VGG16, a final feature map depth of 662 channels, no attention mechanism, and a fully-connected layer of 1931 units. As with the model for panoramic dental x-ray images, the models with attention underperformed compared to their non-attention counterparts. The multi-task variant and the variant with additional demographic information as input did not outperform or match the performance of the presented model. Therefore their results are not presented in this

**Table 6.5:** The results of the age estimation per age group all individual tooth samples. The mean absolute error is shown as $\mu$ and the median absolute error as $\hat{y}$.

| **Age group** | **[20, 25)** | **[25, 30)** | **[30, 35)** | **[35, 40)** | **[40, 45)** | **[45, 50)** | **[50, 55)** | **[55, 60)** | **Overall** |
|---|---|---|---|---|---|---|---|---|---|
| $\mu$ | 8.04 | 6.11 | 5.78 | 5.76 | 6.17 | 7.87 | 8.22 | 10.40 | 6.55 |
| $\hat{y}$ | 6.95 | 4.63 | 4.55 | 4.97 | 5.06 | 6.84 | 7.74 | 10.45 | 5.32 |

**Table 6.6:** The results of the age estimation per age group individual tooth samples with no dental alterations. The mean absolute error is shown as $\mu$ and the median absolute error as $\hat{y}$.

| **Age group** | **[20, 25)** | **[25, 30)** | **[30, 35)** | **[35, 40)** | **[40, 45)** | **[45, 50)** | **[50, 55)** | **[55, 60)** | **Overall** |
|---|---|---|---|---|---|---|---|---|---|
| $\mu$ | 7.32 | 5.27 | 5.31 | 5.61 | 6.53 | 9.29 | 10.14 | 12.37 | 6.15 |
| $\hat{y}$ | 6.38 | 4.03 | 4.24 | 5.02 | 5.57 | 9.28 | 9.58 | 13.04 | 4.94 |

thesis. The model achieves an overall mean absolute error of 6.55 years and an overall median absolute error of 5.32 years for all samples. For samples with no alterations, the overall mean absolute error decreases to 6.15 years, and the overall median absolute error decreases to 4.94 years. For samples with alterations, the best performance is achieved for the samples in the age group between 25 and 30 years, while the best performance for all samples is achieved for those in the age group between 35 and 40 years. An overview of the mean and median absolute errors per age group for all samples can be seen in Table 6.5, and an overview of the mean and median absolute errors per age group for samples with no alterations can be seen in Table 6.6.

In regard to computational performance, the model is capable of processing 100 images in 0.8 seconds. This performance is measured on an Nvidia 2080 Ti GPU. As with panoramic dental x-ray images, the time to estimate the age with current manual measurement methods is measured in minutes per image.

As teeth come in different shapes and sizes, the performance per tooth type varies. However, shape and size are not the only factors. Different teeth experience different wear and tear based on their position and construction. Some of the most durable teeth are canines, which are deeply embedded in the jaw bone and rarely break. On the other hand, incisors tend to be more often damaged, be it mechanically or by decay, due to their central position and relative thinness compared to other teeth. On the other hand, premolars and molars tend to succumb to decay, which results in either heavy alterations like root canals, dental fillings, bridges, crowns, and in extreme cases, complete extraction and removal. The age estimation model achieves the best performance on the molars, specifically the second molar. This tooth is labeled as "X7" throughout the thesis. Given all second molars, the best performing is the upper left molar, with an average of 0.3 years less in mean absolute error. A visualization of these trends of mean absolute errors per tooth type can be seen in Figure 6.8.

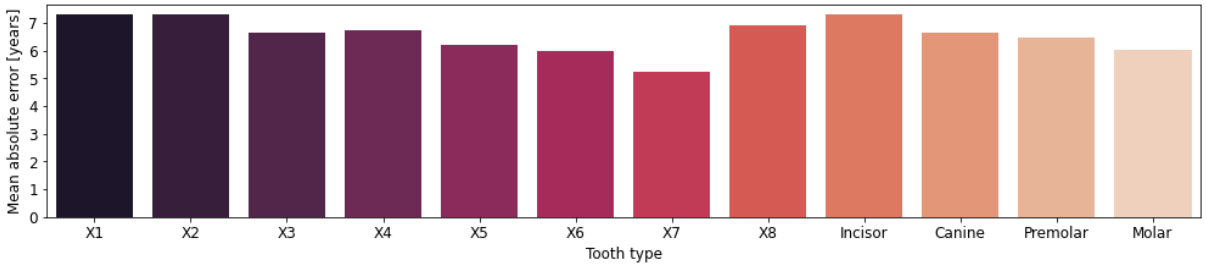


**Figure 6.8:** Age estimation performance results per tooth type. The first eight bars classify teeth into eight types according to their position within a quadrant, and the last four bars classify teeth into the four basic tooth types.

This ranking changes when only healthy teeth are taken into account. For teeth with no alterations, molars still outperform other teeth, but the best performance is achieved by the first molar, which is closely followed by the performance of the second molar. The presence of alterations impacts the average estimation performance. Alterations physically remove part of the natural tooth structures and substitute those parts with replacements that contain no information about the person. In fact, alterations are strongly radio-opaque, resulting in fully white areas in the resulting radiographic image. How much they impact the performance of the model depends on their size and their position. Dental fillings vary in size, depending on the decay and damage that happened to a tooth. However, fillings do not alter the shape or structure of the root. Crowns significantly alter a tooth's shape but leave the entire root untouched. Root canal fillings, on the other hand, change the structure of the root by filling the root canal with radio-opaque material. This in and of itself is not a significant change to the tooth structure; however, to get to the root, it is necessary to drill through the crown of the tooth. In turn, root canal fillings always come paired with either a dental filling or crown, which ultimately results in a significant change to the tooth structure. Bridges and implants are full replacements of a tooth and therefore carry no information specific to the person. Given all this, it naturally follows that tooth alterations decrease the average estimation performance. Missing teeth can only use the surrounding tissue to determine the age and thus perform the worst out of all scenarios, with an average increase in the mean absolute error of 2.6 years. Root canal fillings change the tooth structure significantly, leading to a mean absolute error increase of 0.9 years. Dental fillings and tooth decay are common and small, and thus the model is able to handle those scenarios well, with an average increase of 0.13 years of mean absolute error for tooth decay. Interestingly, the mean absolute error is slightly lower when fillings are present in older samples. This is due to their difference in occurrence with age. Younger samples tend to have less decay, less damage, and thus fewer dental fillings. This means that fillings are more present in older samples, which is a fact the model exploits to resolve uncertain cases. When comparing estimation error between perfect and imperfect tooth samples per age group, it can be noticed that with increasing age, the model better estimates older samples if they have alterations. A visualization of this

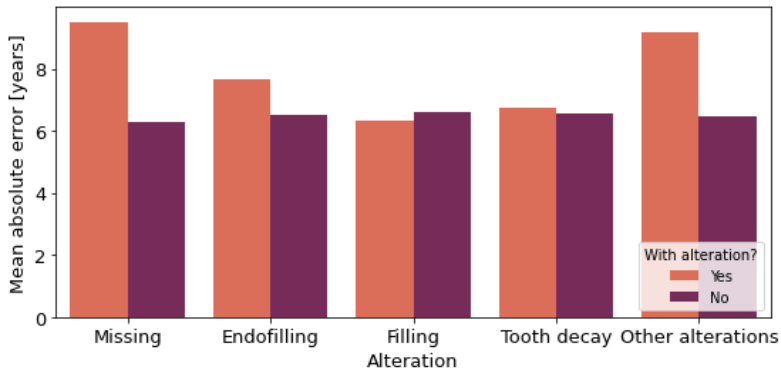


**Figure 6.9:** Age estimation performance per dental alterations. The light-colored bar shows the samples that include the dental alteration, while the darker bar shows results for samples without the alteration.

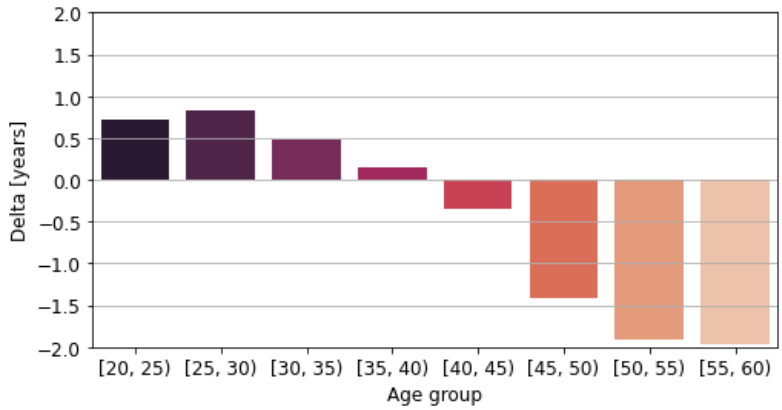


**Figure 6.10:** Difference of mean absolute errors per age group between samples with and without dental alterations. As can be seen, the presence of alterations increases the error in younger samples. However, the model uses dental alteration information to estimate older samples' age correctly.

phenomena can be seen in Figure 6.10. Older samples naturally have more damage and more alterations. Thus the model uses this fact, among others, to estimate the age more accurately. If an older sample has a perfectly healthy tooth, this is effectively an outlier, and thus the model performs worse. A visualization of the impact of alterations on the mean absolute error can be seen in Figure 6.9. A detailed breakdown of the model estimation performance per tooth type and alteration can be seen in Table 6.7 and Table 6.8 for the mean absolute error and for the 4, 8, 16, and 32 tooth-type classification systems, and in Table 6.9 and Table 6.10 for the median absolute error across all four tooth type classification systems.

The multi-task variant and variant with additional demographic input information did not perform better. The multi-task models, models that estimate multiple demographic information simultaneously, did not manage to exploit shared features to improve or even reach equal performance to task-specialized models. The multi-task models have been tested with branching at various points, yet no matter where the branching occurred, the best performance was achieved when only one branch was successful while the others produced estimations with a significant overall error. Effectively, the best multi-task models are models that collapsed into

**Table 6.7:** Overview of the mean absolute error, measured in years, per dental alteration and tooth type (4-, 8-, and 16-class systems). 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 |
| Overall | 6.32 | 7.66 | 9.51 | 6.73 | 9.18 | 6.15 | 6.55 |

**Table 6.8:** Overview of the mean absolute error, measured in years, per dental alteration and tooth type (32-class systems). 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 |
|---|---|---|---|---|---|---|---|
| 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 6.9:** Overview of the median absolute error, measured in years, per dental alteration and tooth type (4-, 8-, and 16-class systems). 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 |
| Overall | 5.12 | 6.24 | 9.01 | 5.73 | 7.99 | 4.94 | 5.32 |

**Table 6.10:** Overview of the median absolute error, measured in years, per dental alteration and tooth type (32-class systems). 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 |
|---|---|---|---|---|---|---|---|
| 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 |

task-specialized models. The models with added demographic input information performed either equally or worse while requiring more input information. Hyperparameter searching did not improve any of those models. It is, therefore, necessary to conclude that those configurations are not successful for the tasks tackled in this thesis.

Similar to panoramic dental x-ray images, the same studies are fundamental in forensic odontology [75, 76, 78]. As discussed in the previous Section, they estimate age from healthy, unaltered teeth with no decay, dental fillings, root canal fillings, or any other type of addition or subtraction of dental tissue. They are methods that require manual measurements and are therefore time intensive. Due to this time requirement, these studies and most studies that derive from them have a relatively small sample size (in the hundreds, less than a thousand). The approach in this thesis is based on a dataset of over 80 000 samples, and imperfect teeth and their impact on model predictions are analyzed. Kvaal et al. [75] is a linear regression model based on different tooth dimension ratios that can be used on the measurements of a single tooth or multiple teeth. As mentioned, depending on the available teeth, the standard error varies between 8.6 to 11.5 years. Drusini et al. [76] model estimate age from molars and premolars using the coronal pulp cavity index [77] and achieves a standard error between 5.88 and 6.66 years. The model presented in this thesis achieves a standard error of 7.95 years on unaltered teeth and 8.51 years on all samples. This is an improvement compared to Kvaal et al. [75] and comparable to Drusini et al. [76] while simultaneously being fully automated, reproducible, and able to process teeth with alterations. Additionally, Drusini et al. [76] focus only on premolars and molars, whereas the model presented in this thesis can process any tooth. If only the proposed models' results on premolars and molars are analyzed, the difference in performance disappears. Cameriere et al. [78] established a linear regression model too, and report their results using the median absolute error instead the standard error of estimate. The model is derived from the 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). This approach achieves a median absolute error of 3.7 years, while the model presented in this thesis achieves an overall median error of 4.94 years on unaltered teeth and 5.32 years on all samples. The difference might stem from the significantly larger sample size (100 samples in Cameriere et al. [78] compared to over 80 000 samples used in this thesis), from the inclusion of all tooth types, and the inclusion of all teeth regardless of their status. As for deep learning-based methods, current literature does not provide any studies in that regard except for the ones published during the research for this thesis. Ultimately, the demonstrated model achieves either a similar or better result than current state-of-the-art methods while being fully automated, reproducible, and multiple orders of magnitude faster.

## 6.2 Sex assessment

### 6.2.1 Panoramic dental x-ray images

Assessing the sex of a person is another fundamental task in forensic odontology. As with age estimation, current methods for sex assessment in forensic odontology are based on manual measurement. The proposed deep learning-based method for sex assessment of panoramic dental x-ray images is fully automated. The image data used to train these sex assessment models is the same as the data used to train the models for age estimation. Section 3.3.1 gives a detailed overview of the data. As described in Chapter 5, a combination of grid and random search is used to determine the best model. The evaluated models are the same, and a detailed overview of their architecture and properties is given in Chapter 4. As described in Section 5.3.3, accuracy is used as the metric to evaluate the efficacy of the proposed method. An analysis of the performance trend per age group is performed, as well as a comparison between the performance of the model with a state-of-the-art feature extractor with attention and the BiFPN-based model, and an interpretability analysis is done using XGradCAM [167], a modern variant of GradCAM [35].

The model with a state-of-the-art based feature extractor with attention achieves an overall accuracy of 96.87%. The hyperparameters for this model are: VGG16 as the feature extractor, a convolutional depth of 256 channels, no attention, and a fully-connected layer of size 128. This model consists of 14M parameters. The best performance is achieved in the age group between 20 and 30 years, which achieved an accuracy of 100%. Achieving 100% is usually an indicator for a model that has overfitted to a dataset with data leakage between the train and test set or an indicator of a small sample size. This result was obtained on the subset of the test set images with an age between 20 and 30 years, which consists of 193 samples. Those images are not present in the train set, and no images of the same person can be in both the train and test set. The BiFPN-based model achieves an accuracy of 97.04%. The hyperparameters for this model are: 2 BiFPN layers with a feature size of 16, convolutions starting with 42 channels and linearly scaling to 128 channels, a feature compressor depth of 1, and an image size of 256. While the performance is slightly better than the previously described model, and despite the huge difference in parameter count, the performance trends per age group behave similarly. A detailed overview per age group and model can be seen in Table 6.11.

The model based on a state-of-the-art feature extractor can assess the sex of 100 images in 1.2 seconds, while the BiFPN-based model can process the same amount of images in 0.11 seconds. Both methods are measured on an Nvidia 2080 Ti GPU. Like age estimation, classical forensic odontology takes between 20 and 30 minutes per panoramic dental x-ray image.

The performance of both models decreases with increasing age. As people age, their dental system accumulates damages and changes, thereby removing natural structures and replacing

**Table 6.11:** Sex assessment results for panoramic dental x-ray images for both model architectures per age group. Samples over the age of 70 are grouped together due to the low sample count.

| Model | [20, 30) | [30, 40) | [40, 50) | [50, 60) | [60, 70) | 70+ | Overall |
|---|---|---|---|---|---|---|---|
| Attention-based model | 100% | 97.9% | 97.1% | 96.9% | 96.1% | 100% | 96.90% |
| BiFPN-based model | 100% | 98.1% | 97.3% | 97.0% | 96.3% | 100% | 97.04% |

them with artificial materials. Those artificial replacement materials do not replicate the internal structure of the tooth, and thus no sex-related information can be extracted from them. In a similar vein, tooth decay breaks the structure of a tooth. It slowly erodes the tooth, again removing the ability to extract relevant information about the sexual dimorphism of the sample. The performance sharply increases in the last age group. However, this is due to a small sample size in that age group. As shown in Section 3.3.1, the available data is biased towards younger samples, thus having fewer samples available in older age groups. Therefore, this sharp performance increase is just a statistical anomaly, which would disappear with more data.

As with age estimation, the model variants with attention underperform compared to their non-attention counterparts. Attention variants that performed well at-best matched the performance of their non-attention variant. The attention variant models that matched the performance converged to a state where their attention module always returned all values very close or equal to 1, effectively disabling any effect the attention module could have. The reason for this behavior of models with attention is the same as for age estimation, which is explained in detail in Section 6.1.1.

