Universidade de São Paulo
FFCLRP - Departamento de Física
Pós-graduação em Física aplicada a Medicina e Biologia

LEONARDO FERREIRA MACHADO

Avaliação do risco de osteoporose, com foco na mandíbula, utilizando radiografia panorâmica dentária e modelos de inteligência artificial

Mandible-focused osteoporosis risk assessment using dental panoramic radiography and artificial intelligence models

Thesis presented to Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto of Universidade de São Paulo, as part of requirements for acquirement the grade of Doctor of Sciences, Area: Physics applied to Medicine and Biology.

Ribeirão Preto - SP
2023

LEONARDO FERREIRA MACHADO

Avaliação do risco de osteoporose, com foco na mandíbula, utilizando radiografia panorâmica dentária e modelos de inteligência artificial

Mandible-focused osteoporosis risk assessment using dental panoramic radiography and artificial intelligence models

Thesis presented to Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto of Universidade de São Paulo, as part of requirements for acquirement the grade of Doctor of Sciences.

Concentration area:

Física aplicada a Medicina e Biologia.

Advisor:

Luis Otavio Murta Junior.

Co-advisor:

Plauto Christopher Aranha Watanabe

.

Rectified version

Original version available at FFCLRP - USP

Ribeirão Preto - SP
2023

FICHA CATALOGRÁFICA

Machado, Leonardo Ferreira

Avaliação do risco de osteoporose, com foco na mandíbula, utilizando radiografia panorâmica dentária e modelos de inteligência artificial / Leonardo Ferreira Machado; orientador Luis Otavio Murta Junior, co-orientador Plauto Christopher Aranha Watanabe. Ribeirão Preto - SP, 2023.

103 f.:il.

Tese (Doutorado - Programa de Pós-Graduação em Física Aplicada a Medicina e Biologia) - Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo, 2023.

1. Osteoporose. 2. Avaliação de risco. 3. Radiografia Panorâmica Dentaria 4. Modelos de Inteligência Artificial

Nome: MACHADO, Leonardo Ferreira

Título: Avaliação do risco de osteoporose, com foco na mandíbula, utilizando radiografia panorâmica dentária e modelos de inteligência artificial

Thesis presented to Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto of Universidade de São Paulo, as part of requirements for acquirement the grade of Doctor of Sciences.

Approved in:____/____/____.

Examination Board

Prof. Dr. : _____ Institution:_____

Judgement:_____ Signature: _____

Prof. Dr. : _____ Institution:_____

Judgement:_____ Signature: _____

Prof. Dr. : _____ Institution:_____

Judgement:_____ Signature: _____

Prof. Dr. : _____ Institution:_____

Judgement:_____ Signature: _____

Prof. Dr. : _____ Institution:_____

Judgement:_____ Signature: _____

To my family.

Acknowledgements

Firstly, I want to give thanks to God, for He was with me guiding me in this unplanned and complicated, but fruitful journey.

Also, to my mother and family who never stopped believing in me and supporting me.

To my professor, advisor, and friend, Dr. Luis Otavio Murta Junior, PhD., for his endless patience and for the trust he put in my work. Definitely it made a huge difference.

To my co-advisor, Dr. Plauto Christopher Aranha Watanabe, PhD., the enthusiast behind the scenes. Without him this project would not be possible.

To my CSIM lab colleagues, specially Vivian and Jackson, for supporting me and being a valuable company during pandemic times.

To my friend Leonardo Vinícius, who was with me in the good and in the bad times during these 7 years of graduation studies. We had such a great time discussing science, politics, religion, music, and fun.

And to all the other friends whom I met here in Ribeirão Preto and who also supported me to continue my work and to do my best,

My sincere gratitude.

"Commit your way to the Lord;
trust in him
and he will do this."
Psalm 37,5.

Resumo

MACHADO, L. F. **Avaliação do risco de osteoporose, com foco na mandíbula, utilizando radiografia panorâmica dentária e modelos de inteligência artificial**. 2023. 103 f. Tese (Doutorado - Programa de Pós-Graduação em Física Aplicada a Medicina e Biologia) - Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto - SP, 2023.

A osteoporose é uma doença sistêmica que provoca perdas na densidade mineral óssea (DMO) que eventualmente causam fraturas ósseas graves. Por ser uma doença silenciosa, muitos são diagnosticados apenas após a fratura. Várias estratégias de diagnóstico oportunistas baseados em imagens estão sendo investigados na literatura em combinação com modelos de inteligência artificial (IA). O presente estudo propõe uma análise de imagens de radiografia panorâmicas odontológicas (PAN) focada na mandíbula usando modelos de inteligência artificial para avaliar o risco de osteoporose. Para este fim, desenvolvemos inicialmente uma ferramenta de segmentação automática da mandíbula para imagens PAN usando um conjunto de algoritmos de aprendizado profundo. Para desenvolver este modelo de segmentação mandibular, usamos dois conjuntos de dados: um conjunto de dados interno preparado com 393 imagens PAN anotadas manualmente por um especialista e um conjunto de dados públicos composto por 116 imagens previamente anotadas. As arquiteturas U-Net e HRNet foram consideradas individualmente e no formato de ensemble com e sem pós-processamento de segmentação. Com esta abordagem, alcançamos o melhor desempenho de segmentação mandibular na literatura com 98,2%, 97,6%, 97,2%, precisão, semelhança de dados e interseção sobre união, respectivamente. No segundo momento deste estudo, usamos esse algoritmo para

extrair a região de interesse (ROI) mandibular de 380 imagens PAN de pacientes que também realizaram exame de densidade mineral óssea (BMD). Esses pacientes foram organizados em dois grupos de acordo com os critérios de diagnóstico da OMS: saldáveis e risco de doença (osteopenia e osteoporose). Treinamos o modelo EfficientNetV2-L usando I) as imagens PAN completas como entradas e depois II) a ROI de segmentação da mandíbula como entrada para separar esses dois grupos. Observamos que o modelo usando a segmentação da mandíbula obteve melhor acurácia e recall (73,9% e 83,0%) do que os modelos treinados com a imagem inteira, o que indica ganhos consideráveis com o uso dessa abordagem focada na mandíbula.

Palavras-chave: 1. Osteoporose. 2. Avaliação de risco. 3. Radiografia Panorâmica Dentaria 4. Modelos de Inteligência Artificial

Abstract

MACHADO, L. F. **Mandible-focused osteoporosis risk assessment using dental panoramic radiography and artificial intelligence models**. 2023. 103 f. Thesis (Ph.D. - Postgraduate Program in Physics Applied to Medicine and Biology) - Faculty of Philosophy, Sciences and Literature, University of São Paulo, Ribeirão Preto - SP, 2023.

Osteoporosis is a systemic disease that provokes bone mineral density (BMD) losses that eventually cause severe bone fractures. Since it is a silent disease, many are diagnosed only after fractures. Several opportunistic image-based diagnoses are being investigated in combination with artificial intelligence (AI) models in the literature. The present study proposes a mandible-focused dental panoramic X-ray image (PAN) analyses using artificial intelligence models to assess the osteoporosis disease risk. To accomplish that, we initially developed an automatic mandible segmentation for PAN images using an ensemble of deep learning algorithms. To develop this mandible segmentation model, we used two datasets: an in-house dataset (IHD) prepared with 393 PAN images manually annotated by a specialist and a third-party dataset composed of 116 images previously annotated. U-Net and HRNet architectures were considered individually and an ensemble format with and without segmentation post processing. With this approach we achieved the best mandible segmentation performance in the literature with 98.2%, 97.6%, 97.2%, accuracy, dice similarity, and intersection over union, respectively. In the second moment of this study, we used this algorithm to extract the mandible image region of interest (ROI) from PAN images from 380 PAN images from patients who also underwent bone mineral density (BMD) examination. Those patients were organized into two groups according to WHO criteria for diagnosis: healthy (no

signs of osteoporosis) and disease risk (osteopenia and osteoporosis). We trained the EfficientNetV2-L model using I) the entire PAN images as inputs and II) the mandible segmentation ROI to separate these two groups. We observed that the model using the mandible segmentation achieved better accuracy and recall (73.9% and 83.0%) than the models trained with the entire image, which indicates considerable gains of using this mandible-focused approach.

Key-words: 1.Osteoporosis. 2.Risk Assessment. 3.Dental Panoramic Radiography. 4.Artificial Intelligence Models.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In 1985, Röntigen experiments using X-Rays allowed the creation of images from inside the body. It was such a breakthrough since it made possible to see and analyze internal parts of the human body without opening the patient. It also gave birth to other important fields of Physics and Medicine: Medical Physics and Radiology. Medical images appeared as an innovative source of information to analyze the human health status. Today, with the up rise of Artificial Intelligence (AI), we might be experiencing a comparable breakthrough: the combination of Medicine and AI to amplify the capabilities of the clinical practice.

In the last decades, we have experienced the appearance AI in a virtually every field of knowledge, science and technological application. The success of artificial intelligence has a direct correlation with an overall increasing of computational processing power at low costs. Another fundamental reason for that is the exponential increasing of the amount of data stored by our appliances in general. This scenario allowed the former gradient-based learning algorithms to became the so-called deep learning algorithms.

Medicine can certainly be heavily impacted by AI given its standardized approach to collect patient information (clinical data) and to store it for taking decisions. Such data have been accumulated over the decades in every clinical institution. This scenario is very favorable to the use of AI techniques to investigate different diagnosis, prognosis, analyses, and automation opportunities that was not possible before.

In the present study, we used AI to investigate alternatives to improve the diagnosis of a disease that is impacting worldwide as the life expectancy increases:

the osteoporosis. Osteoporosis is a silent disease that affects bone mineral density levels. Its silent action is the major explanation why the disease is mostly diagnosed only after a fraction occurrence. Those fractures can lead to comorbidity and even death in the long run. Osteoporosis majorly affects the elder population and is already seen as epidemic especially in developing countries, such as Brazil, where this population grows rapidly. This combination of late diagnosis and a rapid growth of the population at risk results in a growing socioeconomic impact for those countries.

As alternatives to prevent or alleviate the socioeconomic impacts of such epidemic, many diagnoses, prevention, and intervention strategies have been studied over the past decades. Most of them are focusing on the usage of medical images techniques already available in the ongoing clinical practice such as X-Rays, CT, MRI, Ultra-Sound, Dental Panoramic X-ray (PAN) and so on.

PAN images have gained special attention since it is a cheap image modality and largely available worldwide. Further, many studies have already found correlations between some oral structures and the osteoporosis diagnosis, being mandible the oral structure more useful to assess osteoporosis condition. For that reason, many studies have proposed artificial intelligence algorithms to assess bone health status directly from PAN images. Some of them have used the entire image, while others have focused on a rectangular shape regions of interest (ROI) covering the inferior cortical mandibular bone. None of them, though, have investigated the potential gain of using the *entire mandibular bone ROI* to assess the osteoporosis risk with deep learning experiments.

In this study, we present a mandible-focused osteoporosis risk assessment study where we used deep learning algorithms to automatically extract mandible segmentation ROI from PAN images and used it to predict the disease risk. This study is organized in five chapters, being the current introduction the first one. The second chapter brings a contextual introduction of the disease, from the clinical definition to its socioeconomic impact. Next, in the third chapter, we cover the foundations of the AI and the deep learning algorithms used. Also, we go over the published applications of AI to assess osteoporosis through PAN images as to have an understanding of the current state of the problem. The fourth chapter is dedicated to the development of a deep-learning-based automatic mandible segmentation tool.

This chapter is mostly written according to the published version of this study. The Fifth chapter brings the final experiments: the usage of entire PAN image and mandible segmentation only in a deep learning experimentation set to diagnose osteoporosis risk. These last two chapters have their own results, discussion and conclusion sessions that precisely cover all the goals of the study.

Chapter 2

Osteoporosis: clinical, socioeconomic, and practical aspects.

Osteoporosis is one of the diseases of the century. It causes a systematic damage to the skeletal structure that leads to several bone injuries. With high incidence in women over 55 y.o. and in man over 65 y.o., osteoporosis is already considered an epidemic by many authors. High osteoporosis incidence rates come along with high morbidity rates and high treatment costs, especially to developing countries, which are experiencing a rapid elderly population growth. For those reasons, lots of scientific effort and public health politics must be spent on improving osteoporosis diagnosis, related fracture prevention, and disease management. In this chapter, we will go over some basic concepts, diagnosis strategies and their limitations, as the new proposed approaches to enhance osteoporosis pre-treatment and fracture prevention.

2.1 The disease

Osteoporosis is a silent and chronic disease that affects the entire skeletal system making bones fragile and prone to serious injuries. Such injuries, in many cases, are the very first indication of the disease's presence. At the same time, it may also be an indication of how late the disease is being perceived since some fractures may cause serious morbidity and eventually death. Bone frailty is caused by a mineral loss process that inevitably leads to bone tissue micro-architecture damage.

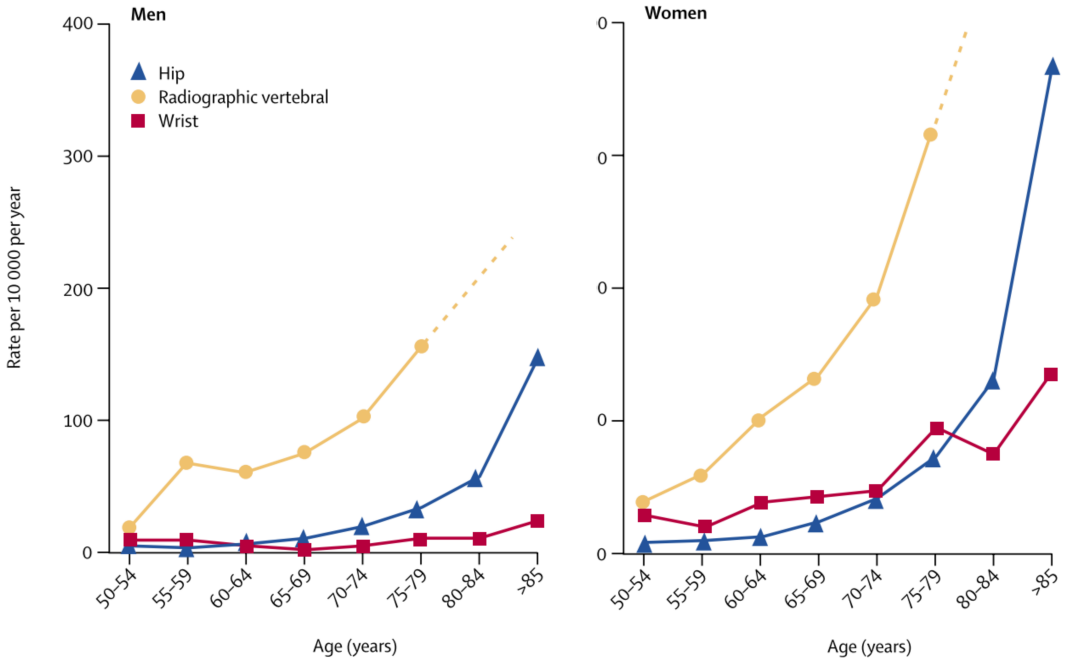

Figure 2.1: *Distal forearm, hip, and radiographic vertebral fracture incidence increase with age, for men (left) and women (right). Adapted from [Sambrook e Cooper 2006].* [1]

This deterioration process is also enhanced by an unbalance between bone mass decreasing rate and the body's bone-recovering capacity [Sambrook e Cooper 2006, Lorentzon e Cummings 2015, Sozen, Ozisik e Basaran 2017].

The incidence of osteoporosis is described by fracture incidence rates according to site, gender, age, and ethnicity. The most common fractures associated with osteoporosis are hip, vertebral, and wrist fractures. The fact that age is the most prominent condition causing an exponential increase of osteoporosis fracture incidence is widely reported in many studies (Fig. 2.1). It may happen given the natural bone content loss and an increasing number of falls associated with ageing. Additionally, studies also converge to a perceived higher incidence of osteoporosis related fractures (ORF) in white population and women (especially post-menopausal). In white population, over 50% of women and 20% of man, older than 50 years old, will experience some ORF [Sambrook e Cooper 2006, Christodoulou e Cooper 2003, Burge et al. 2007].

[1] In this analysis, it is counted only the vertebral fractures identified by imaging studies. Although it is known that over 85% of vertebral fractures do not come to medical attention

Hip fractures, although not the most common, are the most burdensome fracture type, with women being the majorly affected group. In women, the lifetime risk of hip fracture is greater than the lifetime risk of developing breast cancer. Also, 3% of the annual incidence of hip fractures occur in women over 85 years old. Next, incidence and prevalence of vertebral fractures are the highest, although only over a third of all vertebral fractures come to specialists' attention. Some of those fractures result from fall, but most result from routine activities. Their prevalence is the same in men and women. Lastly, wrist fracture, the third most common ORF. They happen mainly in women, being half of them older than 85 years old. In men, their incidence is low and there is no perceived increase with age, as we can verify in Fig. 2.1 (left) [Sambrook e Cooper 2006, Christodoulou e Cooper 2003].

The lifestyle burden experienced by those who suffer osteoporosis related fractures is undeniable. A study carried by WHO and the World Bank measured the disabilities and patient deaths incurred by many diseases allowing a comparison between their burden in Europe through a single loss index. This report showed that osteoporosis causes more losses than hypertension, Parkinson's, multiple scleroses, and all the cancer diseases considered, except lung cancer. Patients who undergo hip fractures, for example, need to be admitted to hospital with serious morbidness, disabilities, and even mortality possibly coming as a result of this fracture. Hip fractures has 10 to 20% mortality rate in the first year after the fracture, with highest death risk in the first 6 months. However, few of those deaths are directly credited to the fracture itself [Sambrook e Cooper 2006, Kanis et al. 2008].

The risk of fracture incidence, of almost all types, is considerably high in individuals with low bone density, notably, elderly people and those who suffer any bone mineral loss disease, such as osteoporosis. For adults who suffered a fracture, there is also a considerable risk of undergoing another fracture of a different type. Elderly people are the fastest growing group worldwide according to the age-adjusted population growth rates published in the 2017 United Nations Report on global population ageing (Fig. 2.2). If the age-adjusted hip fracture rate remains constant, some models predict that the number of hip fractures will rise from 1.7 million in 1990 to 6.3 million in 2050. It all adds to the conclusion that the overall ORF

[Christodoulou e Cooper 2003].

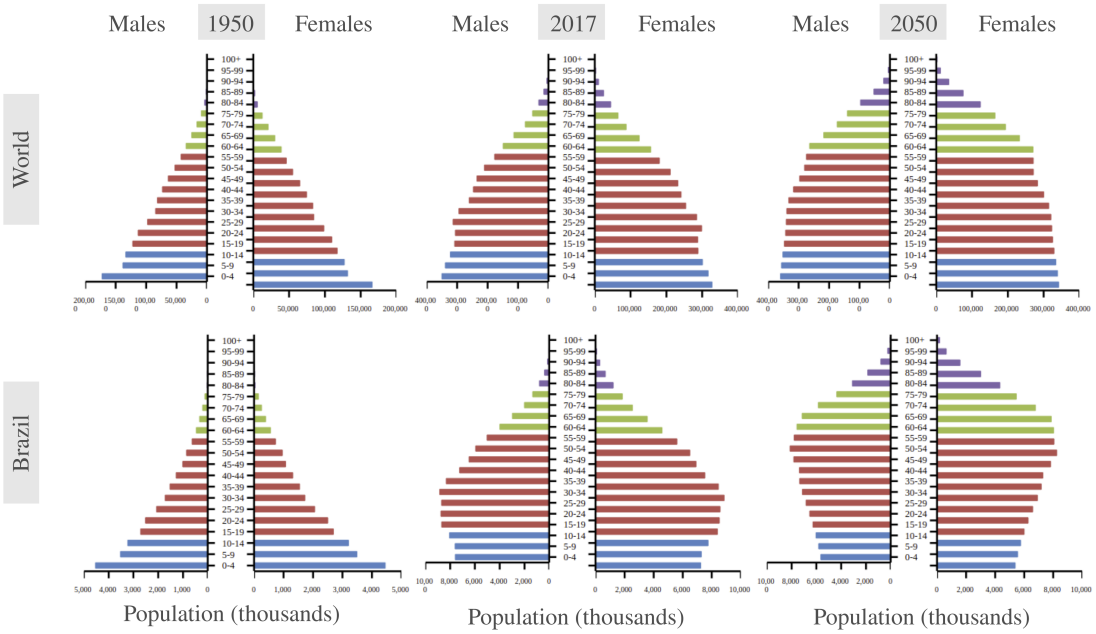

Figure 2.2: *World (top) and Brazil (bottom) Population pyramids estimated in 1950 and 2017, and the predicted one for 2050. Adapted from:* https://population.un.org/ProfilesOfAgeing2017/index.html *Accessed on April 23rd, 2021. The green and purple bars refer to the elderly group. It is possible to notice that those bars become larger from 1950 to 2050.*

numbers will increase in the coming decades and so the economic burden associated with that [Sambrook e Cooper 2006, Economic e Affairs 2017].

2.2 Socioeconomic burden

Many studies tried to evaluate the costs originated from osteoporosis fracture management in some countries, continents, and even in the entire world. Such assessment is complex since some other fractures, such as vertebral and wrist fractures, have not been economically evaluated and reported as much as hip fractures. This last one is the most burdensome ORF type. Hip fractures have much higher chances to provoke temporal or permanent disabilities, longer and costly hospitalizations, and to demand more aftercare than any other ORF. Henceforth, hip fractures related numbers are more available in the literature and many ORF economic reports emphasize particularly those fracture costs.