As this approach to sex assessment is formulated as a classification problem, model interpretability methods can be applied to determine which parts of the image contribute to the model prediction. Understating how and why the deep learning-based model works is as important as the model's performance, especially in medical image analysis. To determine which regions have a significant impact on the model decision, class activation maps are used. The most successful class activation map is GradCAM [35], with XGradCAM being an incremental improvement to the method. XGradCAM, in addition to calculating class activation maps based on model gradients, normalizes those gradients by the activations, giving a clearer picture of what is happening within the neural network. Figure 6.11 shows the XGradCAM saliency maps for both the female and male classes, as well as a panoramic dental x-ray image with annotated anatomical regions that are relevant to sex assessment. A saliency map is calculated for every sample in the test set, which are averaged together to produce a single saliency map for each class. For female samples, the most significant regions are the angle of the mandible and mandibular condyle. The articular eminence of the temporomandibular joint and the styloid process contribute too, but to a lesser degree. Males samples are less impacted by those

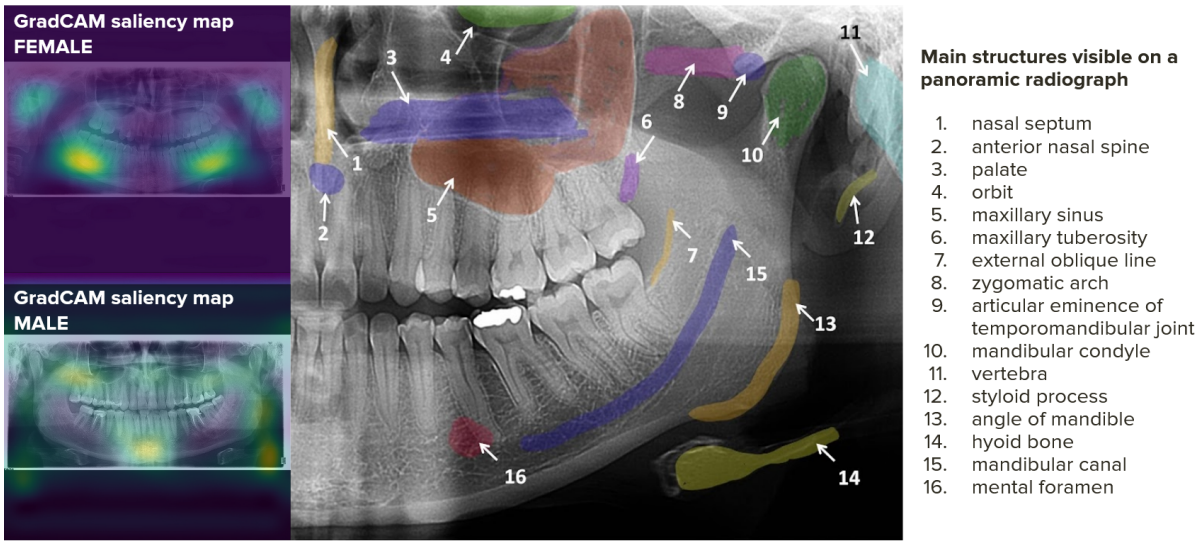


**Figure 6.11:** Anatomical regions of interest for sex assessment visible on a panoramic dental x-ray image and the XGradCAM saliency maps for model interpretability. As can be seen, the regions that the model focuses on for sex estimation are anatomically important for sex assessment.

anatomical elements. Instead, male samples are mostly identified by their chin, angle of the mandible, and maxillary tuberosity. While not strictly part of the dental system, the nasal septum and anterior nasal spine contribute to the assessment of the male category, as well as the hyoid bone, albeit to a lesser extent.

Multiple methods attempt to solve this problem in classical forensic odontology, all of which rely on manual measurements [58, 102, 103, 113]. As those methods require manual measurements, their sample sizes a comparatively smaller than the data used in this thesis, with sample sizes ranging from 100 to 419 samples. Steyn et al. [102], one of the earlier studies, focuses on 12 standard cranial and five mandibular measurements from 91 samples. They used SPSS discriminant function analysis and determined that bizygomatic breadth was the single most indicative measure for the entire skull and that bigonial breadth was the most indicative measurement for the mandible. They also determined that multiple of those 17 measurements can be combined to achieve better accuracy, which ranges from 80% to 86%. Franklin et al. [103] use basic univariate statistics and discriminant function analysis of 10 mandibular parameters to assess the sex of a person. They performed their analysis on 40 samples and determined that the dimensions of the ramus are most indicative of the sex. Depending on the parameters and model used, the results range from 72.5% to 95%. Badran et al. [58] is a newer study that focuses on the mandibular ramus flexure. This study has one of the largest datasets for sex assessment in forensic odontology, consisting of 419 samples. They showed that the mandibular ramus flexure could be used to assess the sex of a person with an overall accuracy of 70.9%. They concluded that the ramus flexure should be considered a supplementary rather than definitive source of information for sex assessment. Deana et al. [113] examines the useability of nonmetrical morphological characteristics of the mandible, like the shape of the chin, the divergence of the

**Table 6.12:** Overview and comparison of the performance of sex assessment methods in literature.

| **Paper** | **Year** | **Sample size** | **Accuracy** |
|---|---|---|---|
| Giles et al. [101] | 1964 | 265 | 85.0% |
| Steyn et al. [102] | 1998 | 91 | 80.0%-86.0% |
| Franklin et al. [103] | 2006 | 40 | 95.0% |
| Dayal et al. [105] | 2008 | 120 | 75.8% |
| Saini et al. [96] | 2011 | 116 | 80.2% |
| Indira et al. [106] | 2012 | 100 | 76.0% |
| Marinescu et al. [107] | 2013 | 200 | 84.0% |
| Badran et al. [58] | 2015 | 419 | 70.9% |
| Bhagwatkar et al. [108] | 2016 | 100 | 76.0%-87.0% |
| Deana et al. [113] | 2017 | 128 | 75.0%-95.2% |
| Maloth et al. [110] | 2017 | 100 | 74.0% |
| Nagaraj et al. [111] | 2017 | 100 | 71.0% |
| Alias et al. [112] | 2018 | 79 | 78.5% |
| Vila-Blanco et al. [118] | 2020 | 3400 | 84.0% |
| Ke et al. [119] | 2020 | 19976 | 94.60% |
| **Proposed approach** | **2019, 2022** | **4035** | **96.9% / 97.0%** |

gonial angle, the profile of the chin, the contour of the base of the mandible, shape of the ramus, and others. Their dataset consists of 128 samples. Seven indicators were analyzed, and their proposed method achieves an accuracy between 75% and 95.2%. Those studies highlight the most popular and successful approaches in forensic odontology, but many more studies examine similar methods across different parameters and population groups, all achieving similar results with slightly different models [96, 105, 106, 107, 108, 110, 111, 112]. It is important to note that all studies discard samples with any kind of damage, be they different rooted in pathologies or any other source. A detailed comparison of the most significant studies in the field of sex assessment in forensic odontology and this thesis is shown in Table 6.12. Regarding deep learning, some progress has been made. Vila-Blanco et al. [118] propose a network that assesses the sex of a person. Their dataset consists of 3400 panoramic dental x-ray images, with 50.32% of the dataset being younger than 20 years old. The proposed model is a simple feed-forward network that estimates age and sex simultaneously. They achieve an overall accuracy of 84%, and their method is fully automated. The method proposed in this thesis provides two models which achieve an accuracy of 96.90% and 97.04%. Ke et al. [119] proposed a model based on multiple feature fusion. Their dataset consists of 19 976 panoramic dental x-ray images, and the dataset is biased towards female samples (61.14%) and toward younger samples (34.57% of the samples are between the ages of 20 and 30). They achieve an accuracy of 94.6%, and their model is fully automated. To conclude, the approach and models proposed in this thesis are fully automated, the dataset is sufficiently large compared to other studies, and they achieve higher performance than all current methods.

### 6.2.2 Individual dental x-ray images

Sex assessment from individual tooth x-ray images is a harder problem, as there is less sexual dimorphism expressed in teeth compared to the surrounding structures. Research shows that sex assessment accuracy from individual teeth might be limited to around 80% [115]. The data used is the same that has been used for age estimation from individual dental x-ray images, and a detailed description of the dataset is given in Section 3.3.2. The dataset consists of 86495 images, which is one of the most extensive datasets of individual teeth in literature, and it includes additional information about tooth alterations. As with sex assessment from panoramic dental x-ray images, accuracy is used as the evaluation metric. The performance and performance trends per age group, tooth type, and tooth alterations are analyzed, and the differences between the model based on state-of-the-art feature extractors with attention and the BiFPN-based model are examined. Finally, XGradCAM [167] is used to determine which regions of the tooth have the highest impact on the model's prediction.

The model with a state-of-the-art based feature extractor achieves an overall accuracy of 75.44% and an accuracy of 76.41% on teeth with no alterations. This model has the following

**Table 6.13:** Results of sex assessment performance for individual tooth x-ray images per age group and per presence of dental alterations for the model based on a state-of-the-art feature extractor and attention.

| Tooth status | [20, 25) | [25, 30) | [30, 35) | [35, 40) | [40, 45) | [45, 50) | [50, 55) | [55, 60) | Overall |
|---|---|---|---|---|---|---|---|---|---|
| Unaltered | 82.09% | 72.54% | 77.89% | 74.49% | 76.52% | 87.64% | 76.77% | 60.78% | 76.41% |
| All | 81.64% | 72.08% | 77.95% | 72.65% | 74.12% | 85.71% | 74.90% | 64.57% | 75.44% |

**Table 6.14:** Results of sex assessment performance for individual tooth x-ray images per age group and per presence of dental alterations for the BiFPN-based model.

| Tooth status | [20, 25) | [25, 30) | [30, 35) | [35, 40) | [40, 45) | [45, 50) | [50, 55) | [55, 60) | Overall |
|---|---|---|---|---|---|---|---|---|---|
| Unaltered | 84.62% | 72.14% | 76.27% | 73.75% | 70.87% | 91.01% | 77.78% | 56.86% | 75.78% |
| All | 82.62% | 70.60% | 75.57% | 71.75% | 69.56% | 88.31% | 69.80% | 56.69% | 73.64% |

hyperparameter values: VGG16 as the feature extractor architecture, a depth of the final feature map of 40 channels, no attention, and a fully-connected layer of 128 units. The best performance is achieved for samples between the ages of 45 and 50 years, with an overall accuracy of 85.71% and an accuracy of 87.64% for teeth without alterations. On the other hand, the BiFPN-based model achieves an overall accuracy of 73.64%, and an accuracy of 75.78% on teeth with no alterations, while only having a number of parameters equal to only 1.2% of the state-of-the-art based feature extractor with attention model. Similar to that model, the BiFPN-based model achieves the best performance for the samples between ages 45 and 50 years, reaching an overall accuracy of 88.31% and an accuracy of 91.01% on unaltered teeth. The hyperparameters for the BiFPN-based model are: 2 BiFPN layers with a feature size of 16, convolutions starting with 8 channels and linearly scaling to 42 channels, a feature compressor depth of 1, and an image size of 256. Despite those differences, the performance trends per age group, tooth type, and impact of alterations are equal between both models. A detailed overview of sex assessment performance per age group for individual dental x-ray images is shown in Table 6.13 for the larger model, and in Table 6.14 for the smaller model.

The model with a state-of-the-art based feature extractor with attention processes 100 images in 1.1 seconds and consists of 14M learnable parameters. The BiFPN-based model consists of 300k parameters and processes 100 images in 0.11 seconds. This performance is measured on an Nvidia 2080 Ti GPU. Sex assessment with manual methods, like all previously mentioned classical forensic odontology methods, is measured in minutes per image.

Table 6.13 and Table 6.14 show the results per age group of larger and smaller model, respectively. Compared to the performance shown for sex assessment from panoramic dental x-ray images, these results do not show a clear trend per age. While the larger model performs slightly better than the smaller model, both models perform significantly better on unal-

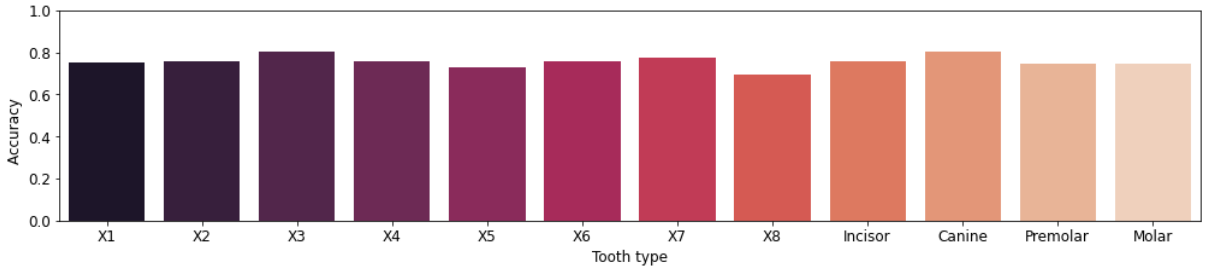


**Figure 6.12:** The accuracy of sex assessment per tooth type. The first eight bars classify teeth into eight types according to their position within a quadrant, and the last four bars classify teeth into the four basic tooth types.

tered teeth. The smaller model is defined by only 2.5% of the amount of learnable parameters compared to the larger model. While the overall performance is worse, their performance on unaltered teeth is acceptably close, especially when taking into account the large reduction of computational requirements. The difference between the performance on all samples and only samples with no alterations is more pronounced for the smaller model. This indicates that the smaller model has successfully learned to identify and measure dental structures expressing sexual dimorphism but did not fully grasp the wide variety of structural disruptions when alterations are present. Additionally, as has been seen in the models trained for age estimation, the model variants with attention did not produce better results. The reasoning for this behaviour is explained in Section 6.1.1.

Sex assessment performance is not consistent across tooth types. As described in Section 5.3.4, teeth can be classified in four different ways. The difference in performance is best observed with the 4-class and 8-class classification approach. As shown in Figure 6.12, the performance is best for canines, followed by incisors, while the performance for premolars and molars is equal. Looking at the individual tooth position, it can be seen that, while there is some minor variance between teeth in the same 4-class group, they follow the presented underlying trend. Equally, when observing the performance difference using other classification systems, the trend persists - there is minor variance, but the most significant difference is between teeth not in the same 4-class group.

Alterations impact sex assessment performance significantly more than tooth type. Overall, tooth alterations worsen performance. Alterations that remove more of the natural tooth structure or significantly alter the shape result in a higher reduction of accuracy. Missing teeth are an interesting case. The dataset has teeth labeled as missing - bounding boxes where a tooth should be if it had not been removed. The performance is, as anticipated, much worse, but it much better than random chance. This indicates that the model is not only taking the tooth into account but also the surrounding tissue. Overall, root canal fillings have the highest impact, as they not only alter the root of the tooth but also require an additional dental filling for the path used to get to the root. The impact of tooth decay has a high variance, as tooth decay

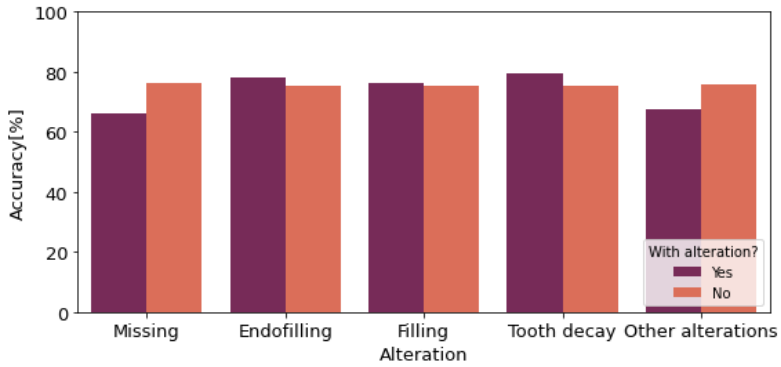


**Figure 6.13:** Sex assessment performance per dental alterations. The light-colored bar shows the samples that include the dental alteration, while the darker bar shows results for samples without the alteration.

can be minor and is often quickly removed but can also progress significantly, with advanced stages of tooth decay bringing extraordinary damage and change to the tooth. A visualization of the impact tooth alterations have on sex assessment performance can be seen in Figure 6.13. Additionally, a detailed overview of performance per tooth alterations and per the 4, 8, and 16 class tooth type classification approach can be seen in Table 6.15, and for the 32 class tooth type classification system in Table 6.16.

Lastly, it is crucial to determine which parts of the tooth the model takes into account for sex assessment. For all images in the test set, the saliency map for all images of the same tooth position (tooth type in the 8-class system) has been calculated using the XGradCAM method. The saliency maps shown in Figure 6.11 are the average saliency map per tooth type, the upper two rows for the female class and the bottom figure for the male class. As can be seen, the most focus is on the crown of the tooth for both classes. This focus is more concentrated for incisors and canines, while premolars and especially molars tend to include the middle section of the tooth as well. For male samples, the focus is more divided between the crown of the tooth and its root. Interestingly, for mandibular teeth and the male class, the model focuses significantly more on the tooth's root than maxillary teeth. As the mandibular parameters can be used for sex assessment, those might subtly influence the tissue around the tooth and the tooth itself, resulting in this behavior.