To glance at those expenses, some cost estimates related to ORF treatment

Country	Number of Fractures in 2018	Number of Fractures in 2022	Percentage increase (2018-2022)
Argentina	141,164	151,457	7.3%
Brazil	413,564	471,445	14,0%
Colombia	64,938	75,485	16.2%
Mexico	220,573	256,701	16.4%
Total	840,239	955,088	13,7%

Table 2.1: *Number of osteoporosis related fractures for four Latin American countries in 2018 and 2022. Adapted from [Aziziyeh et al. 2019]. Notes: 1 - These estimates were not adjusted to account for variations in population size. 2 - The numbers of fractures in 2022 are predictions of the study.*

and after-treatment will be listed. In 1990, the direct and indirect worldwide costs associated with hip fractures were estimated at US$ 34.9 *billion* and are expected to increase to US$ 82.7 billions in 2025, and to US$ 131.5 billions in 2050 [Johnell et al. 1997]. The 2005 estimate of the annual cost with ORF in United States of America was at US$ 17 billion, and it was expected to rise to US$ 25 billions in 2025. Canada spent over US$ 3.6 billions in 2014 with direct medical costs related to ORF. Lastly, that annual expense for 27 countries in the European Union in 2010 was estimated to be US$ 34.5 billion [Burge et al. 2007, Aziziyeh et al. 2019].

Additionally, a very detailed study was carried out by Aziziyeh et al (2019) to estimate the burden of osteoporosis in adults aged 50-89 y.o. in the four most populous Latin American countries: Brazil, Mexico, Colombia, and Argentina. The authors considered almost all the practical financial aspects present when someone suffers a fracture: patient productivity losses, drug and surgical treatment, examinations, and hospitalization costs. The study reports a total annual cost of osteoporosis related fractures for those four countries of US$ 1.17 billion in 2018. Brazil alone accounted for US$ 310 million. The total four country expenses are expected to increase to US$ 1,51 billion a year by 2026, if no intervention is taken, with Brazil reaching the mark of US$ 400 million a year.

Aziziyeh *et al* (2019) also reported the number of fractures registered in

the year 2018 and made predictions about the fracture number increase for each country in the year of 2022, as we can see in table 2.1. Brazil is the leading country in number of fractures (including hip) and Argentina the country with highest population adjusted ORF cost. Those numbers must certainly worry specially the developing countries (where Brazil lies), once their elderly population are the ones with highest increase rates observed. Larger elderly group means higher ORF incidence, higher health care investments, higher outpatient care demand, higher productivity losses, and larger osteoporosis socioeconomic burden [Christodoulou e Cooper 2003, Aziziyeh et al. 2019].

Odén *et al* (2012) assessed the overall number of hip fractures occurred in 58 countries in the year 2010. The authors computed 2.32 million fractures for all causes. From that amount, 50% were associated with osteoporosis and were considered avoidable if the patient's bone mineral densities were corrected by preventive therapies [Odén et al. 2015]. That being said, it is of paramount importance to assess the disease at an early stage and introduce fracture preventive therapies. That certainly sounds obvious. But why is it not that simple to achieve that now? In the next session, we will discuss the key element for all that: the diagnosis.

2.3 Diagnosing the unseen

Since 1994, the principal measure for assessing bone mineral quality and for diagnosing osteoporosis has been the bone mineral density (BMD, the amount of bone mass per volumetric or area unit). The BMD's role in the osteoporosis assessment is like blood pressure when diagnosing vascular diseases. Although, BMD is not the only factor to determine the presence of osteoporosis or osteopenia, BMD information is also largely used as a powerful biomarker for fracture risk assessment and for monitoring treated and untreated patients. Nowadays, many other disease and fracture risk correlated features, such as age and gender, are used to improve osteoporosis management [Kanis et al. 2008, Compston, McClung e Leslie 2019].

The gold standard evaluation for assessing BMD is the Dual Energy X-ray Absorptiometry (DXA) because it is very sensitive to calcium, which bone is

the most important source. In this examination, some radiation doses are used to assess the bone density from peripheral skeletal parts and from the whole skeleton. DXA measures *areal* BMD (g/cm^3), information that explains over two-thirds of bone strength variance. Additionally, DXA measurements can be performed at different sites of human skeleton and yield different BMD contents, allowing different conclusions about patient's bone quality. Alongside DXA, some other BMD assessment tools with comparable performances, such as quantitative ultrasound and quantitative computed tomography, were developed and are sometimes considered. However, those alternative techniques are not adopted as much as BMD by the various geographic-specific guidelines [Kanis et al. 2008, Compston, McClung e Leslie 2019].

Two indexes are derived from DXA BMD measurements: **T-score** and **Z-score**. T-scores is the number of standard-deviations (SD) that a given subject's BMD differs from the *healthy young-adult women* BMD average value. It means the difference between the BMD reference value and the subject's BMD is measured in SDs. This SD is the standard deviation value of the reference population BMD distribution. Z-score, similarly, is the number of SDs that separates a subject's BMD from the average BMD of a population with *same age and sex*. In 1994, WHO defined T-score criteria to diagnose osteoporosis in postmenopausal women and in men over 50 y.o. according to central DXA T-score results. Table 2.2 brings BMD, T-score values, and the associated diagnosis [Kanis et al. 2008, Compston, McClung e Leslie 2019].

WHO definitions expressed in table 2.2 appear to be very straightforward for diagnosing osteoporosis or assessing risk of fractures. However, many subjects experience ORF even having T-scores out of those predefined ranges, meaning high specificity, but low sensitivity. Another important aspect of osteoporosis is that it is an underdiagnosed disease. Many have, but only a few know. And from those who know, most got to know it only after an ORF. This happens mainly because osteoporosis is a silent disease, i.e., it shows no visible signs until a fracture occurs. In general, osteoporosis diagnosis tends to be verified only after fracture occurrence, in its late stage. Additionally, DXA equipment scarcity contributes to osteoporosis underdiagnosis, since DXA is the principal mean for BMD assessment,

Diagnosis	Subject BMD compared to reference value (rv)	Interpreted T-score
Healthy	BMD $\geq$ (**rv** - $1 * SD$)	higher than -1
Osteopenia	BMD $<$ (**rv** - $1 * SD$) & BMD $>$ (**rv** - $2,5 * SD$)	between -2.5 & -1
Osteoporosis	BMD $\leq$ (**rv** - $2,5 * SD$)	less than -2.5
Severe osteoporosis	BMD $\leq$ (**rv** - $2,5 * SD$) with one or more frailty fractures.	less than -2.5 with fractures

Table 2.2: *WHO bone health indications according to T-scores using as reference healthy young women average BMD [Kanis et al. 2008]. **BMD**: Bone Mineral Density of a subject; **SD**: Standard Deviation; **rv**: average BMD of the reference population - healthy young adult women.*

the major indicator of the disease according to WHO. Considering all that, additional strategies have been considered to improve osteoporosis management [Compston, McClung e Leslie 2019, Kanis 2007].

2.3.1 Screening and risk prediction

After many estimations of the economic and social burden reported in successive studies across the world, attention was drawn to the development of diagnosis, prevention, management, and surveillance solutions. Probably, the easiest solution was to prescribe preventive treatment for everyone in a risk group, such as menopause women or elderly people. Test everyone on that risk group with DXA and use WHO criteria. However, those simple solutions do not compose the set of methods currently in use [Kanis 2007].

Screening is considered effective only when high risk subjects can be selected. BMD/DXA screening on groups like women at menopause and elderly people, appear to be appealing since bone mass density knowledgeably decay in such scenarios, making them a reasonable target group for BMD-based screening. Besides the high costs for screening, BMD assessment is not indicated for large-scale screening since it has low sensitivity (WHO criteria) and because DXA is not largely available in most countries. Additionally, menopause patients have shown low continuance

with treatment. It all adds up to a correspondingly low return on investments for screening women in menopause. With elderly people, some interventions are considered even without screening, for every elderly subject, given the fracture risk exponential growth with age. However, such procedures are not widely adopted because the criteria and the screening method vary from guideline to guideline [Kanis 2007].

In this scenario, *opportunistic screening* and *case-finding* strategies appear as alternatives to large-scale screening. In this case, patients are verified for presenting some risk factors and then guided for further assessment. To support that strategy, two groups of tools have been developed: *Osteoporosis prediction* and *fracture risk assessment* tools. Those tools have been considered to decide which patient should go for BDM evaluation, preventive treatment, or even if the subject have osteoporosis. Using such tools seemed to be more precise and efficient than gross screening when it comes to finding high risk patients. Some reports support that this approach saves over 55% BMD tests by dismissing those who do not need further attention [Kanis 2007].

Osteoporosis prediction tools (OPT) have focused on diagnosing the present risk of osteoporosis. Patients found with high risk would be directed to BMD examination so it could be verified if positive or not. Those tools combine several types of risk factors and output a *score* that indicates the presence of osteoporosis. BMD/T-score comes after that evaluation to confirm or dismiss the diagnosis. OPT tools like ABONE, DOEScore, HAQ, NOF, ORAI, OSIRIS, OST(A), POST, SCORE, and SOFSURF were developed focusing specifically high sensitivity results (precisely identifying those patients who *do not* have the disease). Those tools reported high sensitivities ranging 78 to 100%, and low specificities ranging 18 to 58%. The low specificity indicates the necessity of BMD assessment (high specificity test) of those patients pointed to have high risk of disease to close the diagnosis [Kanis 2007].

2.3.2 Fracture risk assessment tools

Fracture risk assessment (FRA) tools compose the other group of tools developed to improve osteoporosis general management. Those tools were designed

to assess the osteoporosis main outcome, the fracture, rather than the presence of the disease itself. Their authors considered assessing such risk more demanding and achievable than closing a precise diagnosis of the disease. in those tools, BMD measurements may or may not be present, it comes in as any other risk factor. Many of those well-validated fracture risk factors have been combined to produce useful subject fracture risk detection tools. Those tools intend to establish practical thresholds for defining risky subjects, dismiss healthy ones, speedup prevention, and improve current group or geographic-specific intervention guidelines [Kanis 2007].

The most widely established tool for FRA is known as FRAX (Fracture Risk Assessment Tool, figure 2.3) with over 33,5 million fracture risk assessments performed up to the moment and acknowledged by WHO [Kanis]. This tool has been incorporated into clinical practice in many countries especially because it allows calibrations by including target population specific fracture rates and overall mortality index. FRAX considers age, sex, body-mass, previous fracture, glucocorticoid usage, secondary osteoporosis, rheumatoid arthritis, parental hip fracture, cigarette smoking, alcohol intake, and femoral neck BMD or T score (optional). That information was gathered over many meta-analysis, reviews, and nine international prospective cohorts. FRAX authors combined all those parameters in a risk assessment function that yields two highly valuable prognostic info: 10-year major osteoporotic fracture (clinical vertebrae, hip, forearm, proximal humerus) and 10-year hip fracture risk [Compston, McClung e Leslie 2019, Kanis 2007, Kanis].

Additionally to FRAX, there is the Garvan Fracture Risk Calculator (GFRC) and QFractureScores-2016 (QFS) tools which were validated with at least one independent cohort. GFRC, differing from FRAX, necessarily needs BMD measures, plus 5 other risk factors, to output 5 or 10-year risks for general and hip fracture occurrence. In the other hand, QFS do not rely on BMD assessments whatsoever, but demands 30 clinical/demographic information, most of them being disease related, to predict 1-10-year risk for hip and general fracture. Many other cohort studies were carried out to evaluate combinations of risk factors for predicting fracture risk (non-osteoporotic fracture, ORF, and hip fracture), however most

[1]

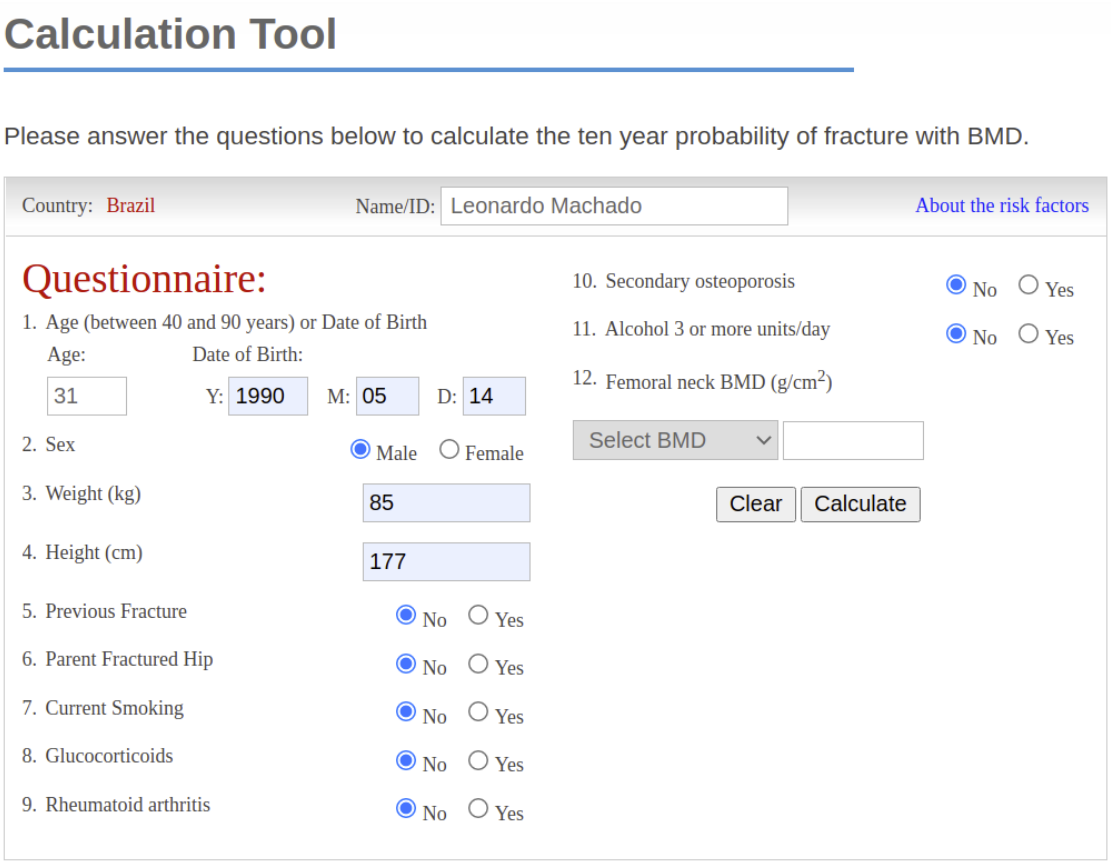

Figure 2.3: *Screenshot from FRAX questionnaire at [Kanis]. FRAX tool is in fact a questionnaire where one can fill clinical and demographic information in combination or not with BMD measures. Note: FRAX can only make predictions for people 40 y.o. and over.*

of them could not be evaluated and validated in different cohorts and some are case-control studies [Compston, McClung e Leslie 2019, Kanis 2007].

FRAX indeed offered a useful tool for FRA with a wide range of applicability given the wide variety present in the cohorts used. However, some validations are yet to be carried out for men and some ethnic group not covered by the study so far. The FRAX validations results also pointed out that FRA can be carried without BMD measures, which is useful especially in places with no DXA equipment in their facilities. Nevertheless, the presence of BMD can improve the performances assessed by gradient risk and receiver operating characteristic (ROC) curve [Kanis 2007].

2.4 Current image-focused trends to improve osteoporosis management

Alongside those described achievements, other research efforts have been put to develop cheaper, direct or indirect, methods to assess BMD and to identify other promising osteoporosis or fracture predictors, hence, improving both the diagnosis and fracture risk assessment. Many of those efforts have relied on various imaging techniques, from various parts of the skeleton, searching for possible radiological markers [Cruz et al. 2018].

Heel X-ray scans were used for separating osteoporotic from control patients [Harrar et al. 2012]; Trabecular features calculated from right femur X-Ray images showed notable accuracy when separating normal and risky from diseased patients [Sapthagirivasan e Anburajan 2013]; Mandible cortical width measured from dental panoramic X-ray were used to diagnose low spinal and femoral neck BMD [Kavitha et al. 2013]; Joint hip ROI extracted from regular hip radiographs were combined with clinical features to improve osteoporosis diagnosis [Yamamoto et al. 2020]; Image texture features extracted from spinal CT were used for separating osteoporotic (with fracture) from non-osteoporotic (without fracture) patients [Valentinitsch et al. 2019]; Mechanical and topological Bone measurements extracted from quantitative MRI scans from femoral neck were evaluate for predicting osteoporosis related fractures in post-menopausal women [Ferizi et al. 2019]. Some of those images are illustrated in figure 2.4.

Those studies point out that imaging techniques will play a remarkable role in osteoporosis management. Not just for detecting fracture, where using images are imperative, but also for predicting fractures and osteoporosis diagnosis itself. Additionally, all the cited studies' methodology relies heavily on a set of trending mathematical, statistical, and technological tools called artificial intelligence (AI), which is one or the cornerstones of the study being presented, and which will be covered in the next chapter.

The medical images just referenced sound very attractive for tackling osteoporosis diagnosis and fracture prediction tasks for two reasons: first, almost

1

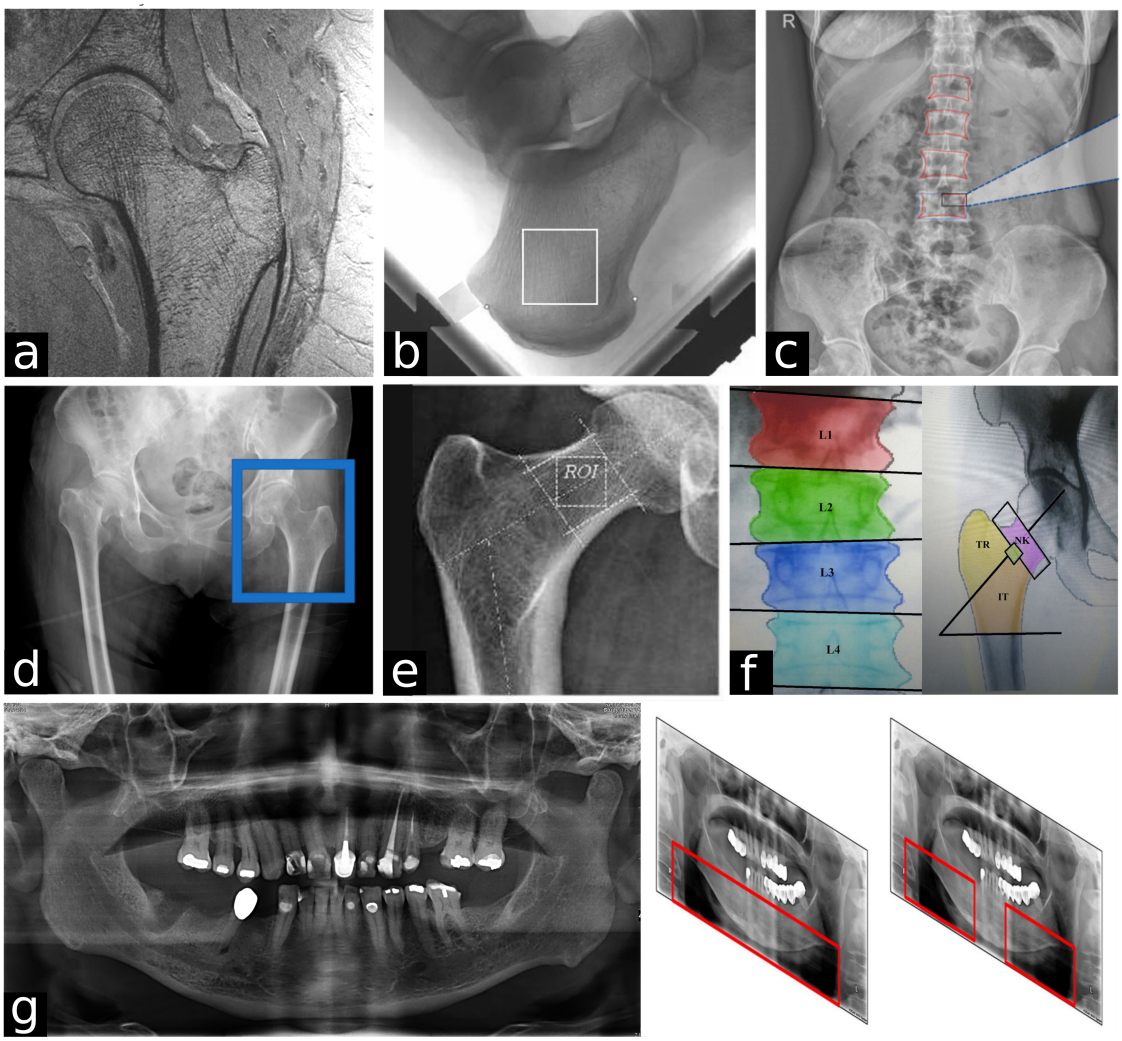

Figure 2.4: *A set of examples of images and sites currently being investigated for improving osteoporosis management. In a) Femural quantitative MRI for topological analysis [Ferizi et al. 2019]; b) Calcaneous X-ray used for texture analysis [Harrar et al. 2012]; c) Lumbar spine X-ray ROI used in deep learning [Zhang et al. 2020]; d) Hip joint area radiograph used in deep learning [Yamamoto et al. 2020]; e) Femural neck X-ray for extracting trabecular features [Sapthagirivasan e Anburajan 2013]; f) DXA produced images for radiomics extraction [Rastegar et al. 2020]; and g) Dental panoramic radiographs used in deep learning [Lee et al.].*

every health care facility in the world have an imaging equipment available; and second, patients regularly take those records as follow-up routine for a variety of abnormalities. Across all those image types and imaged sites mentioned above, we will now pay a especial attention to the imaging modality and site that composes the core of the present study: the dental panoramic radiography and the mandible

bone.

2.4.1 Dental panoramic radiography (PAN) in osteoporosis management

From all those images mentioned earlier, dental panoramic radiography (PAN) shows some remarkably interesting advantages. Firstly, PAN scan equipment is widely available, it includes developing and poor countries where DXA equipment is considerably rare given its high cost. Brazil is the country with the largest amount of PAN scans in the world. Further, PAN images are largely used for dental primary screenings and routine examinations, and it is even promoted by the International Guide to Prescription Radiographs. This fact makes PAN a perfect candidate source of information for opportunistic screening of bone related conditions. MRI, CT and regular X-ray imaging modalities, or even DXA, do not share all those qualities altogether [Watanabe, Watanabe e Tiossi 2012].

The relationship between age, systemic osteoporosis, and changes in quantity and quality of maxilla and mandible bones has been already reported in the literature. PAN images, beyond teeth, catch information from both mandible and maxilla bones alongside other important facial supporting structures. Although, its main usage had been tooth evaluation, in the last decades, it has been used to evaluate bone density quality. Since it pictures bone structures, it can reveal bone radiolucency, thinning and erosion and, for that reason, has been investigated for assessing bone mineral density loss [Watanabe, Watanabe e Tiossi 2012]. Additionally, BMD, a major factor for detecting osteoporosis, was already measured through periapical intraoral radiographs using aluminum densitometric scale and is already used in a commercial system [WATANABE et al. 2008]. Those finds enforce the expectation of the contributions of PAN images for osteoporosis management.

Klemetti (1994) proposed a classification of bone quality based on three indices measured over patterns observed on endosteal margin of mandibular cortical [Klemetti, Kolmakov e Kröger 1994]. After Klemettis's study, several other research proposed additional radiomorphometric indices for that purpose [Watanabe, Watanabe e Tiossi 2012]. Fractal analysis has been used to evaluate morphological patterns of jawbones overtime and more recently has been

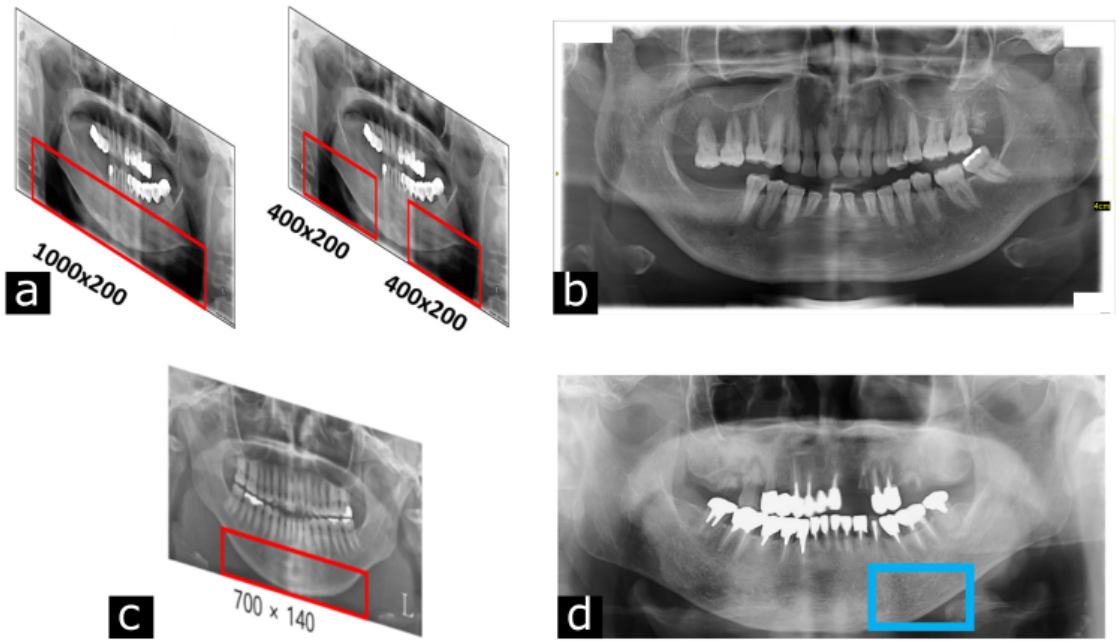

Figure 2.5: *ROI used to leverage mandible bone on PAN images in AI-based studies.*

considered for separating healthy from osteoporotic patients [Franciotti et al. 2021]. Radiomics, has been correlated with bone microarchitectural features in CT and radiographic images for assessing bone quality and osteoporosis screening [Valentinitsch et al. 2019, Ollivier et al. 2013]. Another, recent study proposed a PAN-mandible-based image index (W-Index), which is calculated using the oblique line and the mandibular ramus image regions, to separate different BMD level patient groups achieving prominent results [Watanabe et al. 2022]. Those studies precisely point out the value of mandible analysis for assessing osteoporosis.

Furthermore, some studies have performed artificial intelligence analysis over dental panoramic images for detecting osteoporosis [Lee et al., LING e YANG 2020, Lee et al. 2020, Sukegawa et al. 123]. Since mandible is the oral structure in the PAN image containing the most valuable information for assessing osteoporosis condition, those studies tried different strategies to leverage the mandible bone image region. They used either the entire PAN image (Figure 2.5 b)), a rectangular ROI containing the inferior cortical mandible bone (Figure 2.5 a) and c)), or a rectangular shape containing only the left inferior cortical bone (Figure 2.5 d)). Those studies

have achieved considerably high performances (up to 84.0% accuracy) but with room for improvements. Furthermore, no study have used the entire mandible outline as ROI for training their deep learning algorithms. This was not possible certainly because it is a time-consuming procedure for a large number of images. Further, there is no such tool for automated mandible segmentation openly available to make this task feasible.

In this study, we present an osteoporosis risk assessment trial using artificial intelligence algorithms trained over the mandible-only PAN image ROI. To that end, we firstly developed a deep-learning-based mandible bone segmentation model for PAN images. Then, we used this model to extract the mandible ROI and used it to train deep learning algorithms to predict the osteoporosis risk and compare this approach with using the entire PAN image for the same task. In the next chapter, we provide a deep explanation on the artificial intelligence existent strategies and the deep learning concepts, how to train a model and how to improve it. Then, In Chapter 4, we will present the entire process to develop the automatic mandible segmentation algorithm that we proposed to leverage PAN mandible information for osteoporosis risk assessment. Lastly, in Chapter 5, we present the final AI experiments that evaluate the usage of a mandible ROI PAN image strategy for the osteoporosis risk assessment task.

Chapter 3

Artificial intelligence in Medicine

Medicine is certainly one of the fields of human knowledge that most need constant growth. Although man has an enormous understanding about the human body and its relationship with external agents (drugs, radiation, temperature, ...), tons of questions remain open. From primary care, analysis of early symptoms, up to tertiary and quaternary care, with tumor grading and recurrence prediction. Indeed, the available knowledge is far more incomplete than we usually imagine and demands constant improvement.

With the appearance of medical imaging techniques, e.g., X-ray in 1985, Medicine took a huge step in its treatment power. It became capable of evaluating the inner conditions of the human body without opening it. Later, it became the field known as Radiology. At a comparable level, at this very moment, we are experiencing a new breakthrough in Medicine: the prominent usage of Artificial Intelligence (AI), a set of computational methods and algorithms, to help diagnose and predict outcome as well as to personalize treatment decisions in the clinical practice. Further, the development of medical technologies powered by AI is giving birth to a new field called augmented Medicine, i.e. the improvement of clinical practice with intelligent medical devices [Briganti e Moine 2020]. In the coming topics we will be going into some fundamental concepts on AI and its two principal trends: Machine Learning and Deep Learning.

3.1 What is artificial intelligence?

After more than 60 years of development, Artificial Intelligence, a.k.a. AI, is now a reality. Many systems surrounding us make use of such technology and many more are yet to come. But what is AI after all? According to the first researcher to name it in 1955, John McCarthy, it is about developing machines that behaves as if they were intelligent. Behind this definition is the perception that an outstanding intelligence is one of the remarkable abilities that separates humans from the rest of beings. Making human-like machines meant making human-like intelligent machines. Human intelligence is associated with the capability of learn by repetition, perceive underlying patterns on things, scenes, behaviors and sounds, knowledge transposition, deduction, intuition, and logical elaboration. Hence, a system said to be intelligent should have those capabilities.

In the search to achieve AI systems (computer algorithms) many strategies took place such as understanding how brain works and mimicking it, optimization of problem-solutions, and comprehending high level human thinking through cognitive sciences. The developed AI systems share some important aspects: they learn from data, the more data the more accurate the solution; the algorithms have a huge versatility, they are not problem specific; learned knowledge can be transferred from task to task; many learn strategies became possible given the relentless growth in computational power.

As it grown AI became a term to refer to a diverse collection of computational methods that are applicable to a even large variety of activities. The large areas of AI include natural language processing, i.e., identifying, reading, and classifying human written language; voice recognition, i.e., processing and classifying audio signals; image recognition, i.e., patterns, objects, scenes, and faces detection and classification on digital images; and many others including Machine Learning and Deep Learning, the AI techniques used in this study. Fig. 3.1 represents a conceptual map that points out the principal areas and branches of artificial intelligence.

With such a spectrum of techniques and approaches, AI spread in our daily lives through several applications like internet search engines, virtual secretary, intelligent recommender systems (for songs, movies, TV series, products, scientific

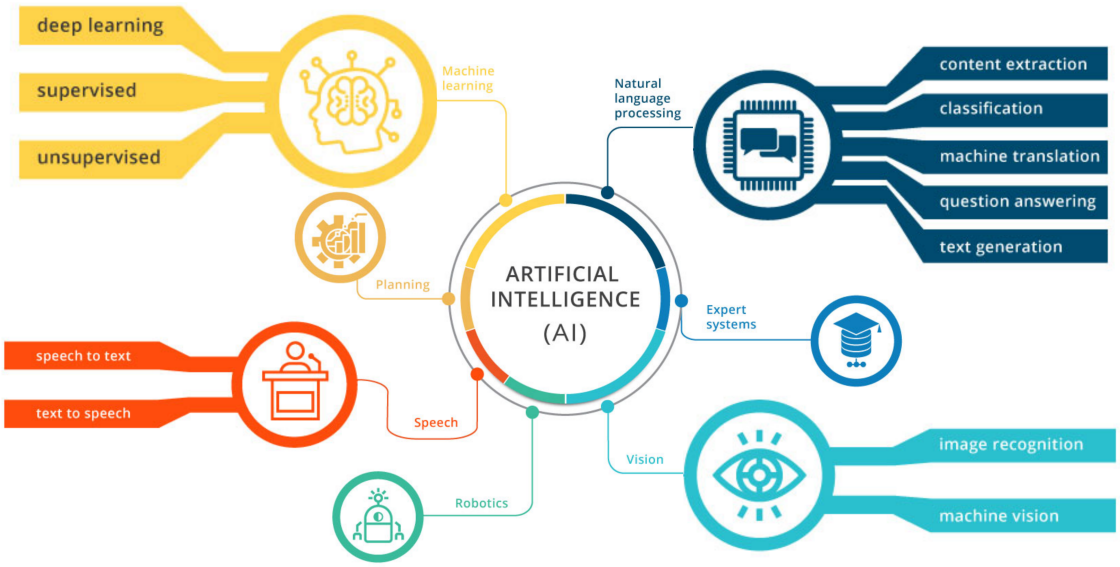

Figure 3.1: *AI areas and branches scheme. Adapted from (https://shorturl.at/iyDX5).*

papers, etc.), face recognition, and even the futuristic self-driving cars.

3.1.1 Why and how AI is in Medicine?

In Medicine it is not different, artificial intelligence is gaining relevance and acceptance as it proves its power and versatility to solve or help to solve a large variety of clinical demands. AI medical applications intend to improve diagnosis and prognosis, reduce disease complications, offer less invasive assessments, reduce repetitive physician work, and reduce hospitalization length. To mention a few, cardiology, pulmonary medicine, endocrinology, neurology, and radiology, are examples of fields in Medicine that is currently making use of AI solutions [Briganti e Moine 2020, Hameed et al. 2021, Chan et al. 2018].

The ongoing technological boom is certainly one of the reasons for this escalating AI usage in medicine. For example, the popularization of wearable technology (e.g., smart watches, wristband, etc.) that generates on-time patient data (e.g., blood pressure, body temperature, heartbeat, etc.) allowed the patient monitoring field to keep adhering to AI solutions. Such patient-data can be analyzed by AI algorithms present in the devices itself or in the cloud, and produce hourly or daily health reports to be evaluated by health practitioners [Briganti e Moine 2020].

Next, we must acknowledge the huge breakthrough AI caused and continue to cause in Radiology and Medical Imaging Analysis in general. By this we refer to cancer and other diseases diagnosis through image (computational tomography (CT), Magnetic Resonance Imaging (MRI), X-ray, ...), histopathology imaging diagnosis, detection of diseased or anomalous anatomy, and automation of time-consuming (manual, user-dependent) tasks such as image segmentation, registration, fusion, and classification. In fact, radiology is a rich soil for AI methods. This happens because image data acquisition and storage is mandatory in clinical radiology and abundant data is the foundation necessary for AI methods to succeed [Hameed et al. 2021].

In the future sections we will go over the practical details of learning algorithms, Machine Learning, Deep Learning, and the principal applications of those techniques as computer aided diagnosis systems.

3.2 Learning algorithms

We can understand the concept of learning with a quite simple example. Imagine you want a friend of yours to recognize your mother at the supermarket. You start to write down many traits and features of her face, and he will try to figure out in his mind which traits he should focus more to find your mother. What a challenging task, is not it? A way around would just give him some pictures of your mother. Your friend would instantly elect some key features that could help spot your mother at the supermarket and would do that without much trouble. This is learning. Learning by examples rather than by clearly stated rules.

When an algorithm can reproducing a decision after "viewing" some examples of the decision being made instead of being taught the rules explicitly, we can say that the algorithm is learning. This is the intelligent behavior we expect algorithms to achieve in machine learning. Most of the tasks that algorithms are classically set to learn are tasks that demand a lot of experience to execute it such as house pricing, credit card fraud detection, insurance calculation, cancer classification, disease diagnosis, face recognition, etc.

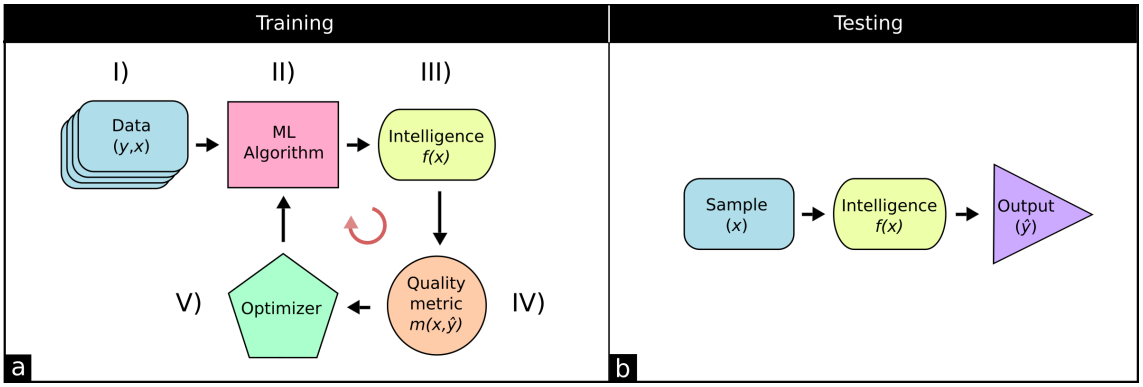

Figure 3.2: *Machine learning experiment stages (training and testing) and elements (a).*

3.2.1 Machine learning

Machine Learning (ML) is a subfield of AI that covers the learning algorithms that can learn *tasks* by *examples*. Let us consider the house pricing example. In this case the task is giving the right price to a house. The examples are composed of set of house features (constructed area, number of bedrooms, number of floors, etc.) and the price for this house according to a specialist. In the case of cancer classification, the task is diagnosing the cancer type. The examples are composed of a set of cancer features (tumor size, location, cell type, etc.) and the cancer classification for those features.

The ML experiments occur in two stages: *training* and *testing* (Figure 3.2). Training is the process where the algorithm will extract intelligence from data examples, it will fit a mathematical function $f(x)$ to explain the data examples, Figure 3.2, a). Testing is the stage where the quality of the final learned function $f(x)$ is tested over unseen data examples, Figure 3.2, b). The ML experiments are basically composed by I) a dataset (data examples), from which the algorithm will learn; II) the ML algorithm itself; III) the intelligence, which is the function $f(x)$ with its weights, the mathematical function that is being fitted over the data examples during training stage; IV) a quality metric, that measure the quality of fit of the function so far fitted; and V) the optimizer that changes the intelligence function weights towards a higher quality metric, Figure 3.2, a). Some ML experiments may show some variance from this structure given some problem-specific subtleties, but in general those elements are always present.

3.2.1.1 Learning paradigms

ML algorithms are characterized according to the learning process used to train it. *Supervised Learning* is the learning process where it is given labeled examples to train the algorithm. A labeled example, in the case of cancer classification, is the one that contains the cancer *features* and the cancer *classification*. *Unsupervised Learning* is the learning style where the algorithm does not have the labels for each example, only their features. This learning style is used generally when the task is not classifying a given sample, but finding an underlying classification pattern. In the case of cancer classification, it would mean having many cancer cases, each one with its specific set of features, but without knowing what class each one of them really belongs to. The algorithm aim would be *to group* or *to cluster* those cancer samples in groups that could reflect an underlying classification, i.e., would be finding a meaningful classification metric for the examples. *Transfer Learning* is another learning style that has became extremely popular. This modality of learning used when there is little example data to train a ML model in a given task, for example car picture classification. Then, it is used a ML model previously trained in a different dataset, e.g., ImageNet, and different tasks, e.g., general image classification, to solve the intended task, car picture classification. This is possible when the ML model used in both tasks are the same and the tasks share some similarities. *Reinforcement Learning* is the fourth learning pillar of ML. Reinforcement learning sets the learning algorithm to work in a trial-and-error approach under a reward-penalization system where it reinforces the desired decisions or behaviors of the algorithm and penalizes undesired solutions. This game-like setting appears as a suitable alternative where the algorithm would need an infinite number of "if-then" statements to apprehend the desired behavior, i.e., vehicle self-driving.

3.2.1.2 Machine learning tasks

Another way of characterizing ML algorithms is according to the style of the task it is being tried to solve with ML. Before we introduce the types of tasks ML is used to solve it is necessary to state that ML training and testing are entirely mathematical (i.e., numerical) process. In other words, for every problem one

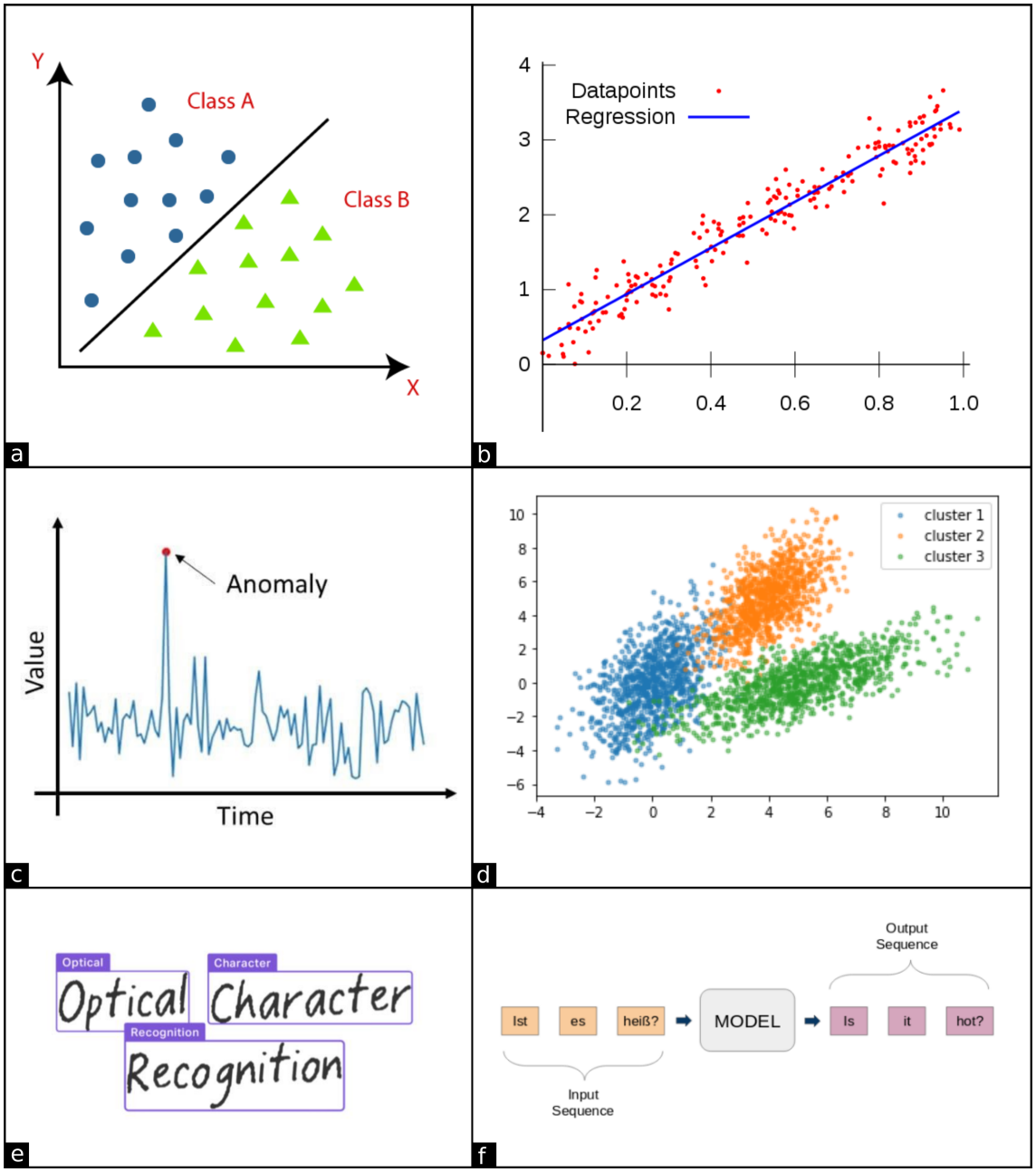

Figure 3.3: *Classical machine learning tasks.a) Classification, b) Regression, c) Anomaly detection, d) Clustering, e) Transcription, and f) Machine translation.*

intends to solve with ML it is necessary to codify the problem into numbers because all the features and information from the examples will be used inside a numerical equation $y_i = f(x_i)$. Where y_i is the *ith* output generated by the algorithm, f is the ML algorithm, and x_i is the vector of quantitative features that describe the sample or example *ith*.