Sex assessment from individual teeth has been studied in forensic odontology, with different approaches reaching high accuracy scores. Nevertheless, teeth are a limited source of information, and therefore sex cannot be assessed with 100% accuracy from just a single tooth. Research shows that, given an analysis of the literature, there seems to be an upper limit of 80% on single tooth sex assessment methods [115]. Karaman uses diagonal teeth measurements to predict the sex in the Turkish population. The dataset consists of 60 samples (30 male and 30 female), from which the diagonal measurements have been taken. Those measurements were analyzed with discriminant function statistics, and it was determined that the most indica-

**Table 6.15:** Overview of the accuracy of sex assessment, per dental alteration and tooth type (4-, 8-, and 16-class systems). 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% |
| Overall | 76.32% | 78.31% | 66.07% | 79.67% | 67.36% | 76.41% | 75.44% |

**Table 6.16:** Overview of the accuracy of sex assessment, per dental alteration and tooth type (32-class systems). 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 |
|---|---|---|---|---|---|---|---|
| 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% |

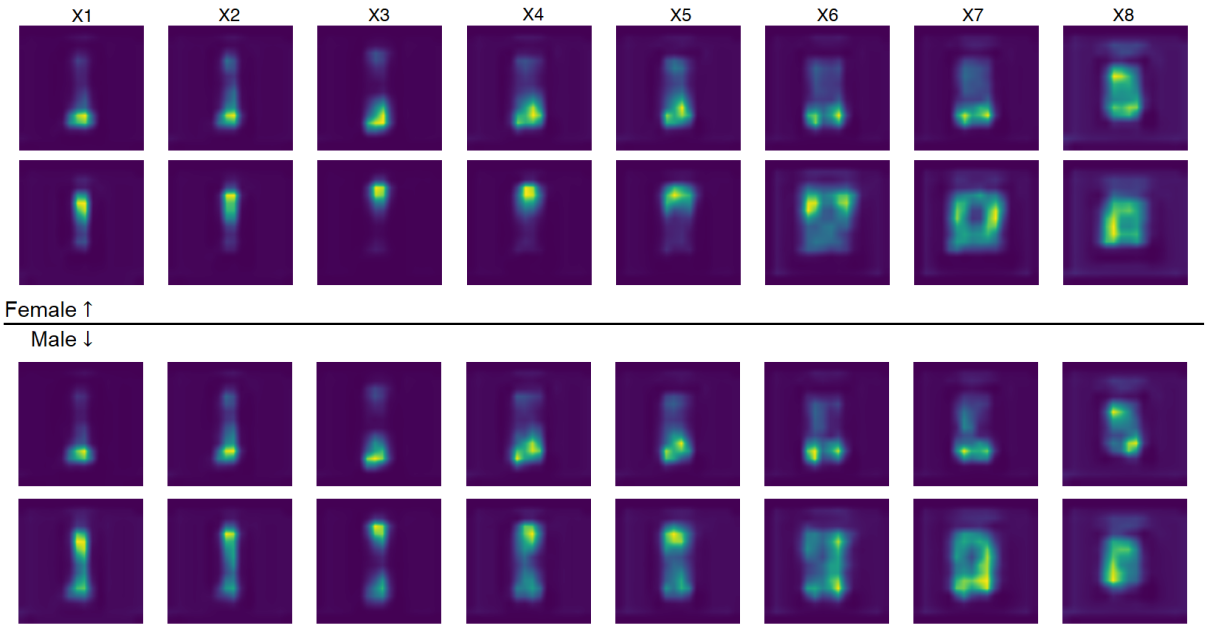


**Figure 6.14:** XGradCAM saliency maps for classification model interpretability. The upper two rows show the saliency map for the female class, and the lower two rows show the saliency for the male class. Each pair of rows represents the average saliency map for the tooth type according to the 16-type classification system, with the upper row of the pair representing the maxillary teeth and the lower mandibular teeth.

tive measurements are: the upper first incisor mesiobuccal-distolingual (MBDL) and distobuccal–mesiolingual diameters, lower second incisor MBDL diameter, and lower canine MBDL diameter. They achieve an overall accuracy of 83.3% and determine that canines significantly express sexual dimorphism. The method proposed in this thesis produces a lower accuracy of 76.41%, but this accuracy is verified on a dataset of 86495 samples instead of 60. Interestingly, the analysis of the proposed deep learning-based model approach also concludes that canines are especially suited for sex assessment. Capitaneanu et al. [168] did a multivariate analysis of the length- and width-related variables for all teeth, which is 212 variables in total. Those measurements are taken from 200 panoramic dental x-ray images, which are equally split between female and male samples. Principal-component analysis was applied, and they determined that an accuracy between 69.0% and 72.5% can be achieved with measurements of a single tooth. The upper end of their estimate is still lower than the results achieved by the proposed deep learning-based model, but if the sample size is taken into account, the results can be considered in-line. Neves et al. [169] performs sex assessment using the mesiodistal width data and verify their results on the Portuguese population. Their dataset consists of 168 samples, with 109 female and 59 male samples. The mesiodistal width was measured on all teeth from the right first molar to the left first molar. They determined that all measurements except for incisors are statistically significantly different. They fit a multivariate logistic regression model using the measurements from the upper left canine, the lower right lateral incisor, and the lower right

canine and achieved an accuracy of 75.0%. This is again in-line with the finding of this thesis, but this method can still not process teeth with alterations. Their accuracy is slightly lower but, again, verified on a dataset that is multiple orders of magnitude smaller. As for deep learning, there currently are no proposed methods for sex assessment from individual tooth x-ray images using deep learning-based methods or any computer vision-based method.

## 6.3 Segmentation of teeth and their alterations

### 6.3.1 Panoramic dental x-ray images

Segmentation is a different approach to using deep learning-based models compared to classification and regression, as the output is not just a vector of predictions but an entire map of predictions. Effectively, segmentation is a per-pixel classification approach used for the precise detection of regions of interest. As described in Section 5.4, the full panoramic dental x-ray images are segmented by alterations and teeth. In other words, the resulting segmentation maps show regions on which either a tooth or some tooth alterations are visible. While many alterations are annotated, the most common ones are tooth decay, crowns, fillings, and root canal fillings. A detailed overview of the data used for segmentation has been given in Section 3.3.3. Multiple models architecture and variants on those architectures were evaluated, as shown in Section 5.4. The primary metric used for the evaluation of model performance is the Sørensen–Dice coefficient. The Jaccard index and Sørensen–Dice coefficient show similar performance. The Jaccard index tends to penalize outliers more than the Sørensen–Dice coefficient. Therefore, the Sørensen–Dice coefficient gives better insight into the average performance and consistency of results. The accuracy score is misleading, as for panoramic dental x-ray images, most of the segmentation map is empty due to the surrounding skeletal structure being of no interest for this task. Therefore, the Sørensen–Dice coefficient is used as the primary metric to present the performance of the proposed approach. An analysis of performance per alteration is performed, as well as a detailed performance comparison between neural network architectures and their variants.

The model variants differ in the number of learnable parameters and thus the amount of compute needed to process an image. UNet and its variants follow the same baseline architecture, with the difference being the number of channels each processing stage produces. In a similar vein, FCN networks differ in the effective upscaling factor between the last feature map and the final segmentation map. For example, FCN8 upscales a feature map by a factor of 8, and FCN32 upscales a feature map by a factor of 32, which is a much more coarse approach. DeepLab v3 prescribes the usage of Atrous Spatial Pyramid Pooling for multi-scale processing but makes no prescriptions about the feature extractor network used. Detailed insight into the

model architectures and variants used is given in Section 5.4.

The best performance is achieved by the large variant of UNet. The Sørensen–Dice coefficient for teeth is 95.12%, and the most successful model for the segmentation of alterations is for fillings with a Sørensen–Dice coefficient of 77.11%. Root canal filling segmentation follows closely with a Sørensen–Dice coefficient of 74.31%, and the segmentation of crowns and tooth decay performs worse, with a Sørensen–Dice coefficient of 44.74% and 24.58% respectively. The overview of the Sørensen–Dice coefficient for every model and segmentation target evaluated is shown in Table 6.17. There is no clear or significant difference in performance in regards to the age of the sample.

**Table 6.17:** The performance for segmentation of teeth and dental alterations per evaluated architecture for panoramic dental x-ray images.

| Model | Tooth | Decay | Crown | Filling | Root canal filling |
|---|---|---|---|---|---|
| Unet Micro | 92.02% | 00.24% | 00.32% | 52.50% | 00.00% |
| Unet Mini | 93.69% | 09.52% | 38.43% | 69.59% | 67.84% |
| Unet | 95.03% | **24.58%** | 45.87% | 76.04% | 71.56% |
| Unet Big | **95.12%** | 20.25% | **47.74%** | **77.11%** | **74.31%** |
| FCN32 | 87.51% | 00.11% | 00.32% | 50.94% | 31.87% |
| FCN16 | 90.75% | 00.10% | 00.32% | 58.97% | 53.10% |
| FCN8 | 93.41% | 00.11% | 01.81% | 66.43% | 55.71% |
| DeepLab V3 - MobileNet | 92.05% | 00.10% | 11.70% | 56.59% | 49.74% |
| DeepLab V3 - HRNet V2 W32 | 94.42% | 00.00% | 43.87% | 73.02% | 70.63% |
| DeepLab V3 - ResNet50 | 93.07% | 04.74% | 44.26% | 70.15% | 62.40% |
| DeepLab V3 - ResNet101 | 92.82% | 00.49% | 31.03% | 67.19% | 59.75% |
| DeepLab V3 - HRNet V2 W48 | 94.45% | 00.55% | 44.77% | 72.98% | 69.99% |

Segmentation is a more computationally demanding computer vision task and thus slower than age estimation and sex assessment. The large UNet variant is the biggest and slowest model and can process 100 images in 2.4 seconds. Manual segmentation of panoramic dental x-ray images is a very arduous task, and it can take between 30 to 40 minutes per image for a trained expert.

All models perform best with the segmentation of the entire tooth. A tooth is well defined, has mostly clear boundaries, and can consequently easily be segmented. This can be observed in Table 6.17, as every model, no matter how large or small, is capable of tooth segmentation with a Sørensen–Dice coefficient close to 95%. Larger models handle less clear and edge cases

better, but overall, segmentation of individual teeth can be done automatically with a very high success rate. All three neural network architectures perform well, with UNet being consistently better, followed by DeepLab v3 and, lastly, FCN. Dental fillings, too, can be automatically segmented with a high success rate, as the best model achieves a Sørensen–Dice coefficient of 77.14%. They are more radio-opaque than teeth and bones, but depending on the amount and type of material used (amalgam or composite fillings), it can be ambiguous which part of a tooth is a dental filling and which is part of the natural dental structure. Root canal fillings have similar properties, except that their position is well defined within the structure of a tooth. They tend to be less radio-opaque, and as less material is used, there are cases where they are hard to spot. Crowns follow a similar trend, where their appearance has enough variance to cause ambiguity and ultimately model errors. Tooth decay performs significantly worse than any other alterations, and the reasons are twofold. Available data for tooth decay is scarce. While tooth decay is common, it is not the primary reason for dental imaging. Effectively, this results in less total data available for tooth decay. Another reason is the insidious nature of how tooth decay presents in dental x-ray images. Tooth decay destroys the dental tissue by slowly spreading from a surface inwards. Its appearance is not uniform as it manifests by thinning and weakening the dental tissue, therefore making it less radio-opaque and less visible on the x-ray image. However, teeth themselves do not have a uniform look in that regard, allowing tooth decay to hide in the natural variance of a tooth's appearance. These two factors combined lead to a low success rate for the segmentation of tooth decay.

The current literature has multiple studies about deep learning-based tooth segmentation of panoramic dental x-ray images. Zhao et al. [127] propose TSASNet, a two-stage attention model for segmentation. They, similar to the approach in this thesis, generate a segmentation map of all teeth present in the panoramic dental x-ray image, and they achieve a Sørensen–Dice coefficient of 93.77% on a dataset of 1500 panoramic dental x-ray images. The study acknowledges that teeth can have different alterations, and they keep track of missing teeth, tooth restorations, dental appliances, and dental implants. However, the study does not perform segmentation of those alterations. Chen et al. [170] propose MSLPNet for tooth segmentation. MSLPNet is a multi-scale model which uses Resnet50 as its feature extractor, which is inspired by multi-scale spatial pyramid pooling [125], an approach similar to DeepLab v3. Similar to Zhao et al., they enumerate the possible alterations the teeth in their dataset can have, but their model only does segmentation of the entire tooth. They achieve a Sørensen–Dice coefficient of 93.01%. Da Silva Rocha et al. [134] use four UNet variants for tooth segmentation of panoramic dental x-ray images. Their UNet variants differ significantly from those used in this thesis, and they diverge from the classical UNet approach. Their dataset consists of 1500 panoramic dental x-ray images. The results vary between variants, with the Sørensen–Dice coefficient ranging from 89.86% to 92.89%. The best-performing model presented in this thesis

achieves a Sørensen–Dice coefficient of 95.12%, firmly outperforming current methods in the literature. UNet Mini is the closest match to the performance of other studies, achieving an accuracy of 93.69%. No study in forensic odontology literature tackles the segmentation of tooth alterations, thus, no direct comparisons can be made.

### 6.3.2 Individual dental x-ray images

The segmentation of individual teeth x-ray images has a similar goal to the segmentation of panoramic dental x-ray images but with a more targeted approach. Panoramic dental x-ray images have a lot of "dead space," areas with tissue not related to teeth. Individual tooth x-ray images focus exclusively on the tooth, allowing for a more detailed segmentation. Additionally, as the focus is on only the tooth, even if the image size is smaller than the image size for panoramic dental x-ray images, the image's effective resolution is much higher. This allows for a more precise segmentation, as more tooth-specific information is present in the input image. The prerequisite to using such models is to have knowledge about the positions of teeth, which can be provided by the proposed model in Section 5.5, and whose results will be presented and analyzed in the upcoming Section 6.4.2 and Section 6.4.3. The overview of the data used is given in Section 3.3.3, and the models and approach used is explained in Section 5.4. An additional variant of data is added by removing all surrounding tissue in an individual tooth x-ray image. The ground truth segmentation map of the entire tooth is used to remove the tissue surrounding the tooth. The metric used to evaluate the model is the same as for panoramic dental x-ray images, the Sørensen–Dice coefficient. The Jaccard index and the Sørensen–Dice coefficient do not differ significantly for the developed models, and the issues with accuracy have been explained in Section 5.4. An analysis of the performance per alteration, per neural network architecture and their variants, per data variants, as well as a comparison to the performance of panoramic dental x-ray images is performed.

Similar to the results of segmentation of panoramic dental x-ray images, the best performing model is the UNet architecture in its large variant. It achieves a Sørensen–Dice coefficient of 95.82% for the segmentation of the entire tooth. The best performance for alterations is for the segmentation of crowns, which achieves a Sørensen–Dice coefficient of 88.54%, closely followed by dental fillings with a Sørensen–Dice coefficient of 84.70%. Root canal filling segmentation performs slightly worse, with a Sørensen–Dice coefficient of 79.47%. Again, tooth decay is the lowest performing task, this time achieving a Sørensen–Dice coefficient of 59.31%. For the second data variant, no models are trained for tooth segmentation as the image is pre-processed using tooth segmentation to remove the surrounding tissue. The models with this variant still perform best with crowns, achieving a Sørensen–Dice coefficient of 88.96%. Dental filling models achieve a Sørensen–Dice coefficient of 83.42%, root canal filling models achieve 78.94%, and tooth decay models achieve 64.63%. Interestingly, for this data variant,

the best performance for tooth decay and root canal fillings is achieved with the baseline UNet model instead of its larger variant. The detailed overview of performance for all alterations and for tooth segmentation for individual tooth images with surrounding tissue is shown in Table 6.18, and the performance for all alterations for the segmentation of individual tooth images without surrounding tissue is shown in Table 6.19.

**Table 6.18:** The performance for segmentation of teeth and dental alterations per evaluated architecture for individual tooth x-ray images which include the surrounding structure.

| Model | Tooth | Decay | Crown | Filling | Root canal filling |
|---|---|---|---|---|---|
| Unet Micro | 95.00% | 06.85% | 18.91% | 77.96% | 72.35% |
| Unet Mini | 95.44% | 32.97% | 72.04% | 81.77% | 76.36% |
| Unet | 95.76% | 56.46% | 88.01% | 84.66% | 79.07% |
| Unet Big | **95.82%** | **59.31%** | **88.54%** | **84.70%** | **79.47%** |
| FCN32 | 94.96% | 06.73% | 7429 | 80.55% | 70.11% |
| FCN16 | 95.44% | 36.75% | 8076 | 82.43% | 75.75% |
| FCN8 | 95.53% | 40.69% | 819 | 83.52% | 76.77% |
| DeepLab V3 - MobileNet | 95.23% | 17.49% | 7694 | 79.75% | 72.19% |
| DeepLab V3 - HRNet V2 W32 | 95.79% | 50.15% | 8025 | 80.74% | 77.95% |
| DeepLab V3 - ResNet50 | 95.53% | 46.84% | 8379 | 82.01% | 75.58% |
| DeepLab V3 - ResNet101 | 95.52% | 34.89% | 8250 | 82.06% | 74.66% |
| DeepLab V3 - HRNet V2 W48 | 95.81% | 14.19% | 8487 | 83.05% | 76.55% |

The processing time for segmentation of individual dental x-ray images is slightly lower than for panoramic dental x-ray images due to the difference in the input image size. The large UNet variant can process 100 images in 1.9 seconds. Manual segmentation is faster than for the panoramic case, as a panoramic dental x-ray image is a series of individual-tooth annotations. Depending on the complexity of the tooth, an expert needs up to 5 minutes to fully segment a tooth. This performance does not change depending on the data variant used, as both variants use images of the same size.

The model for segmentation of the entire tooth is the most successful, achieving a Sørensen–Dice coefficient of 95.82%. As was the case with panoramic dental x-ray images, the success metric does not deviate significantly between neural network architectures and their variants. Overall, smaller capacity models perform slightly worse, but overall the Sørensen–Dice coefficient is between 95% and 96%. Other alterations show a higher variance in their results, but most do not have significant outliers except tooth decay. The performance ranking is consistent across

**Table 6.19:** The performance for segmentation of teeth and dental alterations per evaluated architecture for individual tooth x-ray images which **do not** include the surrounding structure.

| Model | Decay | Crown | Filling | Root canal filling |
|---|---|---|---|---|
| Unet Micro | 40.92% | 70.76% | 77.73% | 73.35% |
| Unet Mini | 50.87% | 87.14% | 82.29% | 75.19% |
| Unet | **64.63%** | 87.93% | 83.24% | **78.94%** |
| Unet Big | 60.73% | **88.96%** | **83.42%** | 78.84% |
| FCN32 | 19.90% | 67.21% | 77.90% | 71.13% |
| FCN16 | 30.26% | 81.67% | 81.02% | 75.17% |
| FCN8 | 53.53% | 83.24% | 81.99% | 76.23% |
| DeepLab V3 - MobileNet | 37.83% | 78.63% | 79.44% | 72.40% |
| DeepLab V3 - HRNet V2 W32 | 57.01% | 86.66% | 80.45% | 77.18% |
| DeepLab V3 - ResNet50 | 57.86% | 85.94% | 80.69% | 75.92% |
| DeepLab V3 - ResNet101 | 55.57% | 85.77% | 80.58% | 75.06% |
| DeepLab V3 - HRNet V2 W48 | 59.00% | 86.71% | 80.14% | 77.54% |

model and data variants, with the best performance being achieved for the segmentation of crowns, followed by fillings, root canal fillings, and tooth decay. There is no significant difference in performance between samples by age nor by tooth type.

Between neural network architecture and their variants, a similar trend is noticeable as is for the segmentation of panoramic dental x-ray images. All but the smallest models perform well, with larger models handling edge cases better and achieving an overall better performance. All neural network architectures perform well, but UNet is still the best performing architecture, followed by DeepLab v3 and, lastly, FCN.

Segmenting an individual tooth in its entirety works well, and results are consistent across neural network architectures and model variants. Given the nature of the images, neighboring teeth are often visible in an image, and those teeth can have alterations of their own. In combination with the surrounding tissue, this can create ambiguous conditions for the model. The pre-segmented data variant avoids these problems by focusing the model solely on an individual tooth, with no external influences. Tooth decay segmentation models achieve a significant performance increase, improving their Sørensen–Dice coefficient from 59.31% to 64.63%. Segmentation of crowns improves, too, albeit with a less substantial increase from 88.54% to 88.96%. The performance for the segmentation models for dental fillings and root canal fillings slightly decreases, from 84.70% to 83.42% and from 79.47% to 78.94%. This is caused by

the fact that fillings and root canal fillings can extend slightly beyond the border of the tooth. By applying the tooth segmentation, those border regions can get cut, which can, across many samples, accumulate in the small performance decrease seen. Another factor is related to the border of the fillings too.