Classification is one the most common ML tasks. The output y of the ML algorithm is categorical. In the learning process, the ML algorithms strive to find a logical/numerical decision boundary in the feature space. After finding such boundary any new sample will be classified according to it. Example: Cancer grading or cancer type classification, Figure 3.3, a). *Regression* is another very popular ML task. In regression the outcome is not a qualitative value (i.e., a class), but a continuous value. The learning process in regression is to define a fitting function that can explain the training data examples. An example of regression is house pricing, Figure 3.3, b). *Anomaly Detection* is the task where the ML algorithm receives an amount of data and scans it to identify anomalies, unusual or atypical samples. It means that the algorithm will identify the probability distribution that the regular data belongs to. In this way, the ML algorithm can point a strange behavior when it differs from the original data distribution. An example of anomaly detection is credit card fraud detection, Figure 3.3, c). *Clustering* is the ML task that intends to identify underlying patterns, classifications, trends, over an amount of unlabeled data. It is mostly related to unsupervised learning. An example of clustering would be the clusterization of articles or papers in a dataset to identify major trending topics, Figure 3.3, d). *Transcription* is the task where the algorithm must learn to receive an unstructured type of data (e.g., an image), and transcribe it into a discrete and structured form of output (e.g., text). An example of transcription is optical character recognition, Figure 3.3, e). In *Machine Translation* task the input is already structured sequence of symbols in some language and the algorithm must learn how to correctly codify the input into another structured sequence of symbols. An example is the translation from English to Portuguese, Figure 3.3, f) [Goodfellow, Bengio e Courville 2016].

3.2.2 Deep learning

Deep learning (DL) is a term in machine learning that comprehends a group of ML algorithms that derive from *neural networks*. DL was initially presented in the form of graph transform networks using gradient-based learning. Its efficiency was verified in the handwritten digit recognition task. The strongest point in neural networks was and still is the ability to perform automatic learning

rather than relying on hand-designed heuristics which is the basis of many other traditional learning approaches [LeCun et al. 1998]. The combination of larger neural networks (multiple layers), the backpropagation algorithm, the abundance of many types of data, and the expressive growth of the computational power allowed deep learning to rise and become one of the most powerful set of tools when it comes learning algorithms to perform human tasks. Since then, DL became state-of-the-art in many tasks such as image classification, object recognition, object detection, and image, video, audio, and speech processing [Lecun, Bengio e Hinton 2015, Goodfellow, Bengio e Courville 2016].

3.2.2.1 Neural Networks

A neural network is a mathematical algorithmic formulation that tries to mimic the way brain learns according to cognitive sciences. Mathematically speaking, it is a function $f(x)$ composed of a set of neurons that correctly associate an input x to an specific output y. A neural network with a single neuron (Figure 3.4, a) is equivalent to a linear regression ($y = ax + b$). The neuron multiplies its input by a weight factor (w_i), adds a bias factor (b_i), and passes this result or not through an activation function (g_i). This is how an individual neuron works.

A collection of neurons defines a layer. A neural network composed of a layer of neurons is equivalent to a logistic regression (Figure 3.4, b). This network can receive not just a single feature number (x), but a feature vector ($x = \{x_0, x_1, x_2, x_3, x_4\}$). Each feature vector component (x_i) is processed by an individual input layer neuron and its output is forwarded to the output neuron (y) that performs a weighted summation, adds a bias factor, and applies an activation function over the result.

When a series of layers are stacked as in a fully-connected approach it creates a *Multi-Layer Perceptron* (MLP). MLP can be understood as the first generation of neural networks and composes a group of networks called *feedforward neural network*, Figure 3.4, c. This nomination has to do with the idea that the input information is propagated through the network only in a single direction. Those networks can certainly learn much more complex input-output relationships than the linear or logistic regression neural nets (Figure 3.4, c). It happens given the presence of

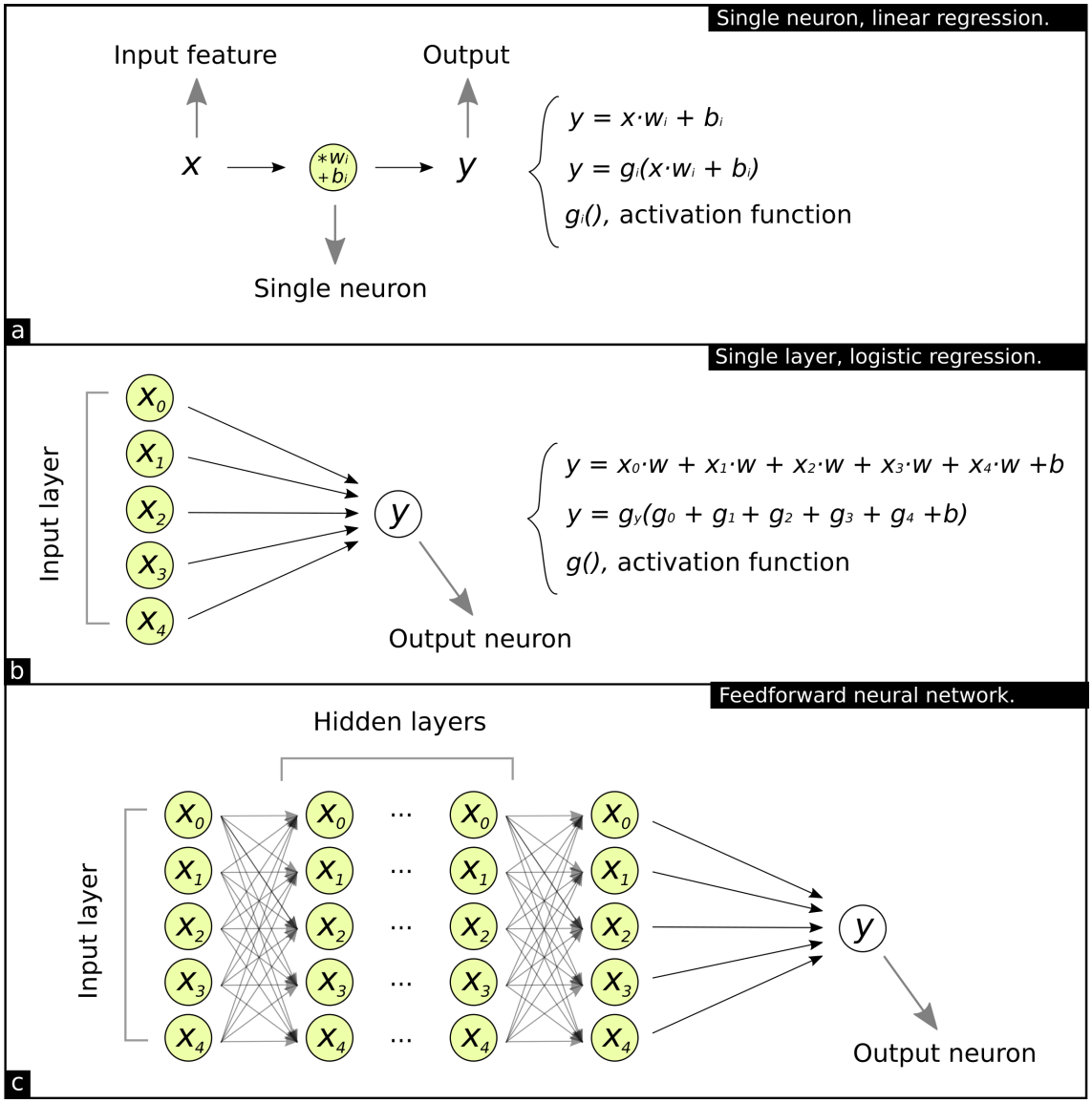

Figure 3.4: *Neural networks basic definitions. a) Single neuron functioning like linear regression. b) One-layer neural network. Works similar to logistic regression. c) Multi-layer perceptron. The first popularized neural network architecture.*

hidden layers that allows higher levels of abstraction for data representation. The MLP and Feedforward nets correspond to the basis of all neural networks concepts latter developed [Goodfellow, Bengio e Courville 2016].

3.2.2.2 The learning mechanism: the backpropagation algorithm

The learning process of a neural network happens by interactively updating the weights and bias associated with each neuron in the network, Figure 3.4, a. But how to update those weights and bias? Well, they must be updated into the direction that minimizes the *loss function*, $L(\mathbf{w},\mathbf{b})$. Loss function is a mathematical formula to measure the error of input-output ($x \rightarrow y$) association for a given set of weights ($\mathbf{w}$) and bias ($\mathbf{b}$). In this way, it is possible to calculate new weight and bias values by deriving $L(\mathbf{w},\mathbf{b})$ according to each w_i and b_i and calculating new values for w_i and b_i that minimizes $L(\mathbf{w},\mathbf{b})$ using a predefined learning rate (α).

$$w_i = w_i - \alpha \frac{\partial L}{\partial w_i} \tag{3.1}$$

$$b_i = b_i - \alpha \frac{\partial L}{\partial b_i} \tag{3.2}$$

The learning rate (α) defines how fast w and b change toward the minimization of $L(\mathbf{w},\mathbf{b})$, at each iteration. Those equations compose the *gradient descent method*, also known as the *backpropagation algorithm* in the context of neural networks. This amazingly simple code constitutes the basis for all the powerful learning algorithms used in deep learning sciences.

3.2.2.3 The introduction of the convolution: The Convolutional Neural Network

Feedforward networks achieved impressive results at its time, but still had some limitations related to the amount or dimensions of the input data. Feedforward could only take one-dimensional vector features, meaning that it could not process daily life images, since they originate from very large input vectors. If we convert an image of 256 x 512 pixels, for example, into a flattened one-dimension feature vector it would yield a 131072-unit vector. A network receiving such an input vector would demand an overwhelming number of neurons (weights and bias) and it means a substantial number of parameters to be optimized in the learning process, which demands a lot of computational power and training time. This limit held neural networks to be effectively used over images for a time until a turning-point concept came into scene: the convolution.

Convolution is a mathematical operation where a filter of dimensions f_x x f_y walks over an image convolving the image, region by region as in Figure 3.5. The filter is a mask, which is a set multiplying factor for every pixel in the filter, numbers in red in the yellow pixel selection, Figure 3.5. At each filter position, it performs the multiplication of the filter weighting factors by the image pixel values, then they are summed up into a single number. This single number is the new pixel of the resulting convolved image. Every filter generates a *feature channel*, and for a set of n filters $(f_0, f_1, f_2, ..., f_n)$ a *feature volume* with depth n is generated. A similar convolutional operation can be performed over this output feature volume. The difference is that the filter would have f_x x f_y x n, with n being the number of feature channels in the feature volume.

The usage of image convolution gave birth to the class of networks known as *Convolutional Neural Networks* (CNN). Those nets can receive as an input two and three-dimensional input images. Convolutional layers allowed CNN to operate over larger input data (e.g., images) with much less parameters to be optimized. CNNs, beside convolutional layers, contain other types of layers such as *max-pooling* layers, which perform dimension reduction, and *fully connected* layers, which operate in a feedforward manner similarly to MLP, Figure 3.6.

3.2.2.4 Neural network architectures

Since the introduction of image convolution in a neural network structure an huge number of increments, layer types, activation functions, ways of information propagation inside the net, were developed to enhance training time, performance, and task adaption. This process gave birth to a wide variety of neural network architectures that keep growing as new tasks are being tackled with neural networks.

Some of the classical neural network's architectures and the most important ones are shortly presented:

- **LeNet5**, seven layers, 60K parameters to optimize, used sigmoid function as the activation function, and was used for hand-written digit recognition;
- **AlexNet**, eleven layers, 60M parameters to optimize, used for image classification on ImageNet dataset;

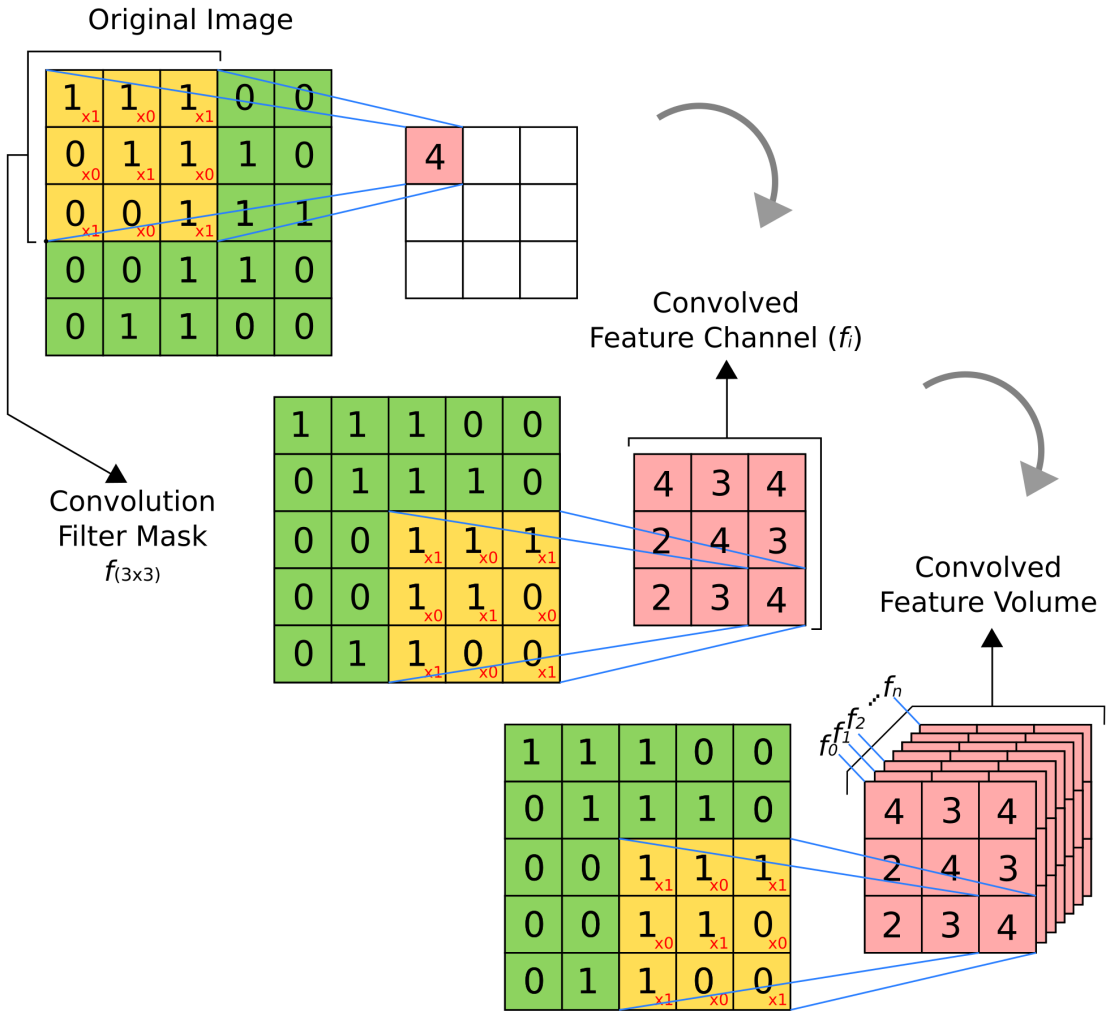

Figure 3.5: *Image convolution on deep learning applications. It illustrates the convolution process over a regular image.*

- **VGG-16**, sixteen convolutional layers, 138M parameters to optimize, used in the image recognition task on ImageNet dataset;

- **ResNets**, implemented in thirty and fifty-layer sizes, introduced the usage of *skip connections* (feed a layer with outputs from whatever previous layer in the architecture, different from feedforward concept). This allowed really *deep* nets to be trained without gradient vanishing or explosion. From now and on we have true deep learning architectures. It used ReLU activation function, and the 1 x 1 convolution;

- **Inception or GoogLeNet**, twenty-seven layers, combined convolutional and

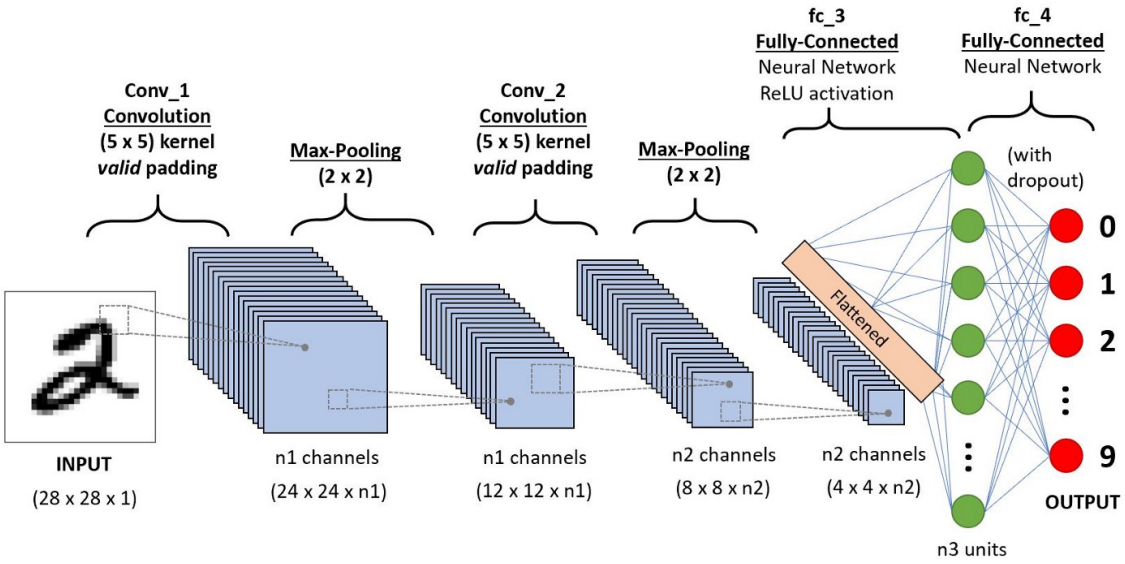

Figure 3.6: *Convolutional neural network example. It is illustrated the architecture of a CNN for handwritten digit recognition task. The input is an image of 28 x 28 x 1 and the output is a 10-unit vector containing the probabilities of the input figure being one of those index numbers. Font: Sumit Saha. A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way. www.towarddatascience.com*

pooling layers of different sizes simultaneously in a single block, the learning algorithm would learn which type of layer would better fit the model, while the other types would be reduced to identity;

- **MobileNet**, designed for environments with short computing resources, introduced the concept of *depth-wise separable convolution* which is much faster than the regular convolution;

- **U-Net**, the first architecture proposed for medical image segmentation, made use of the skip connection and a u-shaped architecture for initially encoding the image information and secondly decoding such information [Lecun, Bengio e Hinton 2015, Goodfellow, Bengio e Courville 2016].

Those architectures can be graphically represented as the one in Figure 3.6. Many other architectures are being developed by the deep learning community every day to suit its necessities as deep learning is being considered to solve new tasks.

Additionally, the development of new architectures is also triggered by the necessity to achieve higher performances in tasks already known.

3.2.3 Setting and fine-tuning a machine learning/deep learning experiment

A machine learning experiment requires identifying and translating the task into a machine learning frame. Next, data separation is mandatory. The data must be separated in three parts: *Training*, *validation*, and *testing* dataset. Then, choosing a network architecture (when using deep learning) which is associated with the task itself. Lastly, *fine-tuning* the model.

Before define fine-tuning, the concept of *bias* and *variance* error must be introduced. Bias error is the difference between the accuracy expected by a human performing the task in question and the accuracy of the learning algorithm in the task. For example, in the task of image recognition, human performance is generally 100%. If the algorithm is scoring 75% in this task, the bias error is 25%. Bias error is a measure of a systematic error. Variance error is the difference between the learning algorithm performance on *training* set and its performance on the *validation* set. Variance error is a measure of how much the performance of the algorithm cannot be generalized to different data. In other words, the variance error tells if the algorithm is overfitting on the training data, if the algorithm learned too many training data traits to a point that it cannot be accurate over new data.

Now, the definition: fine-tuning is the process of optimizing *bias-variance trade-off*. It is a general principle in machine learning. In other words, it is desirable to improve the algorithms performance over the training data to a human-comparable level, but it cannot be so that the algorithm will not be able to succeed over new data. If a neural network trains for a sufficiently large number of steps it will overfit. It will start to perform very well over the training data, but not so well over the validation data. Validation set is used to guide the improvements on the model so we do not keep a model that is overfitting.

In practice, fine-tuning a machine learning application implies in adjusting all the aspects of the experiment: the data used for training, validation, and testing; the number of layers and the architecture itself; and the usage of regularization

techniques;

When dealing with deep learning problems some settings occur:

1. **High Bias Error**: It means a training data problem. A solution might be bigger networks (more layers and or units), training for longer periods, trying different optimization algorithms, and different architectures. That will lead to good performances in the training set, meaning a low bias.

2. **High Variance Error**: validation set problem. More data, regularization methods (L1, L2, and Dropout), as well as different architectures might solve the issue.

3. **Bias-Variance Trade-off**: We may want to reduce bias error, keeping the variance low. Or vice versa. In the deep learning era, large networks and big data tend to be enough solutions for both bias and variation errors.

After fine-tuning is done, testing dataset (unseen data) is used to measure an overall final performance of the network. The testing set is a data not used in the deep learning model development but to evaluate its final performance.

With all those concepts on artificial intelligence (IA), machine learning (ML), and deep learning (DL) now introduced, we can better understand the solutions so far proposed for osteoporosis management on the literature regarding artificial intelligence, and more specifically deep learning. In the next session we will point out some approaches present on the literature and their specificity.

3.3 AI in the osteoporosis scenario

With such versatility and attractive results, AI has been extensively used in medicine to tackle a variety of clinical tasks, as we mentioned in the beginning of this chapter. In osteoporosis management scenario, the focus of the present work, we can find many studies covering it [Wani e Arora 2020, Cruz et al. 2018]. CNNs were used for screening osteoporosis and osteopenia through lumbar and hip radiography [Zhang et al. 2020, Yamamoto et al. 2020]. Hip fracture prediction was performed for post-menopausal women using age, bone mineral density, clinical and

lifestyle factors [Ho-Le et al. 2017]. Deep learning models were used to detect and classify bone fractures in X-ray images [Lindsey et al. 2018, Pranata et al. 2019]. Machine learning algorithms were trained with bone vibroacoustic response signals to assess patient bone quality [Scanlan et al. 2018]. MRI images were used for training Machine Learning algorithms to predict frailty fractures [Ferizi et al. 2019].

In face of these findings that correlate dental panoramic X-rays (PAN) with osteoporosis, some studies in the literature performed analyzes using PAN images to improve osteoporosis detection [Lee et al., LING e YANG 2020, Lee et al. 2020, Sukegawa et al. 123]. In [Lee et al.], the authors trained simple fast-forward convolutional neural networks and used rectangular ROIs to extract the region below teeth on PAN images (Figure 2.5 a)) to separate normal from osteoporotic patients according to Klemetti's criteria. Their algorithm trained very well without any further procedure (e.g., transfer learning or data augmentation), and they achieved a 98.5% accuracy. In [LING e YANG 2020], the authors claim to have used the entire PAN images (Figure 2.5 b)) of 108 patients to train a deep learning architecture using ImageNet pre-training to separate osteoporotic from normal subjects. This study reportedly achieved 92.0% accuracy. However, the authors provided no information regarding the patient demographics, data separation, deep learning architecture, or the criteria used to define osteoporosis. That makes it hard to situate and compare this study's contributions for this task. In [Lee et al. 2020], the authors tested the improvements of using pre-trained weights and partial fine tuning with a VGG-16 in the task of separating normal (including osteopenia) from osteoporosis patients. In this study, they also used a rectangular ROI containing the mandible inferior border (Figure 2.5 c)) to crop the original PAN images before feeding the deep learning algorithms. They achieved 84.0% accuracy with the best model (VGG-16 + transfer learning + partial fine-tuning). Lastly, in [Sukegawa et al. 123], the innovation brought by the study relied in the usage of a single-side rectangular ROI 2.5 c)), the usage of deep learning ensembles (using Efficienet and ResNet variations) and in the combination of deep learning models with clinical parameters (age, height, and mass) in a mixed model. Despite the architectural improvements in the machine learning strategies used there was no improvements in the accuracy (84.5%) for the task if compared to [Lee et al. 2020].

The present study assessed the osteoporosis risk using the EfficientNetV2 architecture pre-trained with ImageNet weights and trained with the mandible segmentation extracted from PAN images. We investigated the osteoporosis risk assessment looking at the PAN images in two different approaches: considering the entire PAN image and the mandible-segmentation ROI. In the next chapter, we will describe the entire workflow used to develop the automatic mandible segmentation model used to extract mandible regions from PAN images. And in the Chapter 5 we will present our final thesis experiments, the AI models trained for assessing osteoporosis risk using the mandible-segmentation ROI on PAN images. These next two chapters contain methodology, results, discussion, and conclusion sessions to fully cover our thesis goals.

Chapter 4

Deep-learning-based automatic mandible segmentation algorithm

In this chapter, we will discuss in detail the workflow and experiments performed for developing a deep-learning-based segmentation algorithm for contouring mandible on dental panoramic X-ray images. This study was already published. For more details check this reference [Machado et al. 2023].

4.1 The patient group and image dataset

Two image datasets were used: an in-house prepared dataset (IHD) and a third-party publicly available dataset (TPD) [Abdi e Kasaei 2017]. The IHD was prepared using 362 patients treated in the Dentistry Department of Hospital das Clinicas da Faculdade de Medicina de Ribeirao Preto (HCFMRP), who at some point needed PAN images for monitoring oral health. All those patients were imaged using the Sirona scan ORTOPHOS XG-3D/Ceph 60-90 kVp. PAN images had 2432x1272 pixel resolution with 32-bit gray-level-intensity depth. To be included in the study the subject should: be an oral patient from the referred healthcare institution, have image(s) available and with diagnosis-quality, have clinical data available, and have mandible segmentation performed by the specialist collaborator in the present study (Figure 4.1, a). Forty-one patients were excluded given some file corruption that caused some mismatch between PAN image and manual segmentation. A total of 321 patients were included. This sample contained a wide range of age [10.68 y.o., 97 y.o.], representative gender distribution (190 men and 131 women) (Table 4.1), and patients with varied clinical conditions (partial to complete endentulous mouth

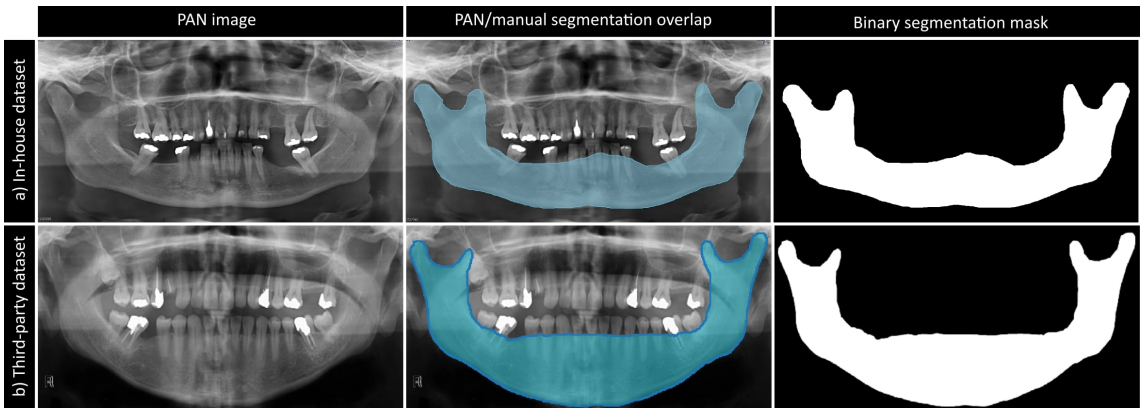

Figure 4.1: *Dental panoramic X-ray, manual segmentation, and binary segmentation mask for in-house dataset (IHD, a) and third- party dataset (TPD), b). The binary mask sets 1 for mandible and 0 elsewhere.*

and patients with implant). The segmentation was done by a dental radiologist with more than 30 years of experience using the Segment Editor module of the image processing software 3DSlicer (4.11 version) [Fedorov et al. 2012]. Some of the patients included had two or more different PAN images that were also manually segmented. We end up with IHD composed of 393 image/manual segmentation pairs. All this data was used under the approval of the HCFMRP's ethics committee.

A third-party dataset (TDP) originally used and published by Abdi *et al* (2015, 2017) was considered. This dataset is composed of 116 PAN images (1250x2900 pixels), each one containing two manual segmentations drawn by different dental radiologists (Figure 4.1, b). Those images were acquired using the Soredex CranexD digital panoramic X-ray scanner at Noor Medical Imaging Center, Qom, Iran. The subjects belonging to this dataset cover a wide variety

Table 4.1: *Age and gender description of in-house dataset (IHD) patients included in the study.*

Gender	***N***	***Image Pairs***	***Age at image (y.o.)***	***Min Age (y.o)***	***Max Age (y.o)***	***(%)***
Male	190	253	69.06 ± 13.48	10.68	97.00	59.19
Female	131	140	57.66 ± 12.49	20.00	91.17	40.81
Overall	321	393	64.89 ± 14.22	10.68	97.00	100

of dental conditions from healthy, to partial and complete edentulous cases [Abdi, Kasaei e Mehdizadeh 2015, Abdi e Kasaei 2017]. Those manual specialist segmentations were fused into a single segmentation (as displayed in Figure 4.1, b) through absolute voting criterion (i.e., the final segmentation image contains only pixels that were marked as mandible regions on both specialists' segmentations). Those unified segmentation masks were paired with each respective PAN image. TPD and IHD datasets were combined to train, validate and test our automatic segmentation model.

4.2 The deep learning models

We considered two deep learning architectures to tackle this segmentation problem: U-Net and HRNet architectures. U-Net is defined as an encoder-decoder architecture that progressively reduces image resolution through convolutional sequences and, later, up-samples it back to the original resolution. As it is up-sampled backward, previous resolution/layer information (skip connection) is added to recover image high-resolution details [Ronneberger, Fischer e Brox 2015]. HRNet (High-Resolution Net) has recently shown impressive results on semantic segmentation, human pose estimation, and visual recognition in public benchmarks, e.g., COCO dataset[1]. HRNet performs segmentation by passing the image through a series of convolutional streams at the same time it keeps the high-resolution stream, three lower-resolution streams in parallel, and exchanges information repeatedly among all the resolution levels in a fully connected approach [Wang et al. 2020].

4.2.1 UNet

Figure 4.2, a) presents U-Net elements. U-Net was designed to receive a 256x512x1 input image and output a same-resolution image segmentation mask. This architecture is composed of two main parts: The encoder and the decoder side. The encoder is composed of 4 encoding convolutional blocks ($c_i, i = 1, 2, 3, 4$) that contain two same-padded, 3x3, convolution that doubles the number of channels C (which is 32, at c_1) and a 2x2 max-pooling that halves the input. The fifth

[1] https://paperswithcode.com/task/semantic-segmentation

stage (c_5) is a transition layer, it does not down-sample its input, but convolutes it and doubles its number of channels. Every convolution block in the first half of the model outputs two image blocks: the max-pooling-halved block and the image block before max-pooling. This last block is the information saved from each level that will be used in the second half of the model to recover high-resolution information through skip connections. The second half of the model is the decoder side. It is composed of four decoding convolutional blocks ($c_i, i = 6, 7, 8, 9$). Each of those blocks is composed of an up-sampling (transposed convolution) that doubles the input resolution and halves the input number of channels; a concatenating layer that adds the up-sampled block and the same-depth encoder-side image block (skip connection); and two same-padded, 3x3, convolutions. The last image block is set to have the same input resolution, but two feature channels. Each feature channel brings the probability of each pixel belonging to class 0 (background) or 1 (mandible segmentation mask).

4.2.2 HRNet

HRNet architecture is presented in Figure 4.2, b). It was implemented according to the definitions in [Wang et al. 2020] with very few modifications. Similarly to U-Net, the input image resolution is 256x512x1 pixels and the same resolution for the output segmentation mask. The input image passes by two successive down-sampling convolutions (followed by batch-normalization and ReLU activation) that halves the initial resolution to 128x256, and then to 64x128 (1/4 of the initial resolution). At the same time, it increases the number of feature channels to C/4, and then to C/2 (with $C = 64$). This 64x128x16 image block is streamed through the main component of the HRNet architecture. The main component is composed of four forward convolutional stages (s_i , $i = 1, 2, 3, 4.$) and four parallel different resolution streams ($r_j, j = 1, 2, 3, 4.$). Each stage comprises four bottleneck-convolution blocks and an additional convolution to a transition layer/block. The bottleneck comprises of a 1x1, 64-channel convolution, a 3x3, 16-channel convolution, and another 1x1, 64-channel convolution. The transition layers are responsible for concatenating all the resolution streams' output by adding the channels but unifying the final resolution to the respective stream resolution

level by up-sampling or down-sampling the stream's previous output blocks. In the first stage, there is only one stream, with only the 1/4 resolution (r_1) stream and C feature channels (Figure 4.2, b).

At the end of the first stage, a down-sampling convolution halves the resolution doubles the number of feature channels (2C), and creates another stream with the halved resolution. The same happens at the end of the second stage and third stage. At the end of the fourth stage, we have four parallel streams with four different resolutions. The outputs of those resolution streams are concatenated into a final output head at the 1/4 resolution with 15C channels. This final head is up-sampled twice to 1/2 resolution and C/2 channels, and then to the original 256 x 512 pixels resolution with two feature channels, similar to the U-Net model.

4.2.3 Models' additional setting

Both U-Net and HRNet were trained using Adam optimizer and sparse categorical cross-entropy loss function. The original implementation of HRNet architecture uses C = 64, but considering the available computational resources, it was used C=32. Next, in the originally proposed architecture, the up-sampling operation is carried out as bilinear interpolation, using the nearest neighbor interpolator. However, we decided to use a more sophisticated up-sampling operation, the transposed convolution, as implemented in the TensorFlow software package, for it offers a smoother up-sampled output.

4.2.4 Models' Implementation

All the deep learning algorithms described here were implemented using the TensorFlow library (2.8.0 version), an open-source library for deep neural networks implementation (Abadi et al., 2016). Additionally, the CUDA library (11.6 version) was used to accelerate the model training step through GPU parallel processing (NVIDIA et al., 2020).

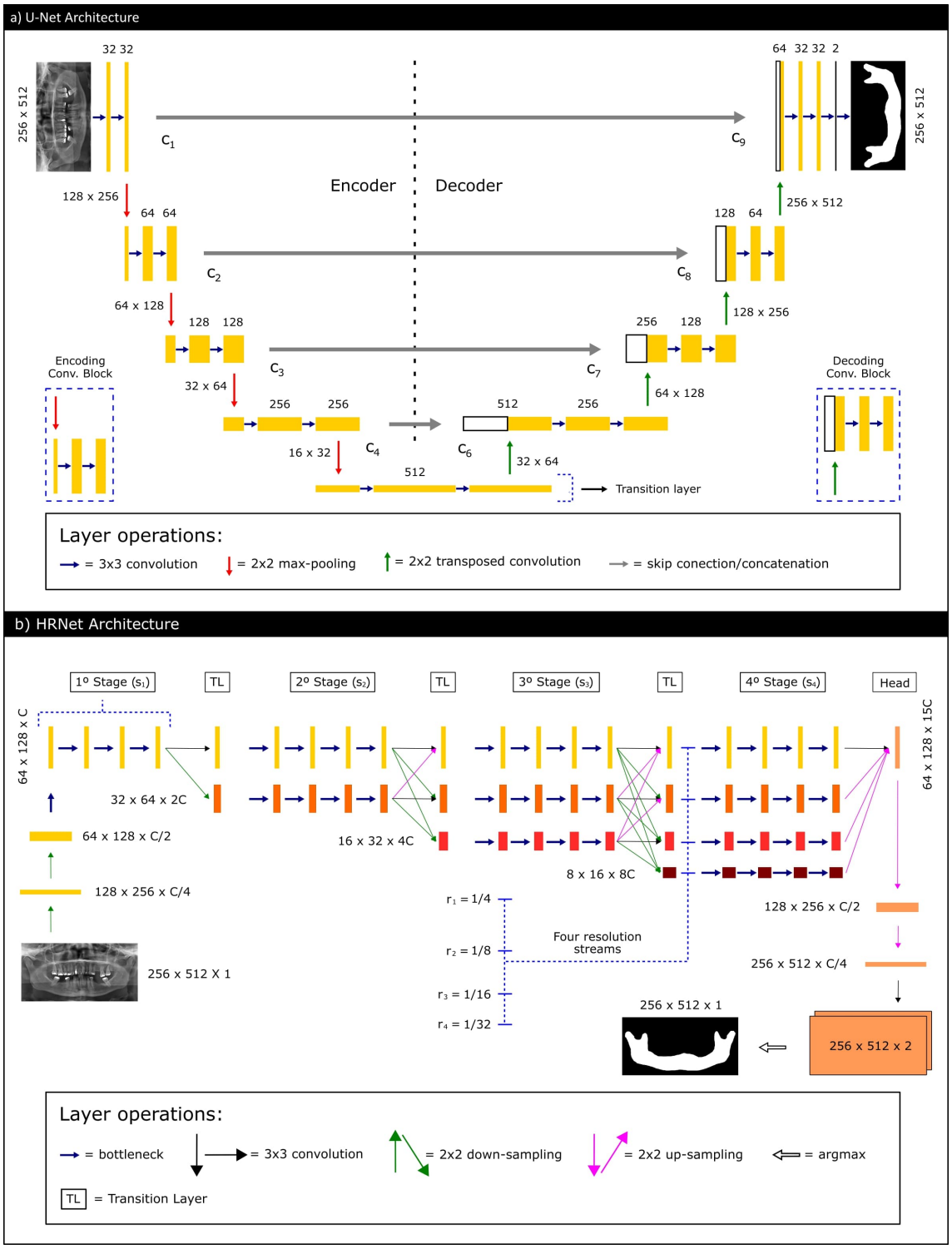

Figure 4.2: *U-Net (a) and HRNet (b) Architectures. Both U-Net and HRNet 3x3 convolutions, and HRNet bottleneck are same-padded layer operations. HRNet bottleneck is a series os 1x1xC, 3x3xC/4, and 1x1xC convolutions, with* $C = 32$*. Argmax is a function that reduces the 256x512x2 input to a 256x512x1 output by keeping the layer index (0 or 1) holding the highest value for each pixel.*

4.3 Metrics

Three metrics that are frequently used to evaluate performances of segmentation algorithms were considered: Accuracy (ACC), Dice Similarity Index (DICE), and the Intersection over Union (IoU) measure. Those three metrics are the most used ones for evaluating segmentation algorithm performances and will allow us to compare our results with previous studies. Here is the mathematical definition of those three metrics:

$$ACC(Predicted, True) = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

$$DICE(Predicted, True) = \frac{2TP}{2TP + FP + FN} \tag{4.2}$$

$$IoU(Predicted, True) = \frac{TP}{TP + FP + FN}. \tag{4.3}$$

Predicted refers to the predicted segmentation mask and *True* to the true specialist's segmentation mask. It is important to recall that using deep learning algorithms for image segmentation means predicting each pixel individually as foreground or background. This idea is present in the rates used to calculate the similarity indexes above: *TP*, true positive rate, the number of pixels correctly classified as foreground; *TN*, true negative rate, the number of pixels correctly classified as background; *FP*, false-positive rate, means the number of pixels misclassified as foreground; and *FN*, false-negative rate, the number of pixels misclassified as background. For ACC, DICE, and IoU metrics, 1 indicates the perfect performance and 0 the worst one.

4.4 Dataset separation and data augmentation (DA)

The 393 image/segmentation mask pairs were stratified for training, validation, and testing models according to the following separation: 253/70/70 (64%/18%/18%, train/validation/test) sets. Following the same proportion, the TPD dataset was separated in 76/20/20 (train/validation/test). The respective

sets from both datasets were concatenated into a single train, validation, and test datasets (329/90/90), as illustrated in Figure 4.3. Data augmentation (DA) is the generation of extra and diverse data from a particular dataset to fit more complex, generalized, and accurate models. We used three image augmentation operations to generate extra data: horizontal and vertical random flip, random rotation, and random contrast (this last one only over gray-scale images) (Figure 4.4). Every image/segmentation-mask pair passed through the augmentation routine in such a way that the mask experienced the same transformations of the image, except for the random contrast. We applied data augmentation only over the training dataset and generated three different transformed datasets (329 pairs each) that were combined with the original image training dataset. Hence, the final augmented training dataset contained 1316 image/segmentation-mask pairs. Validation (90) and test (90) sets remained the very same original images for all the models tested here since they represent the exact real-world images that deep learning algorithms should learn to segment.

4.5 Model Improvements

Two strategies for improving the deep learning segmentation model outputs were considered: The morphological refinement and the ensemble learning approach. Both strategies were performed and evaluated over validation and test set.