Compared to panoramic dental x-ray images, x-ray images of individual teeth offer more details and information about the most important part of the dental system - the teeth themselves. When only the tooth is segmented, and not any alterations, the performance difference is not significantly different. Models for segmentation of teeth from panoramic dental x-ray images achieve a Sørensen–Dice coefficient of 95.15%, while the models for segmentation of teeth from individual tooth x-ray images achieve a Sørensen–Dice coefficient of 95.82%. The area of a tooth is well defined and differentiated from surrounding tissue, thereby not posing too much of a challenge for the segmentation models. For alterations, the performance for individual tooth x-ray images is across the board substantially better compared to panoramic dental x-ray image models. The most significant impact can be seen for tooth decay, where the performance increased from 24.58% to 64.63%. As already discussed, tooth decay can be hard to spot on dental x-ray images, and thus "zooming in" and discarding all unnecessary surrounding tissue allows the models not only to learn features for tooth decay properly but also to work on a higher effective resolution which allows for a more precise segmentation. Segmentation models for crowns too increased significantly, from 47.74% to 88.96%. The reasons are similar to tooth decay, where the noise from the surrounding tissue combined with a low effective resolution for the crown alterations on the image results in a segmentation model with worse performance. The models for dental fillings and root canal fillings also improve but to a lesser degree. Those alterations are less ambiguously imaged with x-rays, thereby reducing the gain from reducing miscellaneous tissue from the input image.

As discussed in Section 6.3.1, not many studies tackle the problem of tooth segmentation of dental x-ray images. Zhao et al. [127] achieve a Sørensen–Dice coefficient of 93.77% with their TSASNet architecture, Chen et al. [170] achieve a Sørensen–Dice coefficient of 93.01% with their MSLPNet architecture, and da Silva Rocha et al. [134] achieve aSørensen–Dice coefficient between 89.86% to 92.89% on using some more exotic UNet variants. The model proposed in this thesis for segmentation of individual tooth x-ray images achieves a dice score of 95.82%, outperforming the current methods for tooth segmentation. The difference between panoramic and individual tooth x-ray images is not as significant for the comparison of results as it might seem at first glance. Panoramic dental x-ray images contain multiple teeth that need to be segmented, but those teeth are of an effectively smaller resolution. Thus more minor segmentation errors do not impact the evaluation metric significantly. Individual tooth x-ray images show only one tooth. However, they have a much higher effective resolution of the tooth images and therefore need to precisely segment all the details, which in turn affect the evaluation

metric to a much higher degree. Additionally, as explained in Section 5.4, the Sørensen–Dice coefficient does not get inflated by true negatives like accuracy and thereby delivers a more stable metric for comparisons, making it a more suitable metric for comparison with other studies. For individual images, studies about automated segmentation of tooth alterations can be found in the literature. Zhu et al. [136] developed CariesNet, which focuses on the segmentation of tooth decay on panoramic dental x-ray images. The architecture of CariesNet is based on full-scale axial attention modules, which are applied on multiple scales, and recombined into the final segmentation map. Their dataset consists of 3217 image regions containing tooth decay, which has been annotated on 1159 panoramic dental x-ray images. CariesNet achieves a Sørensen–Dice coefficient of 93.64%, which outperform the 64.63% of the proposed model for segmentation of tooth decay by a significant margin. While the removal of surrounding tissue improved the results from 59.31% to 64.64%, the gap between CariesNet and all UNet variants is still significant. This is primarily due to a low sample count available for the training of the UNet variants proposed in this thesis. CariesNet is trained on 3217 samples from 1159 panoramic dental x-ray images, while the proposed models are trained on 365 samples from 813 panoramic dental x-ray images. This is nearly an order of magnitude fewer samples. Of all tooth alterations, tooth decay displays the most variety, as it can be of any size, and it can appear on any part of the tooth or multiple parts of the tooth simultaneously. Additionally, while the tooth and other alterations present clearly with x-ray imaging as they are proper independent objects (dental fillings, root canal fillings, crowns), tooth decay is an intricate structural change to the tooth. This makes tooth decay the hardest to detect and delineate its region of influence.

## 6.4 Detection and type determination of teeth in dental x-ray images

### 6.4.1 Determination of tooth type in individual dental x-ray images

Manual forensic odontology methods either compare measurements to reference tables or use them as an input with established linear models that links those measurements to the desired target (age, sex, ...). Those models and reference tables are built with the knowledge of the tooth type that is being measured, as the measurements differ from tooth to tooth. Every position in the dental system has its own function, and thus all teeth slightly differ from each other. There are multiple classification systems used for tooth types. The standard used is the FDI dual notation system [139], which assigns a two-digit label to every tooth, depending on its quadrant within the oral cavity and the position within the quadrant. This gives a unique label to every tooth, resulting in 32 labels. Research studies sometimes use a reduced subset, which results in the 16-class, 8-class, and 4-class classification systems, which are explained in Section 5.3.3.

While the tooth type can be easily determined if all teeth are present, tooth type determination can be a challenge in the case of missing teeth, cases where teeth are found independently from the rest of the remains, or they are the only remains available. The presented results show the performance of the deep learning-based model, which determines the tooth type from individual dental x-ray images, which is built as explained in Section 5.3.4 upon the data shown in Section 3.3.2. The metric used to measure the performance of the model is accuracy. The differences in performance for all classification systems are analyzed, as well as the differences between the model using state-of-the-art feature extractor and attention and BiFPN-based model. An analysis of the performance trend in regard to age is shown, the common classification errors are examined, and the impact of alterations on the model performance is examined.

The model based on a state-of-the-art feature extractor with attention performs tooth type determination with an overall accuracy of 97.99%, 92.40%, 92.23%, and 83.74% for the 4-class, 8-class, 16-class, and 32-class cases, respectively. The model performance improves when only non-altered teeth are taken into account, resulting in an accuracy of 99.15% 4-class approach, 95.53% 8-class approach, 95.45% for the 16-class approach, and 87.24% for the 32-type approach. The precision of those models is, in the same order, 98.02%, 92.48%, 92.28%, and 84.03% for all samples, and 99.18%, 94.92%, 95.11%, and 88.16% for samples with no alterations. Likewise, the recall for those models is 97.99%, 92.40%, 92.23%, and 83.74% for all samples, and 99.15%, 95.53%, 95.46%, and 87.24% for samples with no alterations. The hyperparameters for this model are: VGG16 as the feature extractor, a convolutional depth of 423 channels, no attention, and a fully-connected layer of size 412, which results in a model of 14M learnable parameters. A series of BiFPN-based models were also trained. Their performance ranks similarly to the model based on a state-of-the-art feature extractor, achieving an overall accuracy of 98.24%, 91.89%, 89.10%, and 83.92% for the 4-class, 8-class, 16-class, and 32-class classification approaches, respectively, and an accuracy of 99.17%, 95.00%, 93.08%, and 87.54% on samples with no alterations. The overall precision of those models is 98.20%, 91.96%, 89.51%, and 84.32%, and unaltered samples have a precision of 99.12%, 94.61%, 92.98%, and 88.77%. And finally, the overall recall for those models is 98.24%, 91.89%, 89.10%, and 83.92%, and the recall for samples with no alterations is 99.17%, 95.00%, 93.08%, and 87.54%. The best performing hyperparameters are: 2 BiFPN layers, convolutional depth starting with 16 channels and linearly scaling to 128 channels, a fully-connected layer of size 32, the feature compressor depth of 4, and a BiFPN feature size of 16. Like in previous cases, the BiFPN-based models for tooth type determination achieve a similar results while only requiring a fraction of learnable parameters - in this case, only 2.3% compared to the model based on a state-of-the-art feature extractor. The overview of overall results per age group and classification approach is shown in Table 6.20, and the overview for teeth with no alterations in shown in Table 6.21.

**Table 6.20:** Performance of tooth type determination per type classification system and age group for all individual tooth x-ray samples.

| Age group | Tooth type | | | |
|---|---|---|---|---|
| | Accuracy (4 types) | Accuracy (8 types) | Accuracy (16 types) | Accuracy (32 types) |
| [20, 25) | 98.95% | 96.10% | 95.91% | 88.77% |
| [25, 30) | 98.29% | 93.90% | 93.33% | 85.36% |
| [30, 35) | 98.81% | 95.08% | 94.66% | 85.33% |
| [35, 40) | 98.21% | 92.60% | 93.27% | 87.44% |
| [40, 45) | 96.96% | 85.71% | 85.95% | 75.29% |
| [45, 50) | 96.10% | 89.61% | 88.31% | 76.62% |
| [50, 55) | 95.29% | 84.71% | 84.71% | 75.69% |
| [55, 60) | 91.34% | 78.74% | 78.74% | 60.63% |
| **Overall** | 97.99% | 92.40% | 92.23% | 83.74% |

**Table 6.21:** Performance of tooth type determination per type classification system and age group for individual tooth x-ray samples with no alterations.

| Age group | Tooth type | | | |
|---|---|---|---|---|
| | Accuracy (4 types) | Accuracy (8 types) | Accuracy (16 types) | Accuracy (32 types) |
| [20, 25) | 99.87% | 97.73% | 97.23% | 90.29% |
| [25, 30) | 98.36% | 94.21% | 94.52% | 85.76% |
| [30, 35) | 99.63% | 96.89% | 96.65% | 88.20% |
| [35, 40) | 100.00% | 97.97% | 97.97% | 91.50% |
| [40, 45) | 99.35% | 91.96% | 92.39% | 81.30% |
| [45, 50) | 97.75% | 92.13% | 91.01% | 84.27% |
| [50, 55) | 97.98% | 92.93% | 91.92% | 86.87% |
| [55, 60) | 94.12% | 90.20% | 88.24% | 76.47% |
| **Overall** | 99.15% | 95.53% | 95.46% | 87.24% |

On an Nvidia 2080 TI GPU, the model based on a state-of-the-art feature extractor with attention can assess the tooth type of 100 images in 1.1 seconds, while the BiFPN-based model can process the same amount of images in 0.14 seconds. Manual tooth type determination can take up to a few minutes per tooth.

The performance of the model based on a state-of-the-art feature extractor with attention and the BiFPN-based model are overall in line with each other, except for the case of the 16-class approach. The BiFPN-based model for the 16-class approach has greater difficulties differentiating teeth within the same 4-class category. In other words, it still differentiates well between molars and premolars, but the mistakes happen in accurately labeling if a tooth is, for example, the first or second molar. This holds true for both mandibular and maxillary teeth. However, this trend is not observable when evaluated on the subset of samples with no alterations. The BiFPN-based models for all other approaches do not exhibit such behavior. All other trends which are analyzed in the following paragraphs are identical between both the model based on a state-of-the-art feature extractor with attention and BiFPN-based model.

Tooth type determination is, naturally, easier for the classification systems with fewer categories, which is reflected in the performance of the models. The model for the 4-class approach has the highest accuracy, achieving an overall accuracy of 97.99% and an accuracy of 99.15% on unaltered samples. Interestingly, the performance of the model for the 8-class and the 16-class approach is similar, with the 8-class model reaching an overall accuracy of 92.40% and an accuracy of 95.53% for unaltered teeth, while the 16-class approach achieves an overall accuracy of 92.23% and an accuracy of 95.46% for unaltered teeth. The difference between the 8-class and 16-class approaches is the differentiation between mandibular and maxillary teeth. Differentiation between those categories does not pose a challenge to the created models, as small bits of the surrounding structures are visible in the images, which contain information about the jaw side. As the model can easily differentiate between mandibular and maxillary teeth, the rest of the model capacity is dedicated to, in essence, the 8-class problem, ultimately resulting in a similar performance of the 8-class and 16-class model. This trend can be seen in Figure 6.15, which shows the confusion matrices for the 8-class, 16-class, and 32-class models. As can be seen, the misclassifications happening to the 8-class model are errors between the basic four tooth types - the model rarely misclassifies an incisor as a canine, premolar, or molar, with the vast majority of errors happening between incisors. The same holds true for premolars and molars. Canines rarely get misclassified, as they have a distinct shape and function compared to the other three basic tooth types. The 16-class system shows the same misclassifications, the only difference being that those errors are present both for mandibular and maxillary teeth. However, it is clearly visible that those errors do not happen across jaw sides. This phenomenon persists in the 32-class model. The 32-class model performs the least accurate of all the tooth type determination models, achieving an overall accuracy of 83.74%

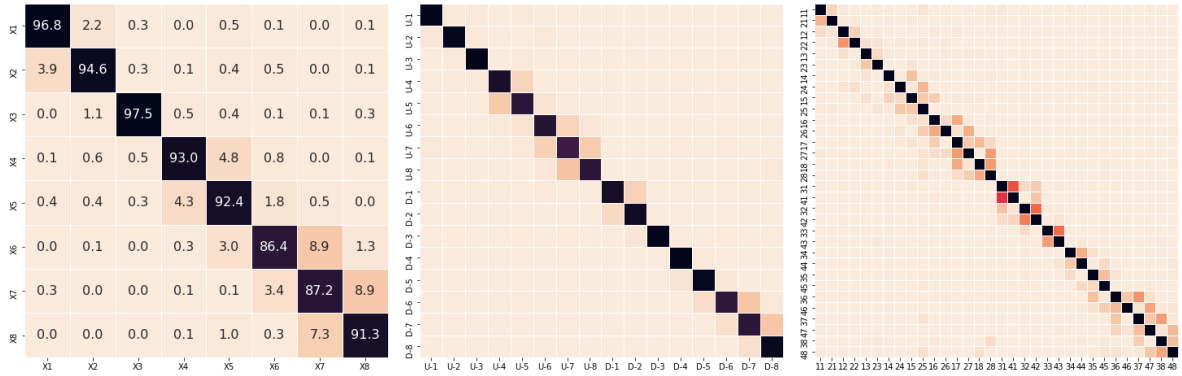


**Figure 6.15:** Confusion matrices for tooth type determination for the 8-, 16-, and 32-class approaches. As shown in Figure 5.9, the rows represent model predictions, and columns represent the ground truth values. For the 16-class case (middle figure), the prefixes "U-" and "D-" represent maxillary (**U**pper) and mandibular (**L**ower) 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). Misclassifications mostly happen between teeth in the same morphological group (4-class system), regardless of the classification system used. For this qualitative overview, values are omitted in the middle and right figure for visual clarity.

and an accuracy of 87.24% on unaltered teeth. Misclassification does not occur across jaw sides but does happen across quadrants. Teeth are divided into four quadrants, but each jaw side is symmetrical, leading to misclassifications. As shown in Figure 6.15, most misclassifications happen between the same 8-class type of the same jaw side. The model can fully differentiate between mandibular and maxillary teeth, as well as the basic four tooth types, with the performance decreasing with an increased abstraction of the tooth type classification approach. All observed trends hold true from the perspective of precision and recall.

As with all previously tackled forensic odontology tasks, performance declines with increased age. In both the overall and unaltered-teeth case, the results remain consistent for samples up to the age of 40, after which performance starts to decline. This holds true for models of all tooth type classification approaches. With age, the number of tooth alterations increases, including damage and decay. Nevertheless, the decrease in performance is not as significant as seen in deep learning-based models previously shown for other forensic odontology tasks. Despite the age-related accumulated damage, the tooth mostly retains its shape and structure, allowing for clear tooth type determination. Comparing the overall results to the results of non-altered teeth, the decrease is less pronounced, with the results having less variance up to the age of 55.

As can be seen in the differences of Table 6.20 and Table 6.21, the tooth type determination models perform better with unaltered teeth. Tooth alterations change the structure of a tooth, which can be to the extent where the tooth type cannot be accurately determined. A dental filling can affect a significant portion of a tooth. However, most dental fillings are not significant enough to consequentially affect the tooth type determination models and affect the results the least. To be precise, samples with dental fillings achieve an accuracy of 99.54% for the 4-class

approach, 93.20% for the 8-class approach, 93.12% for the 16-class approach, and 87.24% for the 32-class approach. The performance for samples with root canal fillings is slightly worse than for fillings, achieving accuracies of 98.31%, 87.46%, 88.48%, and 84.07% for the four classification approaches. The root canal filling by itself does not significantly change the shape and structure of the tooth, but they are always paired with dental fillings, as a path to the root has to be made. This combination of alterations results in an overall lower performance than dental fillings alone. In a similar vein, tooth decay decreases the performance, but not by a significant margin. For the four classification approaches, models produce an accuracy of 99.59%, 94.31%, 95.93%, and 89.02% for samples with tooth decay. Tooth decay is treated early to contain the spread and damage, thus only small areas with tooth decay can be found. Likewise, if tooth decay has been allowed to proceed, the tooth is removed. Removed teeth are an interesting case to examine with this research. Intuitively, if a tooth is missing, we cannot determine its type. Therefore, when evaluated on image samples where the tooth is missing, it would be natural to expect an accuracy of at-best random chance. This is not the case. Models for all classification approaches achieve an accuracy of 88.02%, 72.65%, 69.86%, and 52.69% for samples where the tooth is missing. The tooth is missing, but the surrounding tissue is still present. The performance is significantly degraded, but it is still much better than random. As teeth do not grow at a perfectly straight angle, neighboring teeth can be visible in an analyzed image, thus providing the model information about the position and type of the tooth. Crowns, bridges, tooth germs, leftover roots, and dental implants are other alterations that can be present in the dataset, as described in Section 3.3.2. Their impact is significant, resulting in accuracies of 87.56%, 68.39%, 70.47%, and 60.62%. Depending on the classification approach, the result is either slightly higher or slightly lower than in the case of missing teeth. Those alterations heavily modify the shape of the tooth, with artificial crowns replacing the entire crown of the tooth, bridges linking different tooth elements together, leftover roots missing over half the tooth, tooth germs being undeveloped teeth, and implants fully replacing a tooth. Such significant modifications or direct replacement of teeth erase the information of the tooth type, leading to the model having to rely on the surrounding information, thus producing similar results to the case when the tooth is missing. Not enough samples are available for those alterations, so their impact is grouped together. The detailed overview of the performance per tooth status for the 4-, 8-, and 16-class approach is shown in Table 6.22, and for the 32-class approach in Table 6.23.

Automated tooth type determination has been explored in literature. Oktay [140] is one of the earliest studies of automated tooth type determination. They use a model based on the AlexNet architecture to simultaneously detect tooth positions in the form of bounding boxes and determine the type of the tooth using a 3-class approach where incisors and canines are grouped together. Their dataset consists of 100 panoramic dental x-ray images. On average, they achieve

**Table 6.22:** Overview of the accuracy of tooth type determination, per dental alteration and tooth type (4-, 8-, and 16-class systems). 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%** |

**Table 6.23:** Overview of the accuracy of tooth type determination, per dental alteration and tooth type (32-class systems). 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 |
|---|---|---|---|---|---|---|---|
| 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%** |

an accuracy of 92.84%, Keerthana et al. [141] use projection profile analysis to determine the tooth type using the 4-class approach. They use a dataset of 200 individual tooth samples, split equally between all four tooth types, achieving an accuracy of 92.54%. The model proposed in the thesis achieves an overall accuracy of 97.99%, and an accuracy of 99.15% on samples with no alterations, firmly outperforming both approaches. Chen et al. [142] apply the Faster R-CNN object detection architecture to detect teeth and determine their tooth type on periapical dental x-ray images. They determine tooth type using the 32-class approach, but they correct potential misclassifications using heuristics based on the type determinations of neighboring teeth. Their dataset consists of 1250 periapical dental x-ray images. This approach achieves an accuracy of 79.7% before the heuristic correction and an accuracy of 91.7% after. While the baseline approach does not outperform the method proposed in this thesis, it does outperform it once the correction heuristic is applied. Kim et al. [171] is one of the newest studies tackling this problem. The study uses Faster R-CNN with an Inception v3 feature extractor to detect teeth in panoramic dental x-ray images and determine their tooth type. Their dataset consists of 303 panoramic dental x-ray images. The tooth type determination approach is a bit different than most in the literature. Teeth are classified into three categories: incisors, canines, or molars; premolars and molars are grouped together. Their approach achieves an accuracy of 84.5% and a recall of 84.2%. The closest match to the trained tooth type determination models in the thesis is the 4-class approach, which achieves an overall accuracy of 97.99% and a recall of 97.99%, which strongly outperforms their method. None of the current studies in the literature make information available about possible tooth alterations if such samples were filtered out, nor is the impact on their approach discussed.

### 6.4.2 Detection of teeth without type determination

All previously described methods that use individual tooth x-ray images have the prerequisite of annotating individual tooth positions. While this process does not require a high degree of expertise, it is still a time-consuming and repetitive effort that can be automated. There have been some experiments using state-of-the-art object detection neural networks for the localization of teeth. However, most of those studies use either dated architectures or a small dataset. YOLO v5 [160] is the newest state-of-the-art family of object detection neural network architectures based on the breakthrough YOLO architecture. This family of architectures offers four sizes - small, medium, large, and extra large. An additional variant is available, which extends the backbone with additional layers, adding an extra input for the spatial pyramid pooling, labeled as "P6". The models are named after those two properties. For example, the medium-sized model is labeled *YOLOv5m* for the baseline model and *YOLOv5m6* for the extended model. The models are trained on panoramic dental x-ray images described in Section 3.3.1, with the bounding box positional annotations as described in Sections 3.1 and 3.3. This includes the subset of tooth an-

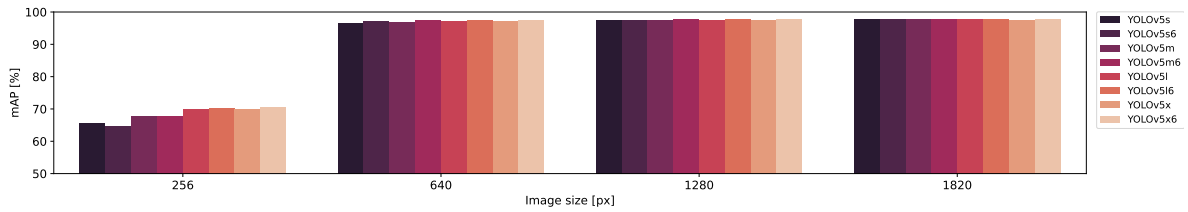


**Figure 6.16:** The performance results for detection with the YOLO v5 model variants for different image sizes. As can be seen, the image size significantly impacts the achieved performance. On the other hand, the model architecture has a lower degree of impact, which diminishes even further with the increase in image size.

notations that contain information about the tooth status. The samples which have been marked as "missing" teeth have been removed from these experiments. The success of these models is measured with established object detection metrics, which are precision, recall, and mAP. All eight variants of the model have been evaluated, and their differences analyzed. Different image sizes have been evaluated, their differences are analyzed, and the best image and model size is determined. Additionally, two training approaches have been tested, with models either being trained from random initialization or by using pretrained weights trained on the COCO dataset [164] as the starting point, the differences in performance are analyzed, and the better training approach for this task is determined.

Overall, the best performing tooth detection model is based on the YOLOv5m6 architecture, using an image size of 1820 px by 1820 px, trained with transfer learning based on weights pretrained on the COCO dataset. The mAP achieved is 97.85%, with a precision of 98.08% and a recall of 94.39%. The model consists of 35.7M learnable parameters, and it can process 100 images in 0.45 seconds. The overview of the performance for every model architecture and image size combination is shown in Table 6.25.

Out of all of the evaluated factors, image size is the most impactful. Models with a higher number of learnable parameters, and thus a higher capacity, tend to perform somewhat better within an image size group. This holds true for the extended models, as they perform slightly better than their non-extended counterparts. Figure 6.16 shows the mAP of all evaluated models grouped by image size, where it can be seen that larger image sizes contribute to better results. This phenomenon diminishes with each increase, leaving just a minuscule increase in performance between the models processing images of 1280 by 1280 px and 1820 by 1820 px. The detailed overview of results can be seen in Table 6.25, where it can be seen that the differences shrink nearly an order of magnitude with every resolution increase. Nevertheless, while model capacity correlates with mAP, the best performance was achieved by the middle-sized model, YOLOv5m6, processing the highest resolution images.

For most use cases, training models with randomly initialized weights instead of pretrained weights is less preferable, as the pretraining is done on enormous general datasets. Those

**Table 6.24:** Peformance results for all YOLO v5 model variants and all evaluated image sizes. The metrics marked with [1] are for models trained from scratch, and the ones marked with [2] are for models trained with transfer learning.

| Architecture | Image size | Precision[1] | Recall[1] | mAP[1] | Precision[2] | Recall[2] | mAP[2] |
|---|---|---|---|---|---|---|---|
| YOLOv5s | 256 | 68.20% | 70.13% | 65.63% | 88.08% | 74.81% | 83.22% |
| YOLOv5s | 640 | 97.60% | 92.92% | 96.52% | 98.02% | 94.11% | 97.20% |
| YOLOv5s | 1280 | 98.37% | 93.70% | 97.37% | 98.32% | 93.85% | 97.41% |
| YOLOv5s | 1820 | 98.37% | 93.80% | 97.74% | 98.37% | 93.99% | 97.83% |
| YOLOv5s6 | 256 | 69.99% | 70.46% | 64.59% | 84.86% | 76.42% | 82.04% |
| YOLOv5s6 | 640 | 97.63% | 93.58% | 97.04% | 98.24% | 94.07% | 97.74% |
| YOLOv5s6 | 1280 | 98.34% | 93.48% | 97.46% | 98.24% | 93.77% | 97.55% |
| YOLOv5s6 | 1820 | 98.42% | 93.70% | 97.77% | 98.29% | 93.87% | 97.82% |
| YOLOv5m | 256 | 69.60% | 71.61% | 67.68% | 86.93% | 74.76% | 83.43% |
| YOLOv5m | 640 | 97.97% | 93.83% | 96.89% | 98.28% | 93.66% | 97.01% |
| YOLOv5m | 1280 | 98.39% | 93.62% | 97.55% | 97.40% | 94.76% | 97.52% |
| YOLOv5m | 1820 | 98.43% | 93.75% | 97.81% | 98.37% | 94.03% | 97.73% |
| YOLOv5m6 | 256 | 72.85% | 71.33% | 67.63% | 89.38% | 76.16% | 84.83% |
| YOLOv5m6 | 640 | 97.65% | 93.80% | 97.28% | 98.10% | 94.09% | 97.67% |
| YOLOv5m6 | 1280 | 98.29% | 93.78% | 97.70% | 98.27% | 93.99% | 97.66% |
| **YOLOv5m6** | **1820** | **98.12%** | **94.10%** | **97.83%** | **98.08%** | **94.39%** | **97.85%** |
| YOLOv5l | 256 | 70.15% | 73.15% | 69.83% | 87.87% | 76.49% | 83.23% |
| YOLOv5l | 640 | 97.93% | 93.62% | 96.98% | 97.83% | 94.26% | 97.29% |
| YOLOv5l | 1280 | 98.43% | 93.47% | 97.58% | 98.10% | 94.24% | 97.53% |
| YOLOv5l | 1820 | 98.43% | 93.91% | 97.73% | 98.04% | 94.12% | 97.71% |
| YOLOv5l6 | 256 | 74.10% | 73.18% | 70.09% | 86.87% | 76.36% | 82.43% |
| YOLOv5l6 | 640 | 98.11% | 93.64% | 97.47% | 98.14% | 94.32% | 97.84% |
| YOLOv5l6 | 1280 | 98.43% | 93.49% | 97.65% | 95.23% | 90.70% | 95.83% |
| YOLOv5l6 | 1820 | 98.34% | 93.88% | 97.78% | 84.63% | 79.01% | 87.87% |
| YOLOv5x | 256 | 69.55% | 73.73% | 69.99% | 91.16% | 74.90% | 85.10% |
| YOLOv5x | 640 | 98.15% | 93.55% | 96.99% | 98.22% | 94.08% | 97.08% |
| YOLOv5x | 1280 | 98.26% | 93.84% | 97.48% | 97.37% | 94.91% | 97.55% |
| YOLOv5x | 1820 | 98.19% | 93.31% | 97.48% | 98.35% | 94.03% | 97.73% |
| YOLOv5x6 | 256 | 76.85% | 71.92% | 70.62% | 85.09% | 76.05% | 80.70% |
| YOLOv5x6 | 640 | 98.17% | 93.83% | 97.55% | 97.12% | 94.98% | 97.60% |
| YOLOv5x6 | 1280 | 98.29% | 94.01% | 97.70% | 97.64% | 94.72% | 97.55% |
| YOLOv5x6 | 1820 | 98.36% | 93.73% | 97.82% | 97.13% | 94.96% | 97.69% |

models are trained for a different task, but given the broadness and size of the dataset, layers closer to the input discover generally useful features that can easily be adapted to other tasks. The dataset used for pretraining, COCO [164], contains natural images taken with ordinary cameras. However, medical images are a wholly different domain than natural images. Those images represent elements that are not encountered in natural-image datasets, especially not in the case of x-ray images. Given the stark difference in domains and the size of the dataset used in this thesis, all models are trained twice, once without pretrained weights and once with. The average performance for models trained with pretrained weights is 90.16%, and the average performance for models trained without pretrained weights is 93.60%. When tested with a paired t-test, the p-value is 0.0097, which firmly allows for the rejection of the null hypothesis that the samples are on average equal. Therefore, the huge difference in image domains and the large dataset used for the training of the models does affect the model performance, and training from random initialization is preferable.

Current studies in literature pair tooth detection with tooth type determination and the details of their approach are therefore already been discussed in Section 6.4.1, as well as in Chapter 2. Oktay [140] is one of the earliest studies on tooth detection. A sliding window approach is used, where every possible image region generated by the sliding window is processed with AlexNet. They do not report an mAP, but they report the precision of every object class. The precision achieved is 92.47% for incisors and canines, 91.74% for premolars, and 94.32% for molars. The overall precision achieved by the best model proposed in this thesis is 98.08%, which firmly outperforms Oktay. Chen et al. [142] do tooth detection on periapical dental x-ray images using the Faster R-CNN architecture. The achieved mAP is "approximately 80%", the overall precision of 90% and the overall recall is 98.5%. While this approach is outperformed by mAP and precision, the recall of the method that Chen et al. propose is significantly higher at 98.5%. Periapical images focus on a much narrower region and therefore show fewer teeth, additionally have less space that is not occupied by teeth, and the effective resolution of a tooth is higher with periapical dental x-ray images. Their model has, therefore, less room for error, which reflects in their recall metric. Kim et al. [171] also uses Faster R-CNN but implements Inception v3 as its feature extractor. They achieve an mAP of 96.7% for tooth detection, which is slightly underperforming compared to the model proposed in the thesis. Laishram et al. [144] uses Faster R-CNN to detect teeth in panoramic dental x-ray images. They achieve an mAP of 91.40%. While this seems like an uncharacteristically low mAP, their dataset is smaller with only 145 images, and their dataset includes panoramic dental x-ray images of children. Panoramic dental x-ray images of children contain overlapping teeth as their permanent teeth develop within the jaw while their primary teeth are still in place.

### 6.4.3 Detection of teeth with tooth type determination

Modern neural network architectures for object detection can simultaneously classify the detected object. This allows for a fast and efficient object detection and classification system, which can share common features between those two tasks. A system like this is end-to-end trainable, allowing for easier and more stable training. Such a combined system has more information available for training, which allows for achieving better overall performance. As with the previously shown detection results, precision, recall, and mAP are the success metrics used to evaluate the model. Given the previous research results, the best performance is achieved with images of size of 1820 by 1820 px, trained from randomly initialized weights, using YOLOv5m6 as the neural network architecture. The difference between architectures is minimal for large images; therefore, the choice of YOLOv5 architecture is the least impactful. The performance of the resulting models is compared to the baseline detection case, and the trends in detection and classification results are analyzed.

The model trained for the 4-class approach achieves a precision of 98.6%, a recall of 98.2%, and an mAP of 98.3%. The models for the 8-class and 16-class approaches achieve a precision of 96.2% and 97.1%, a recall of 96.2% and 96.0%, and an mAP of 97.3% and 96.9% respectively. Finally, the model trained for the 32-class approach achieves a precision of 94.4%, a recall of 94.9%, and an mAP of 95.9%. Like the previously shown results, the model consists of 35.7M learnable parameters, and it can process 100 images in 0.45 seconds. There is no significant inference time difference between the classification of the 4-class and 32-class approaches. The overview of results is shown in Table 6.25.

As expected and as observed in previously shown results, models perform better the lower the number of classes for the tooth type determination approach is. Equally, the performance difference between the 8-class and 16-class approaches is the smallest, as the differentiation between maxillary and mandibular teeth does not pose a significant challenge to deep learning-based models. The model trained for the 32-class approach has the lowest performance. Interestingly, the detection model trained for the 4-class tooth type classification approach outperforms the baseline detection model, which achieved an mAP of 97.85%. The precision of both models is similar, but the recall of the baseline detection model is 94.39%, which is significantly lower than the 4-class model recall of 98.2%. In other words, the additional information about the tooth type enabled the discovery of more robust features, which led to an overall improvement in detection results. The recall for all other tooth type classification approaches is higher, with the difference in the recall of the model for the 32-class approach having the least significant difference, its recall being 94.9%. While the 32-class approach gives more precise information about the location of a tooth, its relative complexity to the 4-class approach did not allow for the learning of features that significantly improve the detection results, as was in the case of the 4-class detection model.

**Table 6.25:** Performance for tooth detection with simultaneous tooth type determination for all tooth type classification systems. For every tooth type classification system, a separate model was trained.

| Tooth type | Precision | Recall | mAP |
|---|---|---|---|
| Incisor | 98.4% | 98.9% | 98.2% |
| Canine | 98.2% | 99.1% | 98.7% |
| Premolar | 98.7% | 97.6% | 98.4% |
| Molar | 99.0% | 97.3% | 98.0% |
| **Overall** | **98.6%** | **98.2%** | **98.3%** |
| X1 | 97.0% | 98.3% | 97.0% |
| X2 | 98.0% | 97.8% | 98.0% |
| X3 | 96.1% | 99.1% | 98.7% |
| X4 | 96.6% | 96.0% | 96.7% |
| X5 | 96.8% | 95.1% | 97.5% |
| X6 | 94.3% | 95.1% | 97.6% |
| X7 | 94.9% | 96.1% | 97.3% |
| X8 | 96.2% | 92.3% | 95.3% |
| **Overall** | **96.2%** | **96.2%** | **97.3%** |
| Down-1 | 97.0% | 97.8% | 96.0% |
| Down-2 | 98.4% | 99.0% | 98.8% |
| Down-3 | 98.7% | 99.7% | 98.7% |
| Down-4 | 98.7% | 99.0% | 99.1% |
| Down-5 | 97.9% | 98.1% | 98.6% |
| Down-6 | 99.0% | 93.3% | 96.4% |
| Down-7 | 94.5% | 95.1% | 96.2% |
| Down-8 | 92.4% | 95.6% | 94.4% |
| Up-1 | 97.4% | 98.5% | 97.1% |
| Up-2 | 97.9% | 97.1% | 97.4% |
| Up-3 | 97.2% | 97.7% | 97.5% |
| Up-4 | 93.5% | 91.1% | 93.9% |
| Up-5 | 98.2% | 91.1% | 96.5% |
| Up-6 | 97.4% | 94.6% | 97.2% |
| Up-7 | 96.9% | 96.5% | 98.0% |
| Up-8 | 97.9% | 91.8% | 94.5% |
| **Overall** | **97.1%** | **96.0%** | **96.9%** |

| Tooth type | Precision | Recall | mAP |
|---|---|---|---|
| 11 | 94.6% | 96.9% | 96.1% |
| 12 | 96.0% | 98.1% | 98.0% |
| 13 | 96.0% | 97.4% | 96.7% |
| 14 | 89.7% | 93.9% | 92.4% |
| 15 | 93.6% | 91.6% | 94.1% |
| 16 | 90.8% | 96.3% | 96.9% |
| 17 | 94.8% | 96.6% | 98.1% |
| 18 | 94.8% | 90.7% | 91.7% |
| 21 | 83.2% | 99.0% | 97.4% |
| 22 | 97.9% | 96.4% | 96.8% |
| 23 | 97.5% | 98.0% | 97.5% |
| 24 | 95.4% | 93.5% | 93.6% |
| 25 | 93.7% | 93.7% | 95.9% |
| 26 | 92.6% | 92.8% | 95.1% |
| 27 | 91.4% | 96.0% | 96.7% |
| 28 | 97.1% | 91.3% | 95.4% |
| 31 | 96.5% | 69.5% | 95.1% |
| 32 | 97.5% | 98.6% | 98.7% |
| 33 | 98.1% | 99.0% | 98.5% |
| 34 | 96.1% | 98.4% | 98.7% |
| 35 | 97.5% | 96.3% | 97.1% |
| 36 | 96.7% | 91.9% | 93.4% |
| 37 | 94.0% | 96.0% | 96.0% |
| 38 | 91.6% | 95.0% | 94.0% |
| 41 | 86.5% | 86.3% | 88.8% |
| 42 | 97.1% | 98.5% | 96.9% |
| 43 | 97.0% | 99.5% | 98.2% |
| 44 | 97.7% | 100% | 98.5% |
| 45 | 96.2% | 99.5% | 98.2% |
| 46 | 96.2% | 94.8% | 95.5% |
| 47 | 92.5% | 96.7% | 95.2% |
| 48 | 91.1% | 93.4% | 94.0% |
| **Overall** | **94.4%** | **94.9%** | **95.9%** |

The overall precision and recall for the detection models are across all tooth type classification approaches higher than the precision and recall for the tooth determination models which classify individual tooth x-ray images. The baseline classification models achieve an overall precision of 98.02%, 92.48%, 92.28%, and 84.03%, while the detection models achieve an overall precision of 98.6%, 96.2%, 97.1%, 94.4%. Likewise, the baseline classification models achieve an overall recall of 97.99%, 92.40%, 92.23%, and 83.74%, while the detection models achieve an overall recall of 98.20%, 96.2%, 96.0%, 94.9%. The decreasing trend with the increasing complexity of the classification approach persists, but the decrease is much lower. The main factor behind this phenomenon is the end-to-end nature of the detection architecture. While the model classifies a proposed region, it can use a wider view of the tooth and its neighborhood to determine its type. As demonstrated in Section 6.4.1, the classification model exploits the visible surrounding structures around a tooth to determine its type, as shown by the high performance of the samples where the tooth is missing. In contrast to the classification models which cannot access additional neighborhood information as it simply is not available in the single tooth x-ray images, the detection models not only define the proposed region of interest they classify, they have access to the entire panoramic dental x-ray image. Other research in literature [142] shows that including information about the neighborhood even in the form of heuristics can heavily influence the performance. Another effect of this model behavior is the success of detecting and determining the tooth type of altered teeth. There is, therefore, no significant change in performance between samples with and without alterations, excluding the samples of missing teeth, as those teeth cannot be found by the detection model.