4.5.1 Morphological refinement (MR)

Segmentation masks are binary image maps (0: background, 1: foreground), as in Figure 1. Morphological operations are shape-based image processing steps performed over binary label maps to change the label's form, structure, borders, etc. The most common ones are the erosion (reduces label map by its borders) and dilation (increases label maps by its borders) operations. As an attempt to correct minor imperfections, such as noisy borders, small-disconnected label objects, and holes on the specialists' segmentation and on the segmentations predicted by the deep learning models, we proposed a morphological refinement (MR) routine composed of two stages: I) island removal, i.e., detection and removal

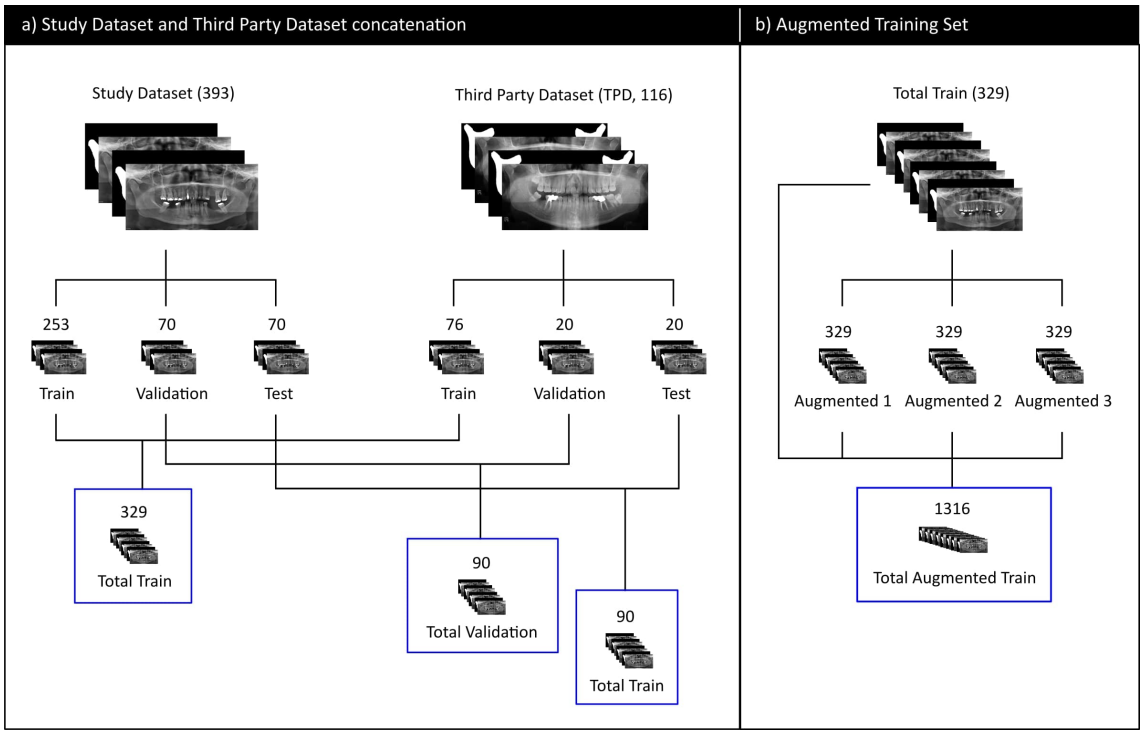

Figure 4.3: *Datasets separation and combination. Total train (329), total validation (90), total test set (90), a), and total augmented train (1316), b), are the datasets used in the experiments with and without augmentation. Total Validation and Total Test set (90 each), a), are composed of the very same images used to validate and test all the models here developed.*

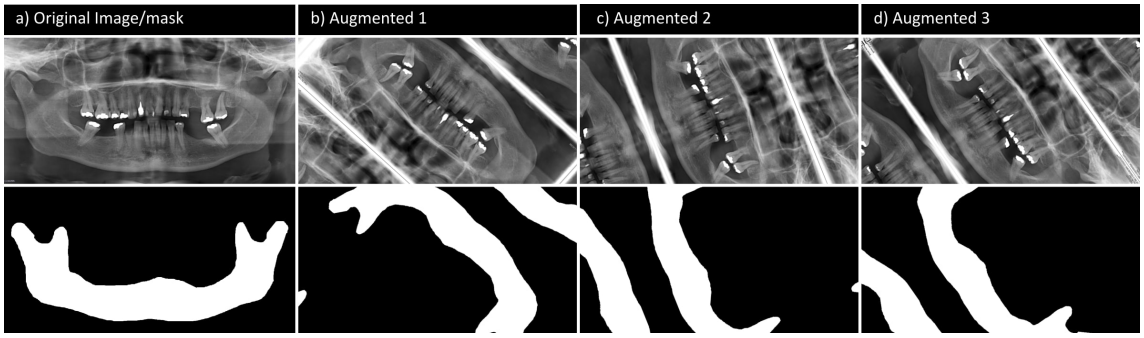

Figure 4.4: *Data augmentation operations applied over the original images. The augmentation operations were applied simultaneously over each PAN image and segmentation mask image. We augmented the original training dataset (a) three times (b, c, and d) using random seeds which yielded three different augmented datasets with 329 pairs each..*

of small-disconnected label objects; and II) border smoothing and hole filling, i.e., a series of dilation and erosion operations to smooth out the predicted segmentation masks' borders and close label object holes. The morphological refinement was implemented using the open-source SimpleITK package (2.1.1 version), a library for digital image processing and analysis [Lowekamp et al. 2013]. MR was also used on specialists' segmentation maps, on both IHD and TPD, before training the deep learning models.

4.5.2 Ensemble Learning

The ensemble is a technique used in machine learning when there are two or more models available to solve a specific task. In semantic segmentation, the segmentation models perform an individual prediction for every single pixel (foreground or background). A set of n semantic segmentation models can be combined into a single ensemble model. The final prediction for each pixel is a decision made over the n predictions for each pixel. The rule can be an absolute voting rule (a pixel is assigned as foreground if it has n foreground predictions), a most voting rule (a pixel is assigned as its value according to the most frequent prediction), or a sufficient voting rule (a pixel is assigned as foreground if it was predicted as foreground a given number of times and above). After training the individual segmentation models, we checked if we could benefit from an ensemble model.

4.6 The experiments performed

A total of four segmentation experiments were performed combining architectures and datasets: I) U-Net architecture using only the total dataset, 329/90/90 (train/validation/test) (Figure 4); II) U-Net architecture with the augmented training set and total validation and test set, 1268/97/97. III) HRNet architecture using only the original dataset, 329/90/90; IV) HRNet architecture with the augmented training set, original validation, and test set, 1316/90/90. The performance of the four trained models was evaluated with and without MR to verify which model benefit from this additional image processing step. And V) an ensemble

model composed of the best-trained models (I to IV) was built and evaluated on total validation and test set. It was also checked if MR could improve the ensemble's final output segmentation.

4.7 Results

Numerical and visual results are presented for each of the four individual models trained and for the composed ensemble. We developed four segmentation models UNet, UNet + DA, HRNet, HRNet + DA. Those models were experimented alone and with morphological refinement (MR). Latter, an ensemble model was proposed with and without morphological refinement.

4.7.1 Deep learning segmentation only

Tables 4.2 presents the performances of the four models alone, without further refinement. U-net and U-Net + DA models achieved the highest performances on both validation and training sets, with U-Net achieving the highest ACC (98.19%), DICE (97.27%), and IoU (97.21%) results on the test set. Table 4.2 also reviews that all the models had excellent performances scoring above 95% in all metrics, for validation and test sets. Data augmented models, U-Net + DA and HRNet + DA, performed better on validation and test set than on training set. It happens because the DA process introduces more complexity to the training set (as in Figure 4.4), naturally making validation and test set (real-world images) look "simpler". It positively impacts the task goals since the validation and test set performances are the most important ones since they define the model's final quality. Further, for the DA models here trained, validation and test sets are the ones that bring real-world data type, distribution, and variability.

Figure 5 displays some segmentation results over two random PAN images chosen from the test set, one from the in-house dataset (IHD) and the other from the third-party dataset (TPD), to illustrate the quality of the results shown in Table 4.2, for each trained model. Every image/output segmentation pair displays its own metrics results. Despite having all performances superior to 95% on validation and test sets (for all three metrics), what should already be a relevant result, the other

three models' predictions still exhibit some imprecision.

Table 4.2: *Performances of the four trained models for each dataset (training, validation, and test set).*

	Performances (%)								
	Training Set			*Validation Set*			*Test Set*		
Models	*ACC*	*DICE*	*IoU*	*ACC*	*DICE*	*IoU*	*ACC*	*DICE*	*IoU*
U-Net	*98.64*	*98.00*	*98.32*	*98.43*	*97.60*	*97.73*	*98.19*	*97.27*	*97.21*
U-Net + DA	*96.75*	*94.92*	*95.46*	*98.17*	*97.04*	*97.23*	*97.96*	*96.81*	*96.78*
HRNet	*98.45*	*97.81*	*97.63*	*98.01*	*97.02*	*96.48*	*97.73*	*96.65*	*95.87*
HRNet + DA	*96.15*	*93.94*	*93.80*	*97.65*	*96.26*	*95.70*	*97.60*	*96.27*	*95.61*

ACC: Accuracy; DICE: Dice similarity index; IoU: Intersection Over Union; DA: Data Augmentation. The orange highlight indicates the best performances for each metric and dataset.

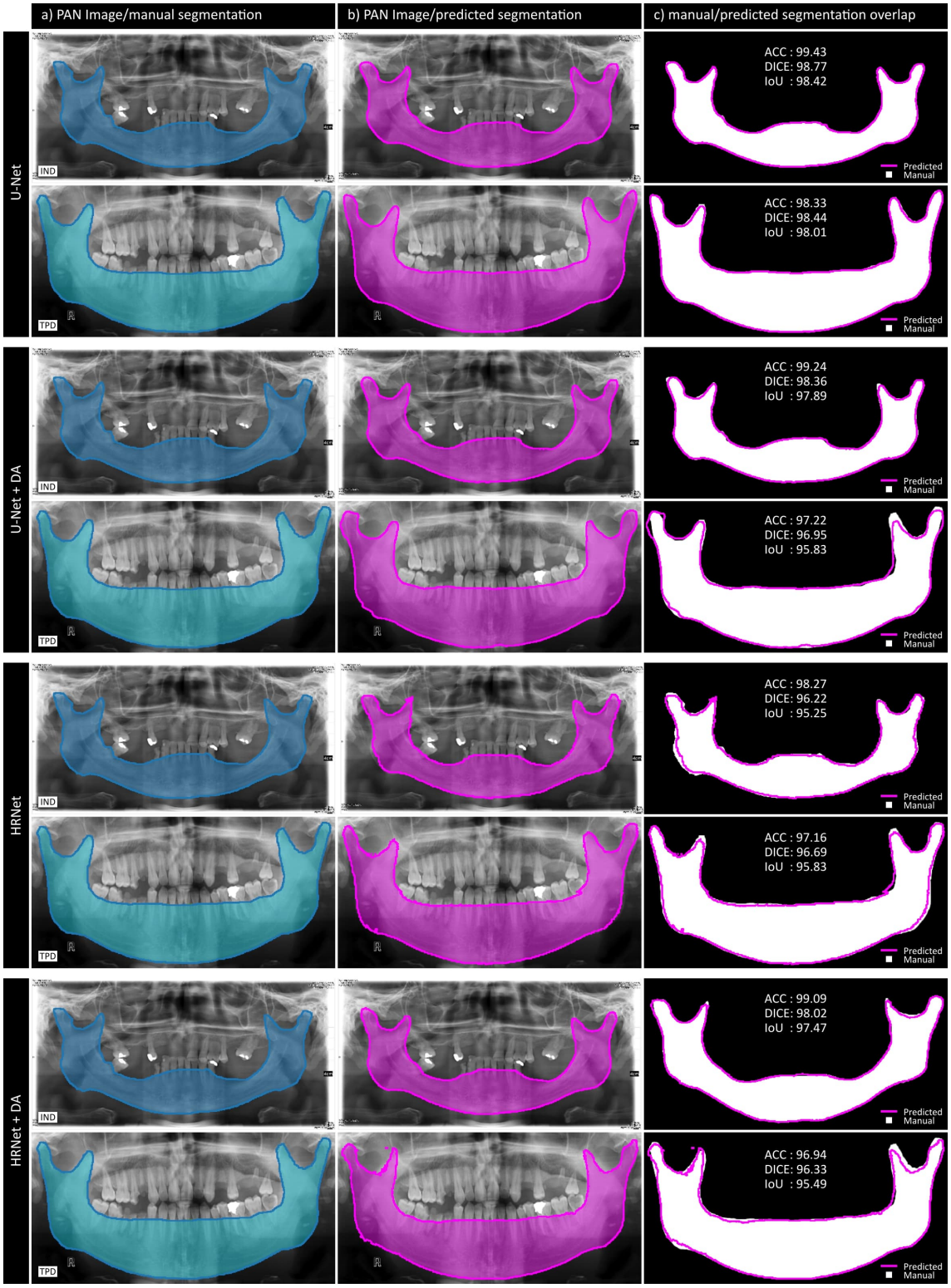

Figure 4.5: *Segmentation output of the four algorithms for two images from the test set, one from the in-house image dataset (IHD) and one from the third-party image dataset (TPD). In Blue (a), the manual segmentation. In purple (b), the predicted segmentation. In (c), the manual and predicted segmentation overlapped. ACC (accuracy), DICE (dice similarity), and IoU (intersection over union) metrics for each segmentation are displayed.*

Table 4.3: *Performances of the four trained models for validation and test sets after morphological refinement.*

	Performances (%)					
	Validation Set			*Test Set*		
Models	*ACC*	*DICE*	*IoU*	*ACC*	*DICE*	*IoU*
U-Net + MR	*98.43*	*97.64*	*97.72*	*98.21*	*97.32*	*97.23*
U-Net + DA + MR	*98.20*	*97.11*	*97.26*	*97.99*	*96.88*	*96.83*
HRNet + MR	*98.06*	*97.21*	*96.52*	*97.78*	*96.81*	*95.89*
HRNet + DA + MR	*97.72*	*96.42*	*95.82*	*97.63*	*96.36*	*95.63*

ACC: Accuracy; DICE: Dice similarity index; IoU: Intersection Over Union; DA: Data Augmentation. MR: Morphological Refinement; All the models improved with MR.

4.7.2 Segmentation results after morphological refinement

Some other segmentation outputs exhibited small-disconnected label pieces, small holes, and or noisy borders. Noisy borders were especially present on HRNet based models. A morphological refinement (MR) routine, a post-processing step, was tested to solve all those imprecisions on the output segmentations. MR was conceived in two stages: I) an island removal routine to detect small and separated amounts of pixels and to change them into background pixels (0). And II) a sequence of two dilate and two erode (using ball structuring element with radius = 5) operations to close holes and to smooth segmentation borders. Table 4.3 describes the numerical similarity metrics recalculated for validation and test sets after applying morphological refinement on models' segmentation outputs. Although small, all the models had an improvement with MR. U-Net models improved up to 0.07%, while HRNet models had improved up to 0.16% if compared with results before morphological refinement. HRNet + DA being the model that most benefited with MR.

When looking at Table 4.3, the improvements made by MR, averaged over the validation and training sets, look numerically small to justify MR. However, when

we inspect some individual image cases it is easier to notice the positive impact of such pre-processing step. Figure 4.6 illustrates four examples of PAN images, two from IHD and two from the TPD dataset, that benefited from MR. In a), it is easy to notice a mislabeled object region separated from the mandible. After MR, this region disappears, an island removal effect and the similarity metrics point to an improvement of 4.74 to 5.35% on DICE and IoU metrics, respectively. In b) and d), it is also possible to notice the island removal effect. In b), however, the similarity metrics point to a performance loss. This happens because the isolated small label object is situated over the actual mandible region. Thus, removing it implies reducing numerical similarity, while gaining, on the other hand, geometrical homogeneity. In c), it is possible to verify another MR effect, the hole filling. A small hole in the left superior side of the mandible region is filled with erosion operation, thus improving similarity scores. In d), it is also possible to check how MR improves noisy borders, which are mostly present on HRNet-based models.

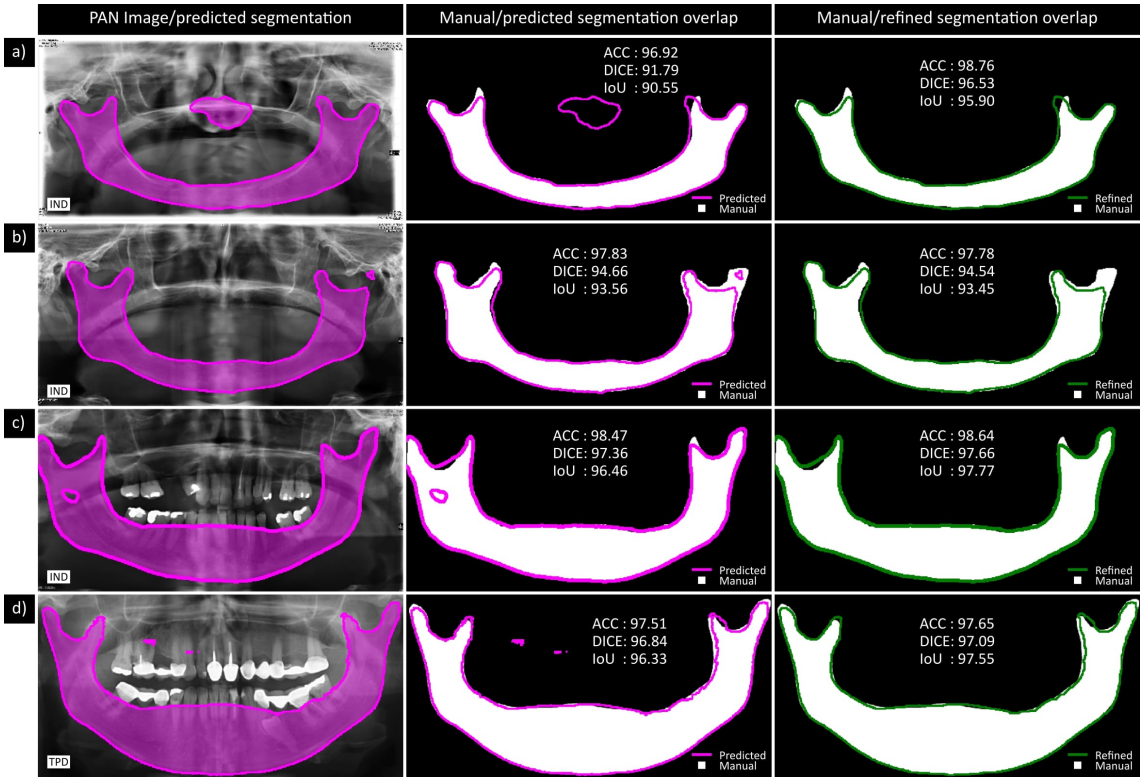

Figure 4.6: *Examples of how morphological refinement improves the predicted output of the developed algorithms. Island removal and smoothing border effects can be observed in a), b) and d). In b), island removal causes a slight loss in similarity metrics because the isolated piece is still placed over the actual mandible region. In c), a small hole located at the left superior side of the mandible is closed with MR. ACC (accuracy), DICE (dice similarity), and IoU (intersection over union) metrics for each segmentation are displayed. IND: In-house image dataset; TPD: Third-party image dataset.*

4.7.3 The limitation of the single model approach

The UNet + MR model alone, the best model individually developed in the present study, outperformed all the previously published studies' results on all similarity metrics available, it will be discussed in detail in the next session. However, despite this impressive result, every model fails in some cases. Figure 7 brings three examples where UNet + MR model fails. As we can see in Figure 4.7, when UNet + MR fails, the other models can offer better segmentation alternatives (marked with green tick sign). It means that, when the UNet + MR's output looks too imprecise, one could count on the other models' segmentation outputs for more accurate segmentation. This is one of the advantages of developing many different

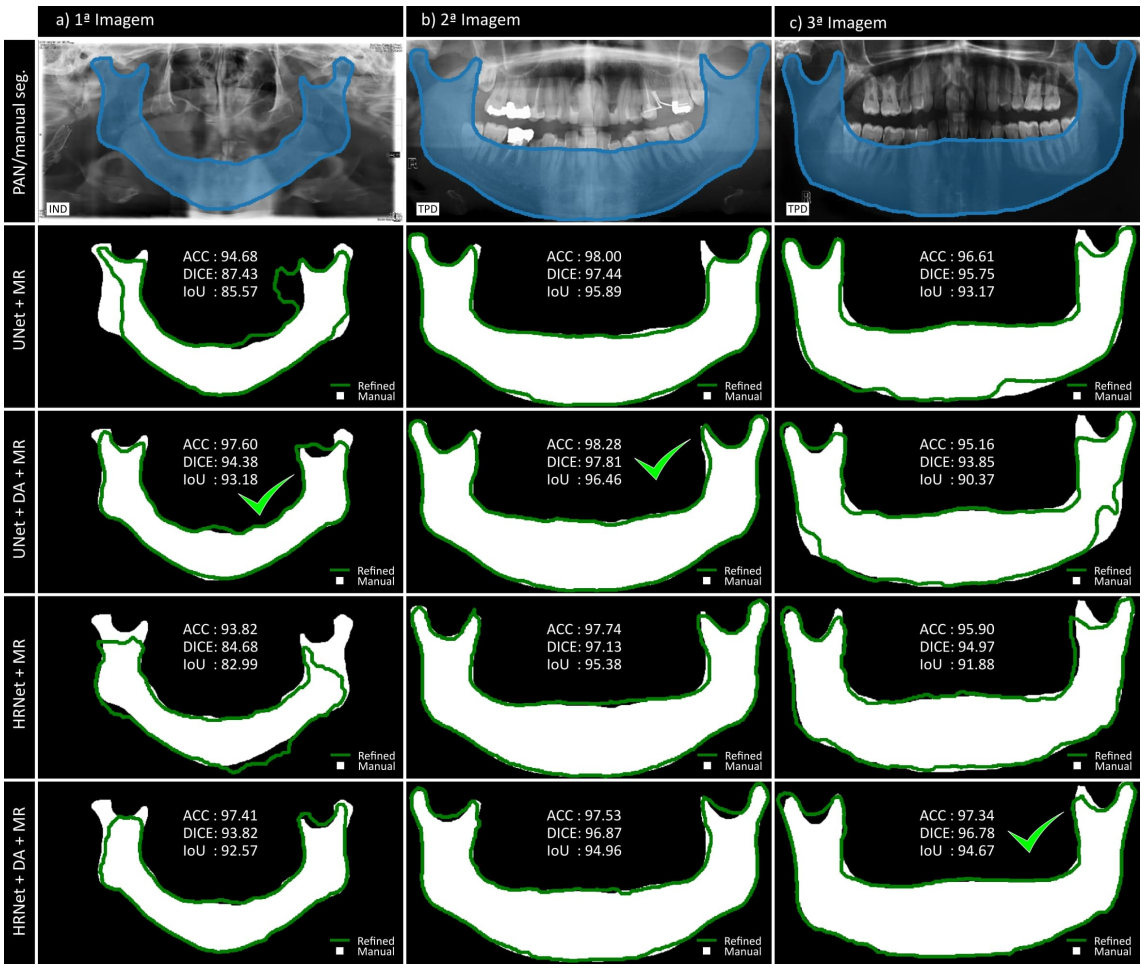

Figure 4.7: *Examples of alternative segmentations when the best-ranked segmentation algorithm fails. In blue is the manual specialist's segmentation. The green outline is the segmentation contour after morphological refinement. ACC (accuracy), DICE (dice similarity), and IoU (intersection over union) metrics for each segmentation are displayed. The green tick sign points to the best alternatives to the UNet + MR failed one.*

segmentation models at the same time.