The most significant studies tackling a similar problem have already been discussed in Section 6.4.1 and Section 6.4.2. Oktay [140] reports a precision of 92.47% for incisors and canines, 91.74% for premolars, and 94.32% for molars, which is significantly lower than the precision achieved by the proposed model for the 4-class approach, which achieves an overall precision of 98.6%. Chen et al. [142] detect teeth in periapical dental x-ray images and determine the tooth type using the 32-class approach. Before the heuristic correction, they achieve an overall precision of 79.7%, and with the correction, they achieve an overall precision of 91.7% and an mAP of "approximately 80%". Laishram et al [144] bases its model on the Faster R-CNN architecture to detect teeth in panoramic dental x-ray images. Their model differentiates between the four basic tooth types and achieves an mAP of 91.40%. Kim et al. [171], too, use Faster R-CNN to detect teeth, but it also uses the Inception v3 architecture as the feature extractor. The model determines the tooth type with a 3-class approach, with no differentiation between premolars and molars. They report an mAP of 96.7% for tooth detection, a precision of 75.5%, and a recall of 84.2%. The model proposed in this thesis achieves an overall precision of 94.4% and an mAP of 95.9% while not relying on heuristic post-processing, thereby outperforming all current methods in the literature.

# Chapter 7

# Conclusion

Forensic odontology studies dental remains intending to identify a person and their demographic information. Age estimation and sex assessment are the two fundamental tasks of a forensic examination, as demographic information is key in identifying a person. Dental remains are particularly suitable for forensic examination due to their extraordinary stability and durability in the face of external force, fire, biological decomposition, and other external factors. In the early stages, destructive methods were used for estimation, but with the introduction of radiographic imaging destructive methods fell out of favor. Still, the most widely recognized and reliable methods in forensic odontology require meticulous manual measurements by highly trained experts. Those methods require precise measurements and a substantial time to analyze, which, combined with their repetitiveness and dependence on reference charts and even atlases, makes them prone to human error. With the rise and success of deep learning-based methods, the field of computer vision improved dramatically and extended its reach as wide as medical image analysis. Given the success in other medical fields, deep learning-based computer vision methods are a natural fit for forensic odontology.

This thesis focuses on the estimation of age, assessment of sex, and determination of tooth type from dental x-ray images, as well as the segmentation of regions affected by artificial and natural dental alterations from adult dental x-ray images using deep learning-based methods. Two convolutional neural network architectures are proposed. One uses state-of-the-art convolutional neural network architectures for image analysis as the feature extractor combined with the attention mechanism, and one is wholly built around the Weighted Bi-directional Feature Pyramid Network module. A combination of grid and random search is used to determine the best model architecture and corresponding hyperparameters. Multiple models are made for panoramic and individual dental x-ray images, processing a wider domain of the problem space. The model search space is exhaustively explored, and the models for each task are extensively and rigorously evaluated on one of the most extensive datasets in literature, including samples with external alterations. The performance and results are assiduously analyzed, which includes

the impact dental alterations have on the correctness and stability of the developed models.

The performance analysis of all developed deep learning-based approaches proposed in this thesis suggests that they perform as well or better than currently used methods while simultaneously being fully automated, reproducible, and multiple orders of magnitude faster than the manual approaches. Analysis of panoramic dental x-ray images consistently, across all tasks, yields more accurate results compared to their single-tooth image counterparts. Age estimation models work well across the board, with the models estimating age as a probability distribution excelling in accuracy and giving insightful indicators about the result confidence. The occlusion-based interpretability analysis of age estimation models suggests that both teeth and the surrounding skeletal structure contain independent informative age indicators, which can be used separately or in conjunction to obtain a more precise estimation. Estimations from individual tooth x-ray images are, on average, negatively impacted by alterations, with some exceptions for common alterations in older samples. Similarly, the results of sex assessment deep learning-based models suggest an improvement in accuracy compared to current methods in the literature. The gradient-based interpretability analysis suggests that mandibular indicators contribute most to the assessment of panoramic dental x-ray images. For individual teeth, the analysis suggests that most female indicators are concentrated in the crown of the tooth. In contrast, male indicators are spread between the crown and the root of the mandibular teeth. Tooth alterations have, across the board, a negative impact on assessment accuracy. The proposed segmentation models perform well, with the results suggesting that models for segmentation of individual teeth outperform current models in literature for both panoramic and individual tooth x-ray images. Except for tooth decay, no other studies currently tackle the automated segmentation of dental alterations. While the proposed models show some weakness for the segmentation of tooth decay, the results suggest a satisfactory success for all other alterations. Removing the surrounding skeletal tissue from individual tooth x-ray images improves the results across the board. Tooth detection is also successful, matching the performance of other studies with a smaller model, with less learnable parameters and requiring less compute. Results for simultaneous tooth type determination and detection suggest that the models proposed in this study outperform the models in current literature in both detection and tooth type determination across all four tooth type classification systems. Equally, results from standalone tooth type determination of individual tooth x-ray images suggest better performance than current automated approaches across all classification systems. Independent from the tooth type classification system, the most common errors were between teeth in the same 4-class group, i.e., misclassifications would happen between two different molars or two different incisors, but very rarely between canines and molars, or incisors and premolars. As with all previously proposed models, tooth type determination performance is negatively affected by the presence of tooth alterations.

The results of this thesis affirm that forensic odontology tasks can be successfully automated with deep learning-based image analysis methods. Those methods can match and even outperform current forensic odontology approaches in relevant performance metrics while broadening which tooth samples are usable by including ones with dental alterations. The proposed approaches may help in practical applications and in gaining further insight into the impact of dental alterations on forensic analysis procedures and the effects of age and the expression of sex in adult teeth. Ultimately, the deep learning-based approaches proposed in this thesis may be used to improve the accuracy, reliability, and speed of forensic estimation of age, sex, and tooth type from adult dental x-ray images.

# Bibliography

[1]Dudar, J. C., Pfeiffer, S., Saunders, S., “Evaluation of morphological and histological adult skeletal age-at-death estimation techniques using ribs”, Journal of Forensic Science, Vol. 38, No. 3, 1993, str. 677–685.

[2]Hu, K.-S., Koh, K.-S., Han, S.-H., Shin, K.-J., Kim, H.-J., “Sex determination using non-metric characteristics of the mandible in Koreans”, Journal of forensic sciences, Vol. 51, No. 6, 2006, str. 1376–1382.

[3]Srivastava, P. C., “Correlation of odontometric measures in sex determination”, J Indian Acad Forensic Med, Vol. 32, No. 1, 2010, str. 56–61.

[4]Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., Sánchez, C. I., “A survey on deep learning in medical image analysis”, Medical Image Analysis, Vol. 42, Dec. 2017, str. 60–88, dostupno na: https://www.sciencedirect.com/science/article/pii/S1361841517301135

[5]Piccialli, F., Somma, V. D., Giampaolo, F., Cuomo, S., Fortino, G., “A survey on deep learning in medicine: Why, how and when?”, Information Fusion, Vol. 66, Feb. 2021, str. 111–137, dostupno na: https://www.sciencedirect.com/science/article/pii/S1566253520303651

[6]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., “Attention is All you Need”, in Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 2017, str. 5998–6008, dostupno na: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[7]Viola, P., Jones, M., “Rapid object detection using a boosted cascade of simple features”, in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Vol. 1, Dec. 2001, str. I–I, iSSN: 1063-6919.

[8]Rosenblatt, F., The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory, 1957.

[9]Rosenblatt, F., “The perceptron: a probabilistic model for information storage and organization in the brain”, Psychological Review, Vol. 65, No. 6, Nov. 1958, str. 386–408.

[10]Cybenko, G., “Approximation by superpositions of a sigmoidal function”, Mathematics of Control, Signals, and Systems, Vol. 2, No. 4, Dec. 1989, str. 303–314, dostupno na: http://link.springer.com/10.1007/BF02551274

[11]Hornik, K., Stinchcombe, M., White, H., “Multilayer feedforward networks are universal approximators”, Neural Networks, Vol. 2, No. 5, Jan. 1989, str. 359–366, dostupno na: https://www.sciencedirect.com/science/article/pii/0893608089900208

[12]Nair, V., Hinton, G. E., “Rectified linear units improve restricted boltzmann machines”, in Proceedings of the 27th International Conference on International Conference on Machine Learning, ser. ICML’10. Madison, WI, USA: Omnipress, Jun. 2010, str. 807–814.

[13]Schmidhuber, J., “Deep learning in neural networks: An overview”, Neural networks, Vol. 61, 2015, str. 85–117, publisher: Elsevier.

[14]Rumelhart, D. E., Hinton, G. E., Williams, R. J., “Learning representations by back-propagating errors”, Nature, Vol. 323, No. 6088, Oct. 1986, str. 533–536, number: 6088 Publisher: Nature Publishing Group, dostupno na: https://www.nature.com/articles/323533a0

[15]LeCun, Y., “Generalization and network design strategies”, Connectionism in perspective, Vol. 19, No. 143-155, 1989, str. 18, publisher: North Holland.

[16]LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., Jackel, L. D., “Backpropagation Applied to Handwritten Zip Code Recognition”, Neural Computation, Vol. 1, No. 4, Dec. 1989, str. 541–551, dostupno na: https://doi.org/10.1162/neco.1989.1.4.541

[17]Cireşan, D. C., Meier, U., Gambardella, L. M., Schmidhuber, J., “Deep, Big, Simple Neural Nets for Handwritten Digit Recognition”, Neural Computation, Vol. 22, No. 12, Dec. 2010, str. 3207–3220, dostupno na: https://doi.org/10.1162/NECO_a_00052

[18]Ruder, S., “An overview of gradient descent optimization algorithms”, dostupno na: http://arxiv.org/abs/1609.04747 ArXiv:1609.04747 [cs]. Jun. 2017.

[19]Kingma, D. P., Ba, J., “Adam: A Method for Stochastic Optimization”, arXiv:1412.6980 [cs], Dec. 2014, arXiv: 1412.6980, dostupno na: http://arxiv.org/abs/1412.6980

[20]Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., Recht, B., "The Marginal Value of Adaptive Gradient Methods in Machine Learning", arXiv:1705.08292 [cs, stat], May 2017, arXiv: 1705.08292, dostupno na: http://arxiv.org/abs/1705.08292

[21]Smith, L. N., "Cyclical Learning Rates for Training Neural Networks", arXiv:1506.01186 [cs], Jun. 2015, arXiv: 1506.01186, dostupno na: http://arxiv.org/abs/1506.01186

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

[23]Koutsoukas, A., Monaghan, K. J., Li, X., Huan, J., "Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data", Journal of Cheminformatics, Vol. 9, No. 1, Jun. 2017, str. 42, dostupno na: https://doi.org/10.1186/s13321-017-0226-y

[24]Snoek, J., Larochelle, H., Adams, R. P., "Practical Bayesian optimization of machine learning algorithms", in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2, ser. NIPS'12. Red Hook, NY, USA: Curran Associates Inc., Dec. 2012, str. 2951–2959.

[25]LaValle, S. M., Branicky, M. S., Lindemann, S. R., "On the relationship between classical grid search and probabilistic roadmaps", The International Journal of Robotics Research, Vol. 23, No. 7-8, 2004, str. 673–692.

[26]Bergstra, J., Bengio, Y., "Random search for hyper-parameter optimization", The Journal of Machine Learning Research, Vol. 13, No. null, Feb. 2012, str. 281–305.

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

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

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

[30]Simonyan, K., Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv:1409.1556 [cs], Sep. 2014, arXiv: 1409.1556, dostupno na: http://arxiv.org/abs/1409.1556

[31]Chollet, F., "Xception: Deep Learning with Depthwise Separable Convolutions", in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, Jul. 2017, str. 1800–1807, dostupno na: http://ieeexplore.ieee.org/document/8099678/

[32]Shahrokh Esfahani, M., Dougherty, E. R., "Effect of separate sampling on classification accuracy", Bioinformatics, Vol. 30, No. 2, Jan. 2014, str. 242–250, dostupno na: https://doi.org/10.1093/bioinformatics/btt662

[33]Shorten, C., Khoshgoftaar, T. M., "A survey on Image Data Augmentation for Deep Learning", Journal of Big Data, Vol. 6, No. 1, Jul. 2019, str. 60, dostupno na: https://doi.org/10.1186/s40537-019-0197-0

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

[35]Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization", in 2017 IEEE International Conference on Computer Vision (ICCV), Oct. 2017, str. 618–626, iSSN: 2380-7504.

[36]Woo, S., Park, J., Lee, J.-Y., Kweon, I. S., "CBAM: Convolutional Block Attention Module", in Computer Vision – ECCV 2018, Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., (ur.). Cham: Springer International Publishing, 2018, Vol. 11211, str. 3–19, series Title: Lecture Notes in Computer Science, dostupno na: http://link.springer.com/10.1007/978-3-030-01234-2_1

[37]Zeiler, M. D., Fergus, R., "Visualizing and Understanding Convolutional Networks", in Computer Vision – ECCV 2014, ser. Lecture Notes in Computer Science, Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., (ur.). Cham: Springer International Publishing, 2014, str. 818–833.

[38]Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J., "Multi-scale Convolutional Neural Networks for Lung Nodule Classification", in Information Processing in Medical Imaging, ser. Lecture Notes in Computer Science. Springer International Publishing, 2015, str. 588–599.

[39]Bali čević, V., Lončarić, S., Cárdenes, R., Gonzalez-Tendero, A., Paun, B., Crispi, F., Butakoff, C., Bijnens, B., "Assessment of Myofiber Orientation in High Resolution Phase-Contrast CT Images", in Functional Imaging and Modeling of the Heart, ser. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, str. 111–119.

[40]Grinsven, M. J. J. P. v., Ginneken, B. v., Hoyng, C. B., Theelen, T., Sánchez, C. I., "Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images", IEEE Transactions on Medical Imaging, Vol. 35, No. 5, May 2016, str. 1273–1284.

[41]Hosseini-Asl, E., Ghazal, M., Mahmoud, A., Aslantas, A., Shalaby, A. M., Casanova, M. F., Barnes, G. N., Gimel'farb, G., Keynton, R., El-Baz, A., "Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network", Frontiers in Bioscience (Landmark Edition), Vol. 23, No. 3, Jan. 2018, str. 584–596.

[42]Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., Navab, N., "AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images", IEEE Transactions on Medical Imaging, Vol. 35, No. 5, May 2016, str. 1313–1321.

[43]Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzebski, S., Fevry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L. L. Y., Ho, K., Weinstein, J. D., Reig, B., Gao, Y., Toth, H., Pysarenko, K., Lewin, A., Lee, J., Airola, K., Mema, E., Chung, S., Hwang, E., Samreen, N., Kim, S. G., Heacock, L., Moy, L., Cho, K., Geras, K. J., "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening", IEEE transactions on medical imaging, Vol. 39, No. 4, Apr. 2020, str. 1184–1194.

[44]Zhu, W., Liu, C., Fan, W., Xie, X., "DeepLung: Deep 3D Dual Path Nets for Automated Pulmonary Nodule Detection and Classification", in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Mar. 2018, str. 673–681.

[45]Altan, A., Karasu, S., "Recognition of COVID-19 disease from X-ray images by hybrid model consisting of 2D curvelet transform, chaotic salp swarm algorithm and deep learning technique", Chaos, Solitons, and Fractals, Vol. 140, Nov. 2020, str. 110071.

[46]Laguarta, J., Hueto, F., Subirana, B., "COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings", IEEE Open Journal of Engineering in Medicine and Biology, Vol. 1, 2020, str. 275–281, conference Name: IEEE Open Journal of Engineering in Medicine and Biology.

[47]Saunders, E., The Teeth a Test of Age, Considered with Reference to the Factory Children: Addressed to the Members of Both Houses of Parliament. Renshaw, 1837.

[48]Krogman, W. M., "The human skeleton in forensic medicine. I.", Postgraduate medicine, Vol. 17, No. 2, 1955, str. A–48.

[49]Solheim, T., "Dental age estimation. An alternative technique for tooth sectioning.", The American journal of forensic medicine and pathology, Vol. 5, No. 2, 1984, str. 181–184.

[50]Röntgen, W. C., "Uber eine neue Art von Strahlen", Sitzungsber Phys Med Ges Wurtzburg, Vol. 9, 1895, str. 132–141.