4.7.4 The ensemble approach

As we can observe in Figure 4.7, developing more than one segmentation model offers the advantage of having additional segmentation guesses with precision possibly superior to the best-ranked model. This fact is also the motivation for the creation of ensemble models on machine learning. I.e., a collection of models combined into a single model may achieve better performance than the individual

Table 4.4: *Performances of the ensemble model on total validation and test sets with and without morphological refinement.*

	Performances (%)					
	Validation Set			*Test Set*		
Models	*ACC*	*DICE*	*IoU*	*ACC*	*DICE*	*IoU*
Ensemble	*98.45*	*97.81*	*97.59*	*98.27*	*97.59*	*97.19*
Ensemble + MR	*98.45*	*97.82*	*97.59*	*98.27*	*97.60*	*97.18*

ACC: Accuracy; DICE: Dice similarity index; IoU: Intersection Over Union; DA: Data Augmentation. MR: Morphological Refinement; All the models improved with MR.

models. Here, we composed an ensemble with the four best individually trained models: UNet + MR, UNet + DA + MR, HRNet + MR, HRNet + DA + MR (Tables 4.3). To compute the final ensemble segmentation prediction, firstly, the segmentation prediction for each one of those component models was acquired. Then, it was counted the frequency (0 to 4) of each pixel was classified as foreground across the four automatic segmentations. Lastly, the ensemble criterion was applied: if a given pixel was marked as foreground two or more times, then, it was assigned as foreground (1) in the final segmentation map, otherwise, it was assigned as background (0), i.e., sufficient rule. This ensemble prediction was performed for every image on the validation and test sets, and the similarity metrics were calculated with and without using morphological refinement on the ensemble segmentation output. The results are displayed in Table 4.4.

Both Ensemble and Ensemble + MR models outperformed the former best-ranked model (UNet + MR) in ACC and DICE metrics. No considerable difference was observed when using MR on the Ensemble model. Further, the ensemble improvement was observed numerical and visually. The UNet + MR failed segmentations in Figure 4.7 are considerably improved by Ensemble + MR (Figure 4.8). Ensemble + MR's segmentation are the most accurate alternatives considering the other four individual models (Table 4.3) and the Ensemble + MR (Table 4.4), for the three UNet + MR failed segmentations in Figure 4.7.

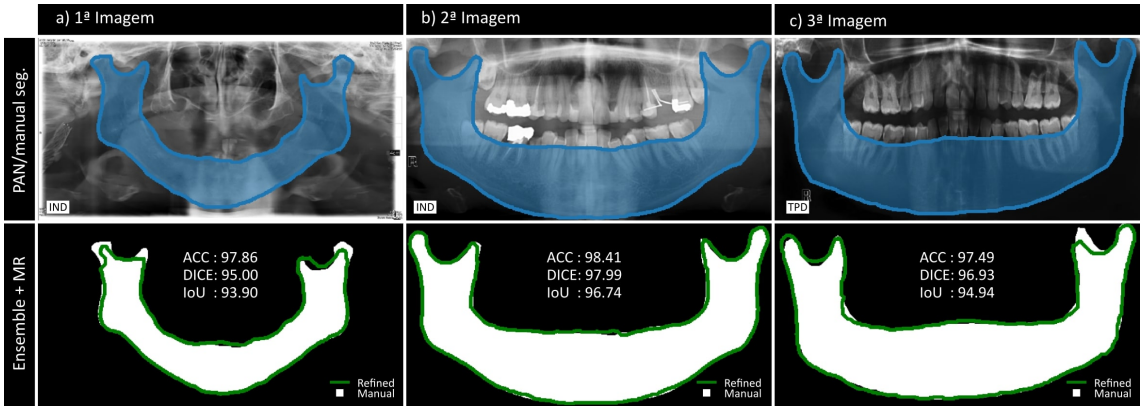

Figure 4.8: *Ensemble + MR segmentation output for the UNet + MR failed segmentations displayed in Figure 4.7. In blue is the manual specialist's segmentation. The green outline is the segmentation contour after morphological refinement. ACC (accuracy), DICE (dice similarity), and IoU (intersection over union) metrics for each segmentation are displayed.*

4.8 Discussion

Assessing oral health through imaging studies has become both highly precise and very common nowadays. Firstly, image scans experienced a quality enhancement in the last decades, and, second, the relatively low-cost of such imaging scans (i.e., digital panoramic X-ray (PAN) scans) increased and facilitated their accessibility by the general population. The facilitated access to oral images is fostering the usage of such images in research that correlates oral structures' conditions with systemic diseases, which has been enhancing the already established perception of oral health as a proxy for other body systems' conditions.

The mandible is one of the oral structures that has been gaining special attention since many studies had already associated some of its properties with osteoporosis, a bone disease commonly under-diagnosed that affects the entire world population [Allen et al. 2007, Alonso et al. 2011, Hastar, Yilmaz e Orhan 2011, Kavitha et al. 2012]. Studies of such kind require precise manual mandible segmentation and analysis, which is a time-consuming and training-required task. In this context, automatic mandible segmentation (AMS) is highly desirable. Indeed, AMS, as well as general automatic imaging analyzes, can avoid fatigue, user-variability, and clinical-hours waste, and facilitate mandible-related complex imaging analyzes to be translated into clinical practice.

Dental panoramic is the mainstay imaging technique for dental health monitoring [Watanabe, Watanabe e Tiossi 2012]. For that reason, most of the studies relating to mandible segmentation are based on PAN analyzes, for example. Nonetheless, PAN images offer many drawbacks regarding its imaging structures definition. As they are two-dimensional images, they suffer from superimposed anatomical structures. In addition, all panoramic radiographs suffer vertical and horizontal distortions that vary according to the anatomical position, variations in the positioning of the patient, and low contrast on some soft-to-solid structures interface. Furthermore, the distance from the focal object to the center of rotation of the X-ray scan systems cause the anterior region in panoramic radiography to undergo greater magnification [Zarch et al. 2011]. All those facts make both manual and automatic mandible segmentation a challenge.

Four studies on AMS on PAN images were found in the literature [Abdi, Kasaei e Mehdizadeh 2015, Naik et al. 2016, Hasan et al. 2016, Cha et al. 2021], Table 4.5. In the present study, we developed a set of deep learning-based algorithms combined with morphological refinements (Table 4.5 and 4.6) that achieved highly accurate performances for mandible segmentation on dental panoramic X-ray images outperforming all the results described on Table 4.5. The Ensemble + MR model achieved the highest performance on the test set: 98.27%, 97.60%, and 97.18%, for ACC, DICE, and IoU metrics, respectively.

All the previously published studies on AMS suffer from limitations that impoverish the generalization of their results, Table 4.5 . The absence of toothless patients, or patients with extended toothless regions in (Abdi et al., 2015) among the developing and tested patient group makes it unfitted for mandible segmentation on elderly people, where this condition is quite common making this algorithm unfitted for osteoporosis studies. (Hasan et al., 2016) and (Naik et al., 2016) do not output a segmentation that contains the whole mandible bone structure, which imitates geometrical feature extraction and imposes information losses for pixel intensity-based features extraction given the absence of some regions (e.g., superior left and right mandibular ramus). Lastly, (Cha et al., 2021) results have limited generalizability due to the small number of patients included in the study. Although the authors used the transfer learning technique to overcome such a small patient

Table 4.5: *Performances of the ensemble model on total validation and test sets with and without morphological refinement.*

Study	***Performances (%)***				***Limitations***
	ACC	*DICE*	*IoU*	*Other*	
(Abdi et al., 2015)	–	*93.22*	–	94.68[1]	Rely on a limited library of already segmented images; and did not consider complex cases: Toothless patients or patients with extended toothless regions.
(Naik et al., 2016)	–	–	–	90.00[2]	The final output is not the complete mandible structure segmentation, but stripes containing mandible and lower jaw edges.
(Hasan et al., 2016)	–	–	–	92.00[3]	The authors used a qualitative criterion to evaluate segmentation. Besides, the final segmentations included teeth and did not include left and right mandibular ramus.
(Cha et al., 2021)	–	–	*89.80*	–	A too-short number of patients. Final segmentation is a mix of the maxillary sinus, maxilla, mandibular canal, mandible, normal tooth, treated tooth, and dental implants rather than mandible alone.

[1]Specificity. The authors also expressed their performances in Sensitivity (94.44%). [2]The performance metric used by the authors is called Success Counts. It is a measure very specific to the study's method. [3]This metric refers to the Percentage of correctly segmented images out of the total amount of images tested. It is a qualitative criterion and, for that reason, has not much precision when describing the performance of segmentation algorithms.

dataset, they achieved only 89.80% on the IoU metric for mandible segmentation, which is a relatively low performance when it comes to semantic segmentation. Additionally, its final segmentation included simultaneously oral structures other

than mandible, which is uninteresting for mandible-specific studies and leads to the same limitation in (Hasan et al., 2016; Naik et al., 2016), an incomplete mandible structure segmentation coverage.

All the limitations pointed out for the studies previously published were solved by the segmentation algorithms presented here. The performances of our final segmentation algorithms show an improvement of 5.05% and 7.28% on DICE and IoU, respectively. Besides, our results were the only ones performed over PAN images from two different scan sources, and the ones using the largest dataset (509 image/segmentation pairs) among the AMS studies, which implies larger robustness to age, gender, oral condition variability, and scan-related image differences. Finally, the development of five deeply trained models (considering the ensemble) allows the calculation of five segmentations simultaneously, which can be useful when the best algorithm (Ensemble + MR) may fail.

4.9 Conclusions

We presented a set of deep learning-based algorithms to perform automatic mandible segmentation with an unseen accuracy and robustness that outperformed all the previously published results and offers a definite solution for AMS: the whole mandible bone structure outline, excluding teeth and other unwanted structures. Further, the large image and patient dataset, which included a wide range of age and gender representativeness, as well as the presence of extreme cases (e.g. toothless patients, treated teeth, dental implant, etc.) and different image sources enforce the robustness of the reported set of algorithms. Next, the segmentation performed by the presented algorithm is perfectly suitable for mandible-specific imaging studies allowing both geometric and pixel intensity-based features. In addition, the level of excellence achieved by the proposed methodology contributes significantly to the translation of automatic imaging analysis tools into the clinical practice improving imaging analyzes reproducibility and reducing human-related variability over such analysis and diagnosis.

The mandible segmentation model developed in this section was the key for the innovation proposed in this study. It was used to extract mandible region of

interest from dental panoramic images used in the osteoporosis risk assessment study. In the next chapter, we will go over the last experiments of this work that investigate the improvements of this disease risk assessment using a mandible-focused ROI to train the artificial intelligence models.

Chapter 5

Diagnosing osteoporosis risk through dental panoramic radiography using artificial intelligence models

In Chapter 2, we described in detail the aspects related to osteoporosis disease and its context. Chapter 3 describes what is Artificial Intelligence (AI) and how it has bee used in medicine as an auxiliary tool for diagnosis and prognosis of many diseases as well as the description of the logic behind the deep learning algorithms. Chapter 4 describes the end-to-end development carried to train an automatic mandible segmentation model, so that we could extract mandible image region form PAN images. In the present chapter, we will cover the final part of this investigation: an AI-based osteoporosis risk assessment focused on the entire mandible ROI extracted from dental panoramic radiography.

5.1 Patient group and image data

For this study, we curated 309 patients clinically followed at the Hospital das Clinicas da Faculdade de Medicina de Ribeirao Preto (HCFMRP) that had both DXA exams and PAN images acquired and stored in the hospital registries. The inclusion criteria: I) to have DXA exam for femoral-neck and spine site (Fig 5.1 a) and b)), and at least one dental panoramic radiography acquired(Fig 5.1 c)); II) the PAN image must have been acquired in period of at most six months before or after the DXA exams. Patients were excluded from the study when: I) their DXA

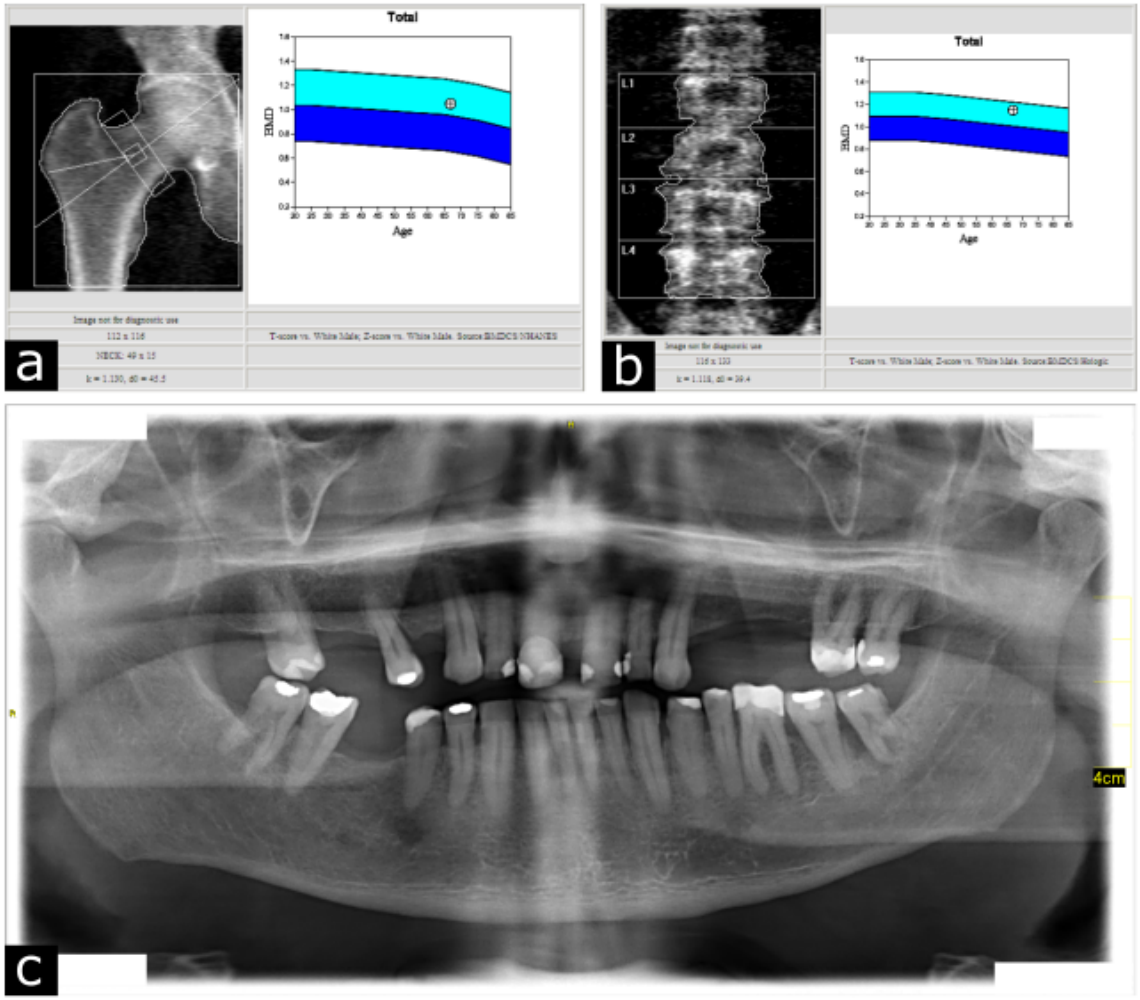

Figure 5.1: *Set of patient exams used in the present investigation. In a), it is the femur neck (hip region) DXA exam carried to measure the bone mineral density (BMD) and from that the T and Z-scores. In b) is the same DXA exam, but for the lumbar spine site (L1-L4). They are both Acquired at the same day. In c), we have the dental panoramic X-Ray acquired in a period of at most 6 months apart from DXA.*

were incomplete, i.e. it did not contain BMD measures, or T-scores were missing for either femur-neck or spine sites; II) The PAN and DXA were acquired with more than six months apart. No gender or age restrictions were considered during either patient including or excluding processes. Some patients had more than one set of PAN and DXA exams eligible for the study as they were followed up for prolonged period of time and were considered more than once in this study. Every eligible pair of PAN image and DXA exams were counted as a different sample. The table 5.1 describes all the patient data considered in this study.

Table 5.1: *Complete description of the patient data used for osteoporosis risk assessment.*

Gender	***N***	***Samples***	***Age at DXA (y.o.)***	***Min. Age (y.o.)****	***Max. Age (y.o.)****	***(%*)***
Male	207	257	69.61 ± 13.97	20.0	90.00	67.63
Female	102	123	55.37 ± 12.96	22.00	88.70	32.37
Overall	309	380	65.00 ± 15.18	20.00	90.00	100

*The Min. Age, Max. Age and Percentage (%) were calculated for all the samples (pair of DXA and PAN exams) included.

The DXA exams were acquired with two different machines. An Hologic 4500 W densitometer (Hologic Inc., Bedford, MA, USA) was used in the acquisition of 333/380 exams. And a GE Prodigy Densitometer (General Electric Company, Milwaukee, WI, USA) was used to acquire another 47/380 DXA exams. All the DXA exams were acquired at the femoral neck and for the lumbar spine (L1-L4) meaning that for each patient there were two bone density measures, two T1-scores, and two bone health diagnosis. All the PAN images were acquired in the same radiography scan used in the study presented in Chapter 3, the Sirona scan ORTOPHOS XG-3D/Ceph 60-90 kVp, and the PAN images were acquired at a 2432x1272 pixel resolution with 32-bit gray-level-intensity depth.

The major innovation of the present study relates to performing experiments to assess osteoporosis risk using only the *mandible image region* extracted from the original PAN images. To that end, we used the segmentation models developed in the first phase of this study, which were presented in chapter 3 and was recently published [Machado et al. 2023]. We extracted the mandible region for all the 380 PAN images included in the study creating a same-size data set composed only of mandible-segmented image regions (Figure 5.2). Same size images but containing only the mandible.

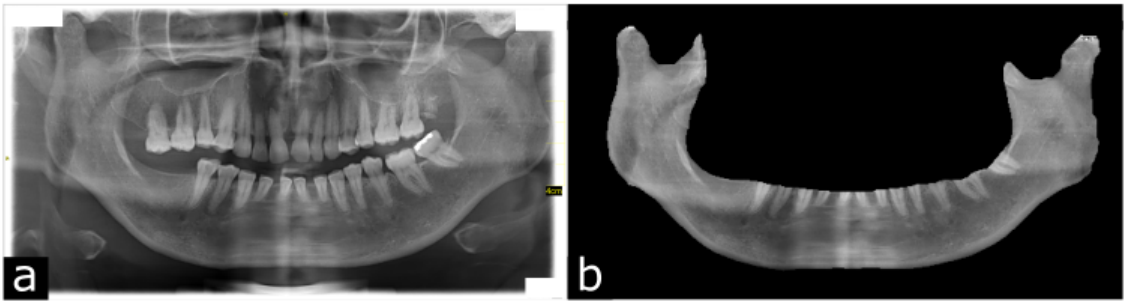

Figure 5.2: *Mandible image region extraction. In a), The original PAN image. In b) The same PAN image, but containing only the mandible.*

5.2 Disease risk: the outcome investigated

The goal of this study is to opportunistically access the osteoporosis risk in a patient using a PAN image of such patient, either with the raw PAN exam or only the mandible image region extracted from the image. So, before stepping into the experiment we need to explicitly define our outcome variable.