[51]Matsikidis, G., Schultz, P., "Altersbestimmung nach dem Gebiss mit Hilfe des Zahnfilms", Zahnarztl Mitt, Vol. 72, No. 2524, 1982, str. 2527–2528.

[52]Eckert, W. G., Garland, N., "The history of the forensic application in radiology", The American Journal of Forensic Medicine and Pathology, Vol. 5, No. 1, Mar. 1984, str. 53–56, dostupno na: https://journals.lww.com/amjforensicmedicine/Abstract/1984/03000/The_history_of_the_forensic_application_in.10.aspx

[53]Singaraju, S., Sharada, P., "Age estimation using pulp/tooth area ratio: A digital image analysis", Journal of Forensic Dental Sciences, Vol. 1, No. 1, 2009, str. 37.

[54]Carvalho, S. P. M., Silva, R. H. A. d., Lopes-Júnior, C., Peres, A. S., "Use of images for human identification in forensic dentistry", Radiologia Brasileira, Vol. 42, No. 2, Apr. 2009, str. 125–130, dostupno na: http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0100-39842009000200012&lng=en&nrm=iso&tlng=pt

[55]Panchbhai, A., "Dental radiographic indicators, a key to age estimation", Dentomaxillofacial Radiology, Vol. 40, No. 4, May 2011, str. 199–212, dostupno na: http://www.birpublications.org/doi/10.1259/dmfr/19478385

[56]Limdiwala, P. G., Shah, J. S., "Age estimation by using dental radiographs", Journal of Forensic Dental Sciences, Vol. 5, No. 2, 2013, str. 118–122, dostupno na: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3826039/

[57]Marroquin, T., Karkhanis, S., Kvaal, S., Vasudavan, S., Kruger, E., Tennant, M., "Age estimation in adults by dental imaging assessment systematic review", Forensic Science International, Vol. 275, Jun. 2017, str. 203–211, dostupno na: https://linkinghub.elsevier.com/retrieve/pii/S0379073817301081

[58]Badran, D. H., Othman, D. A., Thnaibat, H. W., Amin, W. M., "Predictive Accuracy of Mandibular Ramus Flexure as a Morphologic Indicator of Sex Dimorphism

in Jordanians", International Journal of Morphology, Vol. 33, No. 4, Dec. 2015, str. 1248–1254, dostupno na: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-95022015000400009&lng=en&nrm=iso&tlng=en

[59]Shah, N., Bansal, N., Logani, A., "Recent advances in imaging technologies in dentistry", World Journal of Radiology, Vol. 6, No. 10, Oct. 2014, str. 794–807, dostupno na: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4209425/

[60]Wu, C.-H., Tsai, W.-H., Chen, Y.-H., Liu, J.-K., Sun, Y.-N., "Model-Based Orthodontic Assessments for Dental Panoramic Radiographs", IEEE Journal of Biomedical and Health Informatics, Vol. 22, No. 2, Mar. 2018, str. 545–551, conference Name: IEEE Journal of Biomedical and Health Informatics.

[61]Lai, Y., Fan, F., Wu, Q., Ke, W., Liao, P., Deng, Z., Chen, H., Zhang, Y., "LCANet: Learnable Connected Attention Network for Human Identification Using Dental Images", IEEE Transactions on Medical Imaging, Vol. 40, No. 3, Mar. 2021, str. 905–915, conference Name: IEEE Transactions on Medical Imaging.

[62]Kang, D.-Y., Duong, H. P., Park, J.-C., "Application of Deep Learning in Dentistry and Implantology", Implantology, Sep. 2020, str. 148–181, publisher: IMPLANTOLOGY, dostupno na: www.implantology.or.kr/articles/article/RvNO/

[63]Greulich, W. W., Pyle, S. I., Radiographic atlas of skeletal development of the hand and wrist. Stanford university press, 1959.

[64]Roche, A. F., Thissen, D., Chumlea, W., Assessing the skeletal maturity of the hand-wrist: Fels method. Thomas, 1988.

[65]Ciapparelli, L., "The chronology of dental development and age assessment", Practical forensic odontology, 1992, str. 22–42.

[66]Gleiser, I., Hunt Jr, E. E., "The permanent mandibular first molar: its calcification, eruption and decay", American journal of physical anthropology, Vol. 13, No. 2, 1955, str. 253–283, publisher: Wiley Subscription Services, Inc., A Wiley Company New York.

[67]Moorrees, C. F., Fanning, E. A., Hunt Jr, E. E., "Age variation of formation stages for ten permanent teeth", Journal of dental research, Vol. 42, No. 6, 1963, str. 1490–1502, publisher: SAGE Publications Sage CA: Los Angeles, CA.

[68]Cameriere, R., Ferrante, L., Cingolani, M., "Age estimation in children by measurement of open apices in teeth", International Journal of Legal Medicine, Vol. 120, No. 1, Jan. 2006, str. 49–52, dostupno na: https://doi.org/10.1007/s00414-005-0047-9

[69]Sayegh, F., Reed, A., “Calcification in the dental pulp”, Oral Surgery, Oral Medicine, Oral Pathology, Vol. 25, No. 6, 1968, str. 873–882, publisher: Elsevier.

[70]Demirjian, A., Goldstein, H., Tanner, J. M., “A New System of Dental Age Assessment”, Human Biology, Vol. 45, No. 2, 1973, str. 211–227, publisher: Wayne State University Press, dostupno na: https://www.jstor.org/stable/41459864

[71]Nolla, C. M. *et al.*, “The development of permanent teeth”, Doktorski rad, University of Michigan, 1952.

[72]Haavikko, K., “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, 1970, str. 103–170.

[73]Pereira, C. P., Russell, L. M., de Pádua Fernandes, M., Alves da Silva, R. H., de Sousa Santos, R. F. V., “Dental Age Estimation based on Development Dental Atlas Assessment in a Child/Adolescent Population with Systemic Diseases”, Acta Stomatologica Croatica, Vol. 53, No. 4, Dec. 2019, str. 307–317, dostupno na: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6993474/

[74]Cameriere, R., Ferrante, L., De Angelis, D., Scarpino, F., Galli, F., “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”, International Journal of Legal Medicine, Vol. 122, No. 6, Nov. 2008, str. 493–497, dostupno na: https://doi.org/10.1007/s00414-008-0279-6

[75]Kvaal, S. I., Kolltveit, K. M., Thomsen, I. O., Solheim, T., “Age estimation of adults from dental radiographs”, Forensic Science International, Vol. 74, No. 3, Jul. 1995, str. 175–185, dostupno na: http://www.sciencedirect.com/science/article/pii/037907389501760G

[76]Drusini, A. G., Toso, O., Ranzato, C., “The coronal pulp cavity index: A biomarker for age determination in human adults”, American Journal of Physical Anthropology, Vol. 103, No. 3, 1997, str. 353–363, dostupno na: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291096-8644%28199707%29103%3A3%3C353%3A%3AAID-AJPA5%3E3.0.CO%3B2-R

[77]Ikeda, N., Umetsu, K., Kashimura, S., Suzuki, T., Oumi, M., “Estimation of age from teeth with their soft X-ray findings”, Japanese Journal of Legal Medicine, Vol. 39, No. 3, 1985, str. 244–250, publisher: Medico-Legal Society of Japan.

[78]Cameriere, R., Ferrante, L., Cingolani, M., “Variations in Pulp/Tooth Area Ratio as an Indicator of Age: a Preliminary Study”, Journal of Forensic Science, Vol. 49, No. 2,

Feb. 2004, str. 1–3, publisher: ASTM International, dostupno na: http://www.astm.org/DIGITAL_LIBRARY/JOURNALS/FORENSIC/PAGES/JFS2003259.htm

[79]Farah, C., Booth, D., Knott, S., "Dental maturity of children in Perth, Western Australia, and its application in forensic age estimation", Journal of clinical forensic medicine, Vol. 6, No. 1, 1999, str. 14–18, publisher: Elsevier.

[80]Ardakani, F., Bashardoust, N., Sheikhha, M., "The accuracy of dental panoramic radiography as an indicator of chronological age in Iranian individuals.", J Forensic Odontostomatol, Vol. 25, No. 2, 2007, str. 25.

[81]Gulsahi, A., Yuzugullu, B., Imirzalıoglu, P., Genç, Y., "Assessment of panoramic radiomorphometric indices in Turkish patients of different age groups, gender and dental status", Dentomaxillofacial Radiology, Vol. 37, No. 5, 2008, str. 288–292.

[82]Babar, M., Iqbal, S., Jan, A., "Essential guidelines for forensic dentistry", Pakistan Oral Dent J, Vol. 27, 2008, str. 79–84.

[83]Gali ć, I., Nakaš, E., Prohić, S., Selimović, E., Obradović, B., Petrovečki, M., "Dental age estimation among children aged 5–14 years using the Demirjian method in Bosnia-Herzegovina", Acta Stomatologica Croatica, Vol. 44, No. 1, 2010, str. 17–25.

[84]Gali ć, I., Vodanović, M., Cameriere, R., Nakaš, E., Galić, E., Selimović, E., Brkić, H., "Accuracy of Cameriere, Haavikko, and Willems radiographic methods on age estimation on Bosnian–Herzegovian children age groups 6–13", International journal of legal medicine, Vol. 125, No. 2, 2011, str. 315–321, publisher: Springer.

[85]Selmanagi ć, A., Ajanović, M., Kamber-Ćesir, A., Redžepagić-Vražalica, L., Jelešković, A., Nakaš, E., "Radiological Evaluation of Dental Age Assessment Based on the Development of Third Molars in Population of Bosnia and Herzegovina", Acta stomatol Croat, Vol. 54, No. 2, 2020, str. 161–167.

[86]Jeon, H.-M., Jang, S.-M., Kim, K.-H., Heo, J.-Y., Ok, S.-M., Jeong, S.-H., Ahn, Y.-W., "Dental Age Estimation in Adults: A Review of the Commonly Used Radiological Methods", Journal of Oral Medicine and Pain, Vol. 39, No. 4, 2014, str. 119–126, publisher: Korean Academy of Orofacial Pain and Oral Medicine, dostupno na: https://www.koreascience.or.kr/article/JAKO201402148668361.page

[87]Upalananda, W., Wantanajittikul, K., Na Lampang, S., Janhom, A., "Semi-automated technique to assess the developmental stage of mandibular third molars for age estimation", Australian Journal of Forensic Sciences, 2021, str. 1–11, publisher: Taylor & Francis.

[88]Zaborowicz, M., Zaborowicz, K., Biedziak, B., Garbowski, T., "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 (Basel, Switzerland), Vol. 22, No. 2, Jan. 2022, str. 637, dostupno na: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8777593/

[89]Vila-Blanco, N., Carreira, M. J., Varas-Quintana, P., Balsa-Castro, C., Tomas, I., "Deep Neural Networks for Chronological Age Estimation From OPG Images", IEEE transactions on medical imaging, Vol. 39, No. 7, Jul. 2020, str. 2374–2384.

[90]Guo, Y.-c., Han, M., Chi, Y., Long, H., Zhang, D., Yang, J., Yang, Y., Chen, T., Du, S., "Accurate age classification using manual method and deep convolutional neural network based on orthopantomogram images", International Journal of Legal Medicine, Mar. 2021, dostupno na: https://doi.org/10.1007/s00414-021-02542-x

[91]Kim, S., Lee, Y.-H., Noh, Y.-K., Park, F. C., Auh, Q.-S., "Age-group determination of living individuals using first molar images based on artificial intelligence", Scientific Reports, Vol. 11, No. 1, Jan. 2021, str. 1073, number: 1 Publisher: Nature Publishing Group, dostupno na: https://www.nature.com/articles/s41598-020-80182-8

[92]Yang, F., Jacobs, R., Willems, G., "Dental age estimation through volume matching of teeth imaged by cone-beam CT", Forensic Science International, Vol. 159, May 2006, str. S78–S83, dostupno na: http://www.sciencedirect.com/science/article/pii/S0379073806000788

[93]Gulsahi, A., Kulah, C. K., Bakirarar, B., Gulen, O., Kamburoglu, K., "Age estimation based on pulp/tooth volume ratio measured on cone-beam CT images", Dentomaxillofacial Radiology, Vol. 47, No. 1, Jan. 2018, str. 20170239, publisher: The British Institute of Radiology, dostupno na: https://www.birpublications.org/doi/full/10.1259/dmfr.20170239

[94]Hrdli čka, A., "Mandibular and maxillary hyperostoses", American Journal of Physical Anthropology, Vol. 27, No. 1, 1940, str. 1–67.

[95]Galdames, I. C. S., Matamala, D. A. Z., Smith, R. L., "Evaluating accuracy and precision in morphologic traits for sexual dimorphism in malnutrition human skull: a comparative study", Int J Morphol, Vol. 26, No. 4, 2008, str. 876–83.

[96]Saini, V., Srivastava, R., Rai, R. K., Shamal, S. N., Singh, T. B., Tripathi, S. K., "Mandibular Ramus: An Indicator for Sex in Fragmentary Mandible*", Journal of Forensic Sciences, Vol. 56, No. s1, 2011, str. S13–S16, dostupno na: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1556-4029.2010.01599.x

[97]Martin, E. S., “A study of an Egyptian series of mandibles, with special reference to mathematical methods of sexing”, Biometrika, Vol. 28, No. 1/2, 1936, str. 149–178.

[98]Morant, G. M., Collett, M., Adyanthaya, N. K., “A biometric study of the human mandible”, Biometrika, Vol. 28, No. 1/2, 1936, str. 84–122.

[99]De Villiers, H., “Sexual dimorphism of the skull of the South African Banu-speaking Negro”, South African Journal of Science, Vol. 64, No. 2, 1968, str. 118.

[100]Humphrey, L. T., Dean, M. C., Stringer, C. B., “Morphological variation in great ape and modern human mandibles”, Journal of Anatomy, Vol. 195, No. 4, Nov. 1999, str. 491–513, dostupno na: http://doi.wiley.com/10.1046/j.1469-7580.1999.19540491.x

[101]Giles, E., “Sex determination by discriminant function analysis of the mandible”, American Journal of Physical Anthropology, Vol. 22, No. 2, Jun. 1964, str. 129–135, dostupno na: http://doi.wiley.com/10.1002/ajpa.1330220212

[102]Steyn, M., İşcan, M. Y., “Sexual dimorphism in the crania and mandibles of South African whites”, Forensic Science International, Vol. 98, No. 1, Nov. 1998, str. 9–16, dostupno na: http://www.sciencedirect.com/science/article/pii/S0379073898001200

[103]Franklin, D., O’Higgins, P., Oxnard, C. E., Dadour, I., “Determination of Sex in South African Blacks by Discriminant Function Analysis of Mandibular Linear Dimensions: A Preliminary Investigation Using the Zulu Local Population”, Forensic Science, Medicine and Pathology, Vol. 2, No. 4, 2006, str. 263–268, dostupno na: http://link.springer.com/10.1385/FSMP:2:4:263

[104]Teke, H. Y., Duran, S., Canturk, N., Canturk, G., “Determination of gender by measuring the size of the maxillary sinuses in computerized tomography scans”, Surgical and Radiologic Anatomy, Vol. 29, No. 1, Feb. 2007, str. 9–13, dostupno na: http://link.springer.com/10.1007/s00276-006-0157-1

[105]Dayal, M., Spocter, M., Bidmos, M., “An assessment of sex using the skull of black South Africans by discriminant function analysis”, HOMO, Vol. 59, No. 3, Jul. 2008, str. 209–221, dostupno na: https://linkinghub.elsevier.com/retrieve/pii/S0018442X08000164

[106]Indira, A. P., Markande, A., David, M. P., “Mandibular ramus: An indicator for sex determination - A digital radiographic study”, Journal of Forensic Dental Sciences, Vol. 4, No. 2, 2012, str. 58–62, dostupno na: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3669477/

[107]Marinescu, M., Panaitescu, V., Rosu, M., “Sex determination in Romanian mandible using discriminant function analysis: Comparative results of a time-efficient method”, Romanian Journal of Legal Medicine, Vol. 21, No. 4, Dec. 2013, str. 305–308, dostupno na: http://www.rjlm.ro/index.php/arhiv/339

[108]Bhagwatkar, T., Thakur, M., Palve, D., Bhondey, A., Dhengar, Y., “Sex Determination by Using Mandibular Ramus - A Forensic Study”, Journal of Advanced Medical and Dental Sciences Research, Vol. 4, No. 2, 2016, str. 6.

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

[110]Maloth, K. N., Kundoor, V. K. R., Vishnumolakala, S. S. L. P., Kesidi, S., Lakshmi, M. V., Thakur, M., “Mandibular ramus: A predictor for sex determination-A digital radiographic study”, Journal of Indian Academy of Oral Medicine and Radiology, Vol. 29, No. 3, 2017, str. 242.

[111]Nagaraj, T., James, L., Gogula, S., Ghouse, N., Nigam, H., Sumana, C. K., “Sex determination by using mandibular ramus: A digital radiographic study”, Journal of Medicine, Radiology, Pathology and Surgery, Vol. 4, No. 4, 2017, str. 5–8, dostupno na: http://www.jmrps.net/eJournals/ShowText.aspx?ID=99&Type=FREE&TYP=TOP&IN=_eJournals/images/JPLOGO.gif&IID=16&Value=1&isPDF=YES

[112]Alias, A., Ibrahim, A., Bakar, S. N. A., Shafie, M. S., Das, S., Abdullah, N., Noor, H. M., Liao, I. Y., Nor, F. M., “Anthropometric analysis of mandible: an important step for sex determination”, La Clinica Terapeutica, Vol. 169, No. 5, Oct. 2018, str. e217–e223, dostupno na: http://www.clinicaterapeutica.it/ojs/index.php/ClinicaTerapeutica/article/view/259

[113]Deana, N. F., Alves, N., “Nonmetrical sexual dimorphism in mandibles of Brazilian individuals”, Biomedical Research, Vol. 28, No. 9, 2017, str. 4233–4238, dostupno na: http://www.biomedres.info/abstract/nonmetrical-sexual-dimorphism-in-mandibles-of-brazilian-individuals-7343.html

[114]Karaman, F., “Use of Diagonal Teeth Measurements in Predicting Gender in a Turkish Population”, Journal of Forensic Sciences, Vol. 51, No. 3, 2006, str. 630–635, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1556-4029.2006.00133.x, dostupno na: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1556-4029.2006.00133.x

[115]Joseph, A. P., Harish, R. K., Mohammed, P. K. R., Vinod Kumar, R. B., "How reliable is sex differentiation from teeth measurements", Oral & Maxillofacial Pathology Journal, Vol. 4, No. 1, 2013, str. 289–92.

[116]Capitaneanu, C., Willems, G., Jacobs, R., Fieuws, S., Thevissen, P., "Sex estimation based on tooth measurements using panoramic radiographs", International Journal of Legal Medicine, Vol. 131, No. 3, May 2017, str. 813–821.

[117]Capitaneanu, C., Willems, G., Thevissen, P., "A systematic review of odontological sex estimation methods", The Journal of Forensic Odonto-stomatology, Vol. 35, No. 2, Dec. 2017, str. 1–19, dostupno na: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6100233/

[118]Blanco, N. V., Vilas, R. R., Nouche, M. J. C., Carmona, I. T., "Towards deep learning reliable gender estimation from dental panoramic radiographs", in Proceedings of the 9th European Starting AI Researchers' Symposium 2020 co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago Compostela, Spain, August, 2020, ser. CEUR Workshop Proceedings, Rudolph, S., Marreiros, G., (ur.), Vol. 2655. CEUR-WS.org, 2020, dostupno na: http://ceur-ws.org/Vol-2655/paper22.pdf

[119]Ke, W., Fan, F., Liao, P., Lai, Y., Wu, Q., Du, W., Chen, H., Deng, Z., Zhang, Y., "Biological gender estimation from panoramic dental x-ray images based on multiple feature fusion model", Sensing and Imaging, Vol. 21, No. 1, 2020, str. 1–11.

[120]Son, L. H., Tuan, T. M., "Dental segmentation from X-ray images using semi-supervised fuzzy clustering with spatial constraints", Engineering Applications of Artificial Intelligence, Vol. 59, Mar. 2017, str. 186–195, dostupno na: https://linkinghub.elsevier.com/retrieve/pii/S0952197617300039

[121]Osadcha, O., Trzcionka, A., Pacho ńska, K., Pachoński, M., "Detection of Dental Filling Using Pixels Color Recognition", in Information and Software Technologies, Damaševičius, R., Vasiljevienė, G., (ur.). Cham: Springer International Publishing, 2018, str. 347–356.

[122]Yadav, G., Maheshwari, S., Agarwal, A., "Contrast limited adaptive histogram equalization based enhancement for real time video system", in 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Sep. 2014, str. 2392–2397.

[123]Otsu, N., "A Threshold Selection Method from Gray-Level Histograms", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, Jan. 1979, str. 62–66, conference Name: IEEE Transactions on Systems, Man, and Cybernetics.

[124]Zichun, Y., Qunfei, Z., Zisheng, T., Wenjun, X., "CBCT image segmentation of tooth-root canal based on improved level set algorithm", in Proceedings of the 2020 International Conference on Computers, Information Processing and Advanced Education, ser. CIPAE 2020. New York, NY, USA: Association for Computing Machinery, Oct. 2020, str. 42–51, dostupno na: https://doi.org/10.1145/3419635.3419654

[125]Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., "Pyramid Scene Parsing Network", in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, str. 6230–6239, iSSN: 1063-6919.

[126]Jader, G., Fontineli, J., Ruiz, M., Abdalla, K., Pithon, M., Oliveira, L., "Deep Instance Segmentation of Teeth in Panoramic X-Ray Images", in 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Oct. 2018, str. 400–407, iSSN: 2377-5416.

[127]Zhao, Y., Li, P., Gao, C., Liu, Y., Chen, Q., Yang, F., Meng, D., "TSASNet: Tooth segmentation on dental panoramic X-ray images by Two-Stage Attention Segmentation Network", Knowledge-Based Systems, Vol. 206, Oct. 2020, str. 106338, dostupno na: https://www.sciencedirect.com/science/article/pii/S0950705120304950

[128]Lee, S., Woo, S., Yu, J., Seo, J., Lee, J., Lee, C., "Automated CNN-Based Tooth Segmentation in Cone-Beam CT for Dental Implant Planning", IEEE Access, Vol. 8, 2020, str. 50 507–50 518, conference Name: IEEE Access.

[129]Ronneberger, O., Fischer, P., Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation", in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, ser. Lecture Notes in Computer Science, Navab, N., Hornegger, J., Wells, W. M., Frangi, A. F., (ur.). Cham: Springer International Publishing, 2015, str. 234–241.

[130]He, K., Gkioxari, G., Dollár, P., Girshick, R., "Mask R-CNN", in 2017 IEEE International Conference on Computer Vision (ICCV), Oct. 2017, str. 2980–2988, iSSN: 2380-7504.

[131]Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., "Path Aggregation Network for Instance Segmentation", in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, str. 8759–8768, iSSN: 2575-7075.

[132]Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Loy, C. C., Lin, D., "Hybrid Task Cascade for Instance Segmentation", in

2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019, str. 4969–4978, iSSN: 2575-7075.

[133]Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., Smola, A., "ResNeSt: Split-Attention Networks", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2022, str. 2736–2746, dostupno na: https://openaccess.thecvf.com/content/CVPR2022W/ECV/html/Zhang_ResNeSt_Split-Attention_Networks_CVPRW_2022_paper.html

[134]da Silva Rocha, E., Endo, P. T., "A Comparative Study of Deep Learning Models for Dental Segmentation in Panoramic Radiograph", Applied Sciences, Vol. 12, No. 6, Jan. 2022, str. 3103, number: 6 Publisher: Multidisciplinary Digital Publishing Institute, dostupno na: https://www.mdpi.com/2076-3417/12/6/3103

[135]Silva, G., Oliveira, L., Pithon, M., "Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives", Expert Systems with Applications, Vol. 107, Oct. 2018, str. 15–31, dostupno na: https://linkinghub.elsevier.com/retrieve/pii/S0957417418302252

[136]Zhu, H., Cao, Z., Lian, L., Ye, G., Gao, H., Wu, J., "CariesNet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic X-ray image", Neural Computing and Applications, Jan. 2022, dostupno na: https://doi.org/10.1007/s00521-021-06684-2

[137]Phulari, R. G., Textbook of Dental Anatomy, Physiology and Occlusion. JP Medical Ltd, Nov. 2013.

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

[139]ISO Central Secretary, "Dentistry — designation system for teeth and areas of the oral cavity", International Organization for Standardization, Geneva, CH, Standard ISO 3950:2016, 2016, dostupno na: https://www.iso.org/standard/68292.html

[140]Betul Oktay, A., "Tooth detection with Convolutional Neural Networks", in 2017 Medical Technologies National Congress (TIPTEKNO), Oct. 2017, str. 1–4.

[141]Keerthana, K. M., Rajeshwari, B., Keerthi, S., Menon, H. P., "Classification of tooth type from dental X-ray image using projection profile analysis", in 2017 International Conference on Signal Processing and Communication (ICSPC), Jul. 2017, str. 394–398.

[142]Chen, H., Zhang, K., Lyu, P., Li, H., Zhang, L., Wu, J., Lee, C.-H., "A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films", Scientific Reports, Vol. 9, No. 1, Mar. 2019, str. 3840, number: 1 Publisher: Nature Publishing Group, dostupno na: https://www.nature.com/articles/s41598-019-40414-y

[143]Ren, S., He, K., Girshick, R., Sun, J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", in Advances in Neural Information Processing Systems, Vol. 28. Curran Associates, Inc., 2015, dostupno na: https://papers.nips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

[144]Laishram, A., Thongam, K., "Detection and Classification of Dental Pathologies using Faster-RCNN in Orthopantomogram Radiography Image", in 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), Feb. 2020, str. 423–428, iSSN: 2688-769X.

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

[146]Tan, M., Pang, R., Le, Q. V., "EfficientDet: Scalable and Efficient Object Detection", in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, Jun. 2020, str. 10 778–10 787, dostupno na: https://ieeexplore.ieee.org/document/9156454/

[147]Kanopoulos, N., Vasanthavada, N., Baker, R., "Design of an image edge detection filter using the Sobel operator", IEEE Journal of Solid-State Circuits, Vol. 23, No. 2, Apr. 1988, str. 358–367, conference Name: IEEE Journal of Solid-State Circuits.

[148]Eldan, R., Shamir, O., "The Power of Depth for Feedforward Neural Networks", in Conference on Learning Theory. PMLR, Jun. 2016, str. 907–940, iSSN: 1938-7228, dostupno na: https://proceedings.mlr.press/v49/eldan16.html

[149]Hendrycks, D., Gimpel, K., "Gaussian Error Linear Units (GELUs)", arXiv, Tech. Rep. arXiv:1606.08415, Jul. 2020, arXiv:1606.08415 [cs] type: article, dostupno na: http://arxiv.org/abs/1606.08415

[150]Elfwing, S., Uchibe, E., Doya, K., "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning", Neural Networks, Vol. 107, Nov. 2018, str. 3–11, dostupno na: https://www.sciencedirect.com/science/article/pii/S0893608017302976

[151]Ramachandran, P., Zoph, B., Le, Q. V., "Searching for Activation Functions", arXiv, Tech. Rep. arXiv:1710.05941, Oct. 2017, arXiv:1710.05941 [cs] type: article, dostupno na: http://arxiv.org/abs/1710.05941

[152]Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T., "Visualizing the Loss Landscape of Neural Nets", in Advances in Neural Information Processing Systems, Vol. 31. Curran Associates, Inc., 2018, dostupno na: https://proceedings.neurips.cc/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html

[153]Tan, M., Le, Q., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", in Proceedings of the 36th International Conference on Machine Learning. PMLR, May 2019, str. 6105–6114, iSSN: 2640-3498, dostupno na: https://proceedings.mlr.press/v97/tan19a.html

[154]Keeler, J. D., "Basins of attraction of neural network models", AIP Conference Proceedings, Vol. 151, No. 1, Aug. 1986, str. 259–264, publisher: American Institute of Physics, dostupno na: https://aip.scitation.org/doi/abs/10.1063/1.36277

[155]Pan, H., Han, H., Shan, S., Chen, X., "Mean-Variance Loss for Deep Age Estimation from a Face", in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, str. 5285–5294, iSSN: 2575-7075.

[156]Long, J., Shelhamer, E., Darrell, T., "Fully convolutional networks for semantic segmentation", in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015, str. 3431–3440, iSSN: 1063-6919.

[157]Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., "Rethinking Atrous Convolution for Semantic Image Segmentation", dostupno na: http://arxiv.org/abs/1706.05587 ArXiv:1706.05587 [cs]. Dec. 2017.

[158]Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications", dostupno na: http://arxiv.org/abs/1704.04861 ArXiv:1704.04861 [cs] version: 1. Apr. 2017.

[159]Sun, K., Xiao, B., Liu, D., Wang, J., "Deep High-Resolution Representation Learning for Human Pose Estimation", in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019, str. 5686–5696, iSSN: 2575-7075.

[160]Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., TaoXie, Michael, K., Fang, J., Imyhxy, Lorna, Wong, C., Yifu, Z., V, A., Montes, D., Wang, Z., Fati, C., Nadar, J., Laughing, UnglvKitDe, Tkianai, YxNONG, Skalski, P., Hogan, A.,

Strobel, M., Jain, M., Mammana, L., Xylieong, "ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations", dostupno na: https://zenodo.org/record/3908559 Aug. 2022.

[161]He, K., Zhang, X., Ren, S., Sun, J., "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", in Computer Vision – ECCV 2014, ser. Lecture Notes in Computer Science, Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., (ur.). Cham: Springer International Publishing, 2014, str. 346–361.

[162]Wang, C.-Y., Mark Liao, H.-Y., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., Yeh, I.-H., "CSPNet: A New Backbone that can Enhance Learning Capability of CNN", in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Jun. 2020, str. 1571–1580, iSSN: 2160-7516.

[163]Smith, L. N., Topin, N., "Super-convergence: very fast training of neural networks using large learning rates", in Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, Vol. 11006. SPIE, May 2019, str. 369–386, dostupno na: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11006/1100612/Super-convergence--very-fast-training-of-neural-networks-using/10.1117/12.2520589.full

[164]Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., "Microsoft COCO: Common Objects in Context", in Computer Vision – ECCV 2014, ser. Lecture Notes in Computer Science, Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., (ur.). Cham: Springer International Publishing, 2014, str. 740–755.

[165]Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., "ImageNet: A large-scale hierarchical image database", in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2009, str. 248–255, iSSN: 1063-6919.

[166]Hu, J., Shen, L., Sun, G., "Squeeze-and-Excitation Networks", in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, str. 7132–7141, iSSN: 2575-7075.

[167]Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., Li, B., "Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs", dostupno na: http://arxiv.org/abs/2008.02312 ArXiv:2008.02312 [cs, eess]. Aug. 2020.

[168]Capitaneanu, C., Willems, G., Jacobs, R., Fieuws, S., Thevissen, P., "Sex estimation based on tooth measurements using panoramic radiographs", International Journal of Legal Medicine, Vol. 131, No. 3, May 2017, str. 813–821.

[169]Neves, J. A., Antunes-Ferreira, N., Machado, V., Botelho, J., Proença, L., Quintas, A., Mendes, J. J., Delgado, A. S., "Sex Prediction Based on Mesiodistal Width Data in the Portuguese Population", Applied Sciences, Vol. 10, No. 12, Jan. 2020, str. 4156, number: 12 Publisher: Multidisciplinary Digital Publishing Institute, dostupno na: https://www.mdpi.com/2076-3417/10/12/4156

[170]Chen, Q., Zhao, Y., Liu, Y., Sun, Y., Yang, C., Li, P., Zhang, L., Gao, C., "MSLPNet: multi-scale location perception network for dental panoramic X-ray image segmentation", Neural Computing and Applications, Vol. 33, No. 16, Aug. 2021, str. 10 277–10 291, dostupno na: https://doi.org/10.1007/s00521-021-05790-5

[171]Kim, C., Kim, D., Jeong, H., Yoon, S.-J., Youm, S., "Automatic Tooth Detection and Numbering Using a Combination of a CNN and Heuristic Algorithm", Applied Sciences, Vol. 10, No. 16, Jan. 2020, str. 5624, number: 16 Publisher: Multidisciplinary Digital Publishing Institute, dostupno na: https://www.mdpi.com/2076-3417/10/16/5624

# List of Figures

# List of Tables

# Biography

Denis Milošević was born in Böblingen, Germany, in 1993. He attended the High School of Natural Sciences and Mathematics in Osijek and then went on to study computer science at the University of Zagreb, Faculty of Electrical Engineering and Computing, where he got his master's degree in 2017 with his graduate thesis "Object Tracking in Video Sequences Using Deep Networks ."After three years of working on computer vision problems in the industry, in 2019 he enrolled in Doctoral Studies for computer science at the University of Zagreb, Faculty of Electrical Engineering and Computing, under the mentorship of Professor Marko Subašić, Ph.D., where he does research in the field of medical image analysis with deep learning. He is currently employed as a young researcher with the DATACROSS project of the Research Centre for Excellence in Data Science and Cooperative Systems. Along with his research project, he works as a teaching assistant for the courses "Information Processing," "Biomedical Image Analysis," and "Deep learning". His research interest includes computer vision, image processing, machine learning, and deep learning. He participated in the organization of several international conferences, workshops, and summer schools. He is an author or co-author of four journal publications and two conference papers.

## List of publications

### Journal papers

1. **Milošević, D.**, Vodanović, M., Galić, I., Subašić, M., "A Comprehensive Exploration of Neural Networks for Forensic Analysis of Adult Single Tooth X-Ray Images", IEEE Access, Vol. 10, July 2022, pp. 70980-71002, doi: 10.1109/ACCESS.2022.3187959
2. **Milošević, D.**, Vodanović, M., Galić, I., Subašić, M., "Automated Estimation of Chronological Age from Panoramic Dental X-ray Images Using Deep Learning", Expert Systems with Applications, Vol. 189, March 2022, pp. 116038, doi: 10.1016/j.eswa.2021.116038

3.Banjšak, L., **Milošević, D.**, Subašić, M., "Implementation of Artificial Intelligence in Chronological Age Estimation from Orthopantomographic X-ray Images of Archaeological Skull Remains", Bulletin of the International Association for Paleodontology, Vol. 14, No. 2, December 2020, pp. 122-129.

4.Banjšak, L., **Milošević, D.**, Subašić, M., Brkić, H., Vodanović, M., "Artificial Intelligence Implementation in Tooth Identification from X-ray Images", International Dental Journal, Vol. 71, September 2021, pp. S35, doi: 10.1016/j.identj.2021.08.007

## Conference papers

1. **Milošević, D.**, Vodanović, M., Galić, I., Subašić, M., "Automated Sex Assessment of Individual Adult Tooth X-Ray Images", in 2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA), September 2021, pp. 72-77, doi: 10.1109/ISPA52656.2021.9552124
2. **Milošević, D.**, Vodanović, M., Galić, I., Subašić, M., "Estimating Biological Gender from Panoramic Dental X-ray Images", in 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA), September 2019, pp. 105-110, doi: 10.1109/ISPA.2019.8868804

# Životopis

Denis Milošević je rođen 27.06.1993. u Böblingenu, Njemačka. Pohađao je Matematičku i prirodoslovnu gimnaziju u Osijeku, te je nakon toga studirao računarsku znanost na Sveučilištu u Zagrebu, na Fakultetu elektrotehnike i računarstva i elektrotehnike, gdje je diplomirao 2017. godine sa završenim radom "Praćenje objekata u videosnimkama pomoću dubokih mreža". Nakon tri godine rada na problemima računalnog vida u industriji, 2019. započinje doktorski studij pod mentorstvom prof. dr. sc. Marka Subašića, gdje se bavi analizom medicinskih slika pomoću metoda dubokog učenja. Trenutno je zaposlen kao Mlađi istraživač na DATACROSS projektu Znanstvenog centra za izvrsnost i znanost o podacima i kooperativne sustave. Uz istraživački projekt radi kao asistent na predmetima Obrada informacija, Analiza slika u biomedicini, i Duboko učenje. Njegovi interesi uključuju računalni vid, obrada slika, strojno učenje, i duboko učenje. Sudjelovao je u organizaciji nekoliko međunarodnih konferencija, radionica, i ljetnih škola. Autor ili su-autor je četiri publikacije u časopisima i dva konferencijska rada.