All the patients had DXA assessments for two sites: femoral-neck and lumbar spine. From now on, we are going to refer to those sites as *femur* and *spine*. For each site, this exam outputs a bone mineral density measure (BMD, g/cm^2) and the equivalent T-score. The T-score was used in accordance with WHO criteria (Table 2.2) to get a bone health diagnosis: *healthy*, *osteopenia*, and *osteoporosis*. Osteopenia and osteoporosis diagnosis were grouped in a single group called *disease risk*. This last one would contain all patients who would be considered either with osteoporosis or in its intermediate state. This approach allowed us to treat this investigation a binary classification problem: 0 - healthy, 1 - disease risk[1]. Having a DXA for two sites means that we can have two bone health diagnosis for the same patient and they can diverge. Table 5.2 bring the distribution of patient-samples per *diagnosis*/*site*.

[1]The advantages of such an approach will be further analyzed in the results and discussions session.

Table 5.2: *Complete description of the patient data used for osteoporosis risk assessment.*

DXA Site	***Healthy***	***Osteopenia***	***Osteoporosis***	***Disease Risk****
Femur	142	193	45	238
Spine	164	116	100	216

*This is not a WHO diagnosis but rather the grouping of two sets of patients with osteopenia and osteoporosis diagnosis so that we can use a binary classification approach and simplify the problem.

5.3 The experiments

Since we had two BMD measures (i.e., two diagnoses) for each patient, the deep learning experiments were set to achieve femur spine diagnosis separately. In this way, we performed four experiments to evaluate our thesis:

1. Separate healthy from disease-risk patients, according to **Femur diagnosis**, using the **original PAN image**;

2. Separate healthy from disease-risk patients, according to **Femur diagnosis**, using the **mandible image**;

3. Separate healthy from disease-risk patients, according to **Spine diagnosis**, using the **original PAN image**;

4. Separate healthy from disease-risk patients, according to **Spine diagnosis**, using the **mandible image**;

As they are stated, the goal and the context of those experiments are self-explained. We intend to investigate how well deep learning algorithms can classify patients in *healthy* and osteoporosis *disease risk* (binary problem) according to both femur and spine diagnosis (Table 5.2.In this way, we can evaluate which site's bone health (femur or spine) appears to be more correlated with the PAN image findings. By performing the experiments with original PAN and with mandible region only, we can investigate if we observe not just the general correlation between

pan image and osteoporosis diagnosis, but if this bone health information is present majorly in the mandible.

5.3.1 EfficientNetV2: The deep learning model used

For the current investigation, we opted to use the EfficientNNetV2 model for all the experiments here performed [Tan e Le 2021]. It is a convolutional neural network model that was published in 2021 that has scored 86.8% accuracy in ImageNet dataset and 98.7% in CIFAR-10 dataset for image classification task[2]. This model is promptly available in the TensorFlow model library version 2.10, what makes it extremely practical to set up and run experiments.

TensorFlow also brings pre-trained weights built-in for all the models in its model library. All the EfficientNetV2 model family (small, medium, and large) have pre-trained weights for the ImageNet dataset. It means that our models will make use of the *knowledge* learned in a different larger dataset. We had to make a few adaptations, though. As the the classification task in the ImageNet dataset is a problem of 1000 classes, the entire set of weights and model cannot be used. We remove the head of the model so that we can adapt to our task. This is the gist of transfer learning.

Figure 5.3 describes the architecture of the model we used for our experiments. The first component is the EfficientNetV2-L (with image net pre-trained weights) without ImageNet classification task head. The second part is the a fully convolutional planar network to treat the EfficientNet output to an output that fits the task (binary classification). Lastly is the output layer, a single neuron that outputs a number in the range [0,1], which is the probability of the input to belong the class 1 (disease risk).

5.3.2 Data separation, data imbalance, and data augmentation

To understand the experimentation that we are going to present, we need to pay especial attention to how we perform *data separation*, i.e., the number of samples in *train*, *validation*, and *test* set. It is important to stress that we are managing a

[2] thttps://paperswithcode.com/paper/efficientnetv2-smaller-models-and-fasterext

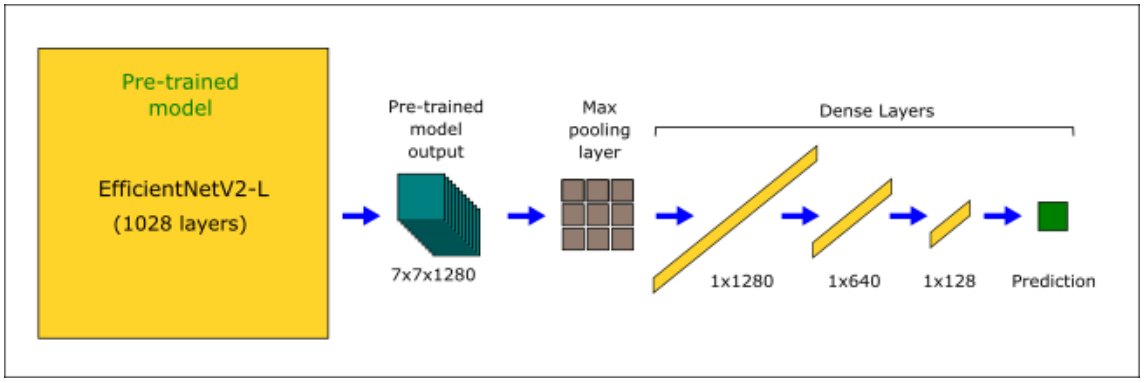

Figure 5.3: *Deep learning experiment architecture used in this study. The pre-trained model's output is passed through a planar series of dense layers, and a single number is output.*

very imbalanced data (Table 5.2), i.e., the number of samples in each class are quite different for both Femur and Spine diagnosis. We have much less image samples in the healthy group. Henceforth, an effort will be carried out to separate the data in a way that all the training datasets contain same amounts of each class, which implies managing the data imbalance.

The data separation is illustrated in Figure 5.4. We defined the proportion 76%/12%/12% for train/validation/test sets. Initially, we performed data separation for healthy and disease risk groups separately to later reunite them. Further, we over-sampled an amount of image samples for the class with less images (healthy group, check Table 5.2). Next, we applied an augmentation transformation (same as in Figure 4.4) over those oversampled images. After, we added the augmented data back to the respective sets. Finally, we added each class set to have a final train, validation, and test sets.

The process illustrated in Figure 5.4 refers to Experiment 1. It is the data separation according to Femur diagnosis using original PAN images. The same approach was used to separate the data and correct class imbalance in the other three experiments. Table 5.3 brings the data separation per image set (train, validation, and test) and per diagnosis in both original and balanced data distribution.

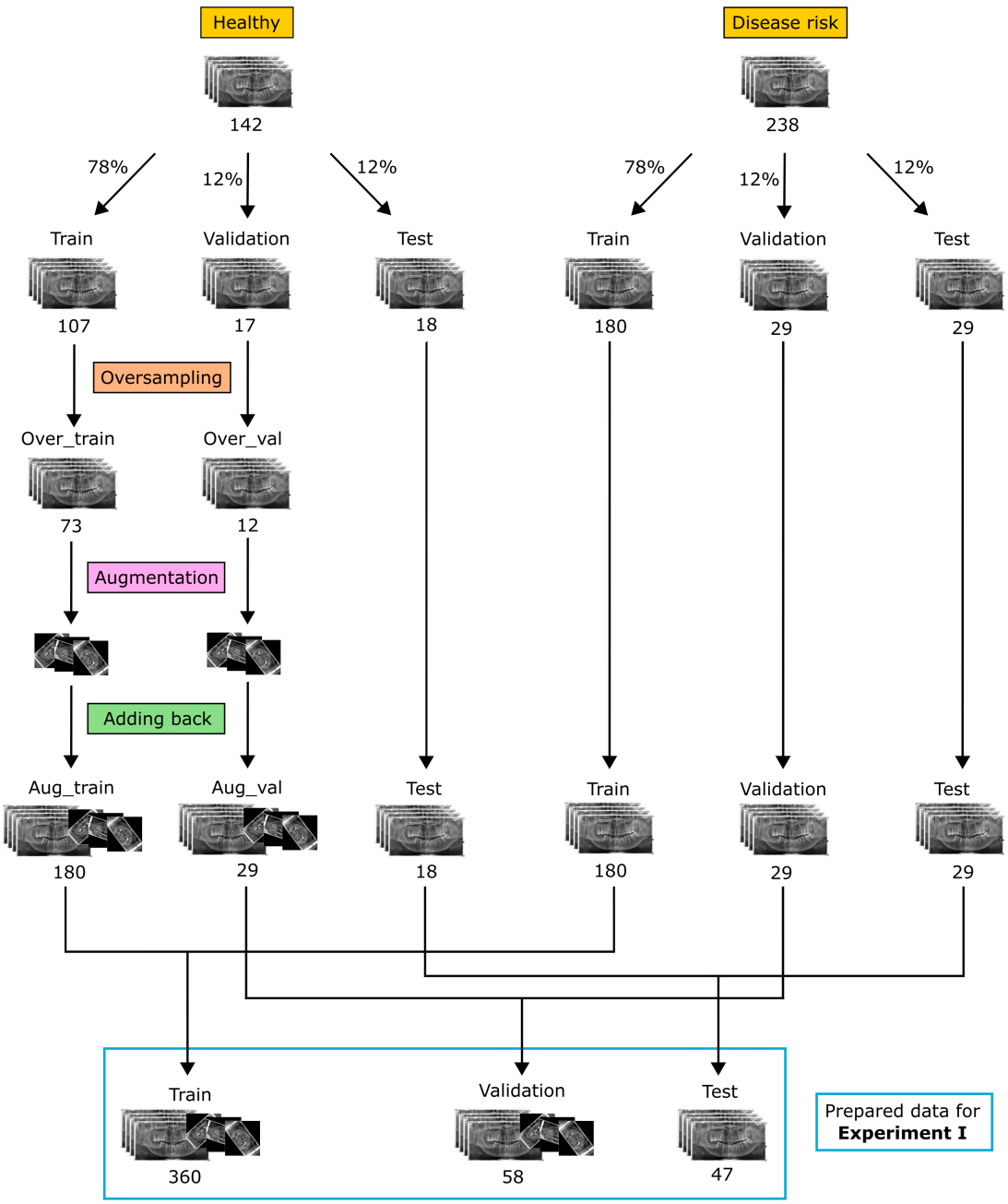

Figure 5.4: *Data separation, oversampling and augmentation for Experiment I - Femur diagnosis using complete PAN images. The other experiments used the same data separation scheme.*

5.3.3 Metrics and hyper-parameters

Four metrics were used to evaluate the performance of the model. Accuracy (ACC), balanced accuracy (BACC), precision (PRE), and recall (REC). Following

Table 5.3: *Data separation for the Femur and Spine diagnosis distribution.*

Dataset	***Diagnosis***	***Train***		***Validation***		***Test***	
*Classes** →		0	1	0	1	0	1
Original	*Femur*	107	180	17	29	18	29
Balanced		180	180	29	29	18	29
Original	*Spine*	124	164	20	26	20	26
Balanced		164	164	26	26	20	26

* The classes refer to the study groups: 0 - Healthy; 1 - Disease Risk. The data was balanced using both oversampling and data augmentation technique as illustrated in Fig. 5.4.

the mathematical formulation of each one of them:

$$ACC(y_{true}, y_{pred}) = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

$$BACC(y_{true}, y_{pred}) = \frac{1}{n_{classes}} \sum_{n=1}^{n_{classes}} \frac{TP_n}{TP_n + FN_n} \tag{5.2}$$

$$PRE(y_{true}, y_{pred}) = \frac{TP}{TP + FP}. \tag{5.3}$$

$$REC(y_{true}, y_{pred}) = \frac{TP}{TP + FN}. \tag{5.4}$$

y_{pred} refers to the predicted diagnosis by the model and y_{true} is the patient's diagnosis according to the respective site DXA. $n_{classes}$ refers to the number of classes in the problem (2). The component rates are defined as follows: *TP*, true positive rate, the number of samples/images correctly classified as *disease risk* group. *TN*, true negative rate, the number of samples correctly classified as *healthy*; *FP*, false-positive rate, means the number of samples misclassified as *disease risk*; and *FN*, false-negative rate, the number of samples misclassified as *healthy*. For ACC, BACC, PRE, and REC metrics, 1 indicates the perfect performance and 0 the worst.

Balanced Accuracy[3] is an important metric to use here given the reality of data imbalance in the present study. BACC gives us a measure of how well the model is doing without any bias coming from more frequent classes, which tend to be easier to predict. Accuracy, when compared to BACC, tells us if the model is succeeding better in one class or another. If BACC $\approx$ ACC, means the algorithm is performing similarly on both classes. If BACC $<$ ACC, means the model is predicting better for the most frequent class, while if BACC $>$ ACC, the model must be predicting better in the least frequent class.

For all those experiments, the *Adam* optimizer and the *binary-cross-entropy* loss function was used. All models were trained for 400 epochs and the best model was selected by checking the validation accuracy (ACC) metric. During training, we could only calculate accuracy (ACC), since the data imbalance was corrected with data augmentation (Figure 5.4), ACC had same effect as BACC. However, in the evaluation stage, we used BACC to evaluate the performance on *test* set and compare the results for all the experiments.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html

5.4 Results

In this section, we are going to present the performances achieved by the deep learning models for the Femur and Spine diagnosis using complete PAN and mandible region image. Next, Table 5.4 brings the results for all the four experiments described in session 5.3.

Let's analyze the results per diagnosis. We will Look at Femur first (Exp. 1 and 2), Table 5.4. We can see that the deep-learning model obtained the same ACC (0.723) using Complete PAN and Mandible Image and that the BACC is higher for complete PAN (Exp. 1). However, for Exp. 1, BACC > ACC, which means that the model performed better in the less frequent class (0, healthy) than in the more frequent and more important class, the disease risk group (1). It can also be observed by comparing the PRE and REC for Exp. 1 and 2. As we can see, Exp. 2 achieved higher REC (0.830) than Exp. 1 (0.660). It means that Exp. 2's model can detect more disease risk patients, making it better than Exp. 1's model.

Now Looking at Spine diagnosis, we see that Exp. 4, in which we use mandible image to separate healthy from disease risk patients, produced much better

Table 5.4: *Deep Learning models' performances for osteoporosis risk assessment.*

Exp. #	***Diagnosis site***	***Image***	*Epochs**	***BACC***	***ACC***	***PRE***	***REC***
1	*Femur*	Complete PAN	213	0.744	0.723	0.860	0.660
2	*Femur*	Mandible Image	367	0.691	0.723	0.750	0.830
3	*Spine*	Complete PAN	62	0.556	0.543	0.630	0.460
4	*Spine*	Mandible Image	36	0.735	0.739	0.770	0.770

*This value points to the epoch in which the model achieved the highest validation set BACC/ACC score during training. All the models were trained for 400 epochs. BACC: Balanced Accuracy; ACC: Accuracy; PRE: Precision for class 1 - Disease risk; REC: Recall for class 1 - Disease risk.

results (BACC: 0.735, ACC: 0.739, PRE: 0.770, REC: 0.770) than the model trained with Complete PAN image (BACC: 0.556, ACC: 0.543, PRE: 0.630, REC: 0.460), for every metric used. Also, Exp. 4 achieved the highest Accuracy (0.739) among the four experiments performed.

Another important result is the epoch in which the best result was achieved. We can observe that the Deep-learning algorithm achieved the best performance for Spine diagnosis faster (62 and 36 epochs) than for Femur Diagnosis (213 and 367 epochs). And that, for Spine diagnosis, Mandible Image experiment achieved the best results faster (36 epochs) than Complete PAN image experiment (62 epochs), while for Femur diagnosis, the opposite was observed.

Those results point to a correlation between the mandible region image and the BMD levels assessed in both Femur and Spine sites, since the performances for the models trained with Mandible Image were better than the performance of the models trained with the entire image. Also, this correlation is better observed for the Spine site, since the deep-learning models achieved their best performances much faster for Spine diagnosis.

In the next session, we will discuss the presented results in contrast with thesis questions and the results already published on the diagnosis of osteoporosis using dental panoramic X-ray and deep learning models.

5.5 Discussions

Dental Panoramic X-ray (PAN) imaging is proving to be a very promising image modality to assess osteoporosis condition. Many studies have correlated oral bone conditions, easily observed in PAN images, with BMD level changes, henceforth the osteoporosis diagnosis [Watanabe, Watanabe e Tiossi 2012, Klemetti, Kolmakov e Kröger 1994, Franciotti et al. 2021, Valentinitsch et al. 2019, Ollivier et al. 2013, WATANABE et al. 2008, Watanabe et al. 2022]. Mandible is one of the oral bones that has gained considerable interest from researchers as such correlations continue to be observed. Some studies have used PAN images in combination with deep learning models to perform osteoporosis diagnosis. Let us briefly review the main findings of such studies.

In a preliminary study, Lee *et al* (2019) trained convolutional neural networks (fully convolutional without pre-training) using rectangular ROIs extracted from the region below teeth on PAN images to separate old, post-menopausal (72.2 $\pm$ 8.5 y.) osteoporotic female from young, healthy (32.8 $\pm$ 12.1 y.) female patients [Lee et al.]. In this study, the osteoporosis diagnoses were obtained using Klemetti criteria [Klemetti, Kolmakov e Kröger 1994]. It achieved 98.5% accuracy for its best model. The study group is very clearly separated not just by diagnosis, but also by age and restricts itself to female group. The cropping strategy they choose is very well directed by the Klemetti criteria that evaluate erosion in the inferior mandibular cortex (border). It means that the deep-learning algorithms are reproducing a classification that is intrinsically present in the image. It brings no new knowledge in the scene of osteoporosis analysis through PAN images since the classification criteria is already visible in the image. The algorithms in this study study are likely just reproducing oral radiologists diagnoses rather than uncovering any underlying relations between PAN images and osteoporosis disease.

Ling *et al* (2020) used the entire PAN images as inputs to their deep learning models to separate osteoporotic from normal subjects [LING e YANG 2020]. This study reportedly achieved 92.0% accuracy using transfer learning and leave-one-out cross validation approach. The authors used only 108 patients in their entire

analysis, but provided no further information on patient demographics, data stratification nor the deep learning architecture used. The absence of such information leaves no space for comparing their results with the ones available in the literature.

Lee *et al* (2020) tested a variety of deep-learning models with dental panoramic X-rays to achieve osteoporosis diagnosis according to WHO criteria [Lee et al. 2020]. The patients were separated into non-osteoporosis (healthy and osteopenia) and osteoporosis patients. The study does not explain which site was used for BMD assessment. The authors evaluated the improvements obtained by using transfer learning on the osteoporosis diagnosis task. The best model achieved 84.0% ACC and 90.0% REC using transfer learning. Further, they investigated which region of the image ROI they used contained the most relevant information for deep learning algorithms to make their decisions. According to authors, the right and left mandibular cortex were the most relevant part of the images for the correct predictions, which is aligned with many previous studies. However, they used only a rectangular PAN ROI containing only the mandibular cortex for training the algorithms and producing an importance-based feature map. Additionally, their conclusions were based on randomly spotted cases rather than an average importance-based feature map of over all the predictions, what makes their conclusions less generalized. Nevertheless, the image regions pointed by this study as the most correlated with the osteoporosis (border of the mandibular cortical bone) is in accordance with other studies that had defined radiomorphometric indices based on those same regions, e.g., mandibular cortical index and Klemetti classification.

Sukegawa *et al* (2022) assessed osteoporosis diagnosis through pan images using EfficientNet and ResNet deep learning models alone and as an ensemble combined with clinical parameters (age, weight, and body mass index (BMI)) [Sukegawa et al. 123]. In this study, the authors also used a manually drawn rectangular cropped region that covers the mandibular cortical bone. The osteoporosis classification was obtained from Spine and Hip DXA using the WHO criteria (Table 2.2). The group was divided in non-osteoporosis and osteoporosis, with osteopenia being included in the non-osteoporosis group. The models achieved 83.2 and 71.6% (ACC and REC) for image-only single model approach and 84.5

and 74.9% (ACC and REC) for ensemble combined with clinical features. This study brings the innovation of combining clinical data with the AI model features in an ensemble approach to perform osteoporosis if compared to [Lee et al. 2020]. However, despite the considerable improvements, the final performance is basically the same (84.5% against 84.0%).

The present study brings the possibility of an evaluation using the entire mandibular bone imaged in a dental panoramic X-ray as its greatest innovation. The automatic mandible segmentation model developed in Chapter 4 ([Machado et al. 2023]) provides us with an image with same dimensions as the original one but containing only the mandible bone. This segmentation model was used to extract the mandible for every patient image in the present study. The EfficientNetV2-L was used to separate patients with risk of disease from those who did not have any risk of disease, according to two different DXA analyzes: femur-neck and Spine. We tested the diagnosis performance of the algorithms for both complete PAN images and the mandible-only image. The best models we trained achieved 73.9% ACC, and 83.0% REC. In our experiments we could also observe that the models trained with mandible image achieved the best results, what indicates that using mandibular segmentation ROI offer gains in comparison to use the entire PAN. Further, since it contains the mandible image portion captured by the rectangles used in the previous studies ([Lee et al., Lee et al. 2020, Sukegawa et al. 123]) and other important regions such as the oblique line and mandibular ramus [Watanabe et al. 2022], it can be a better alternative to rectangular ROIs specially as it can be automatically generated. We also identified a slightly better correlation between the Spine diagnosis and the mandible, since the algorithms converged much faster for Spine diagnosis. In fact, this is another originality brought by the present analysis.

The short number of image-diagnosis samples was one limitation faced in this research. Also, for the effect of a precise comparative between our results/approach and the previous studies, we would have to reproduce the rectangular ROI extraction and trained the same models under the same conditions so that we could directly compare the ROI approach. However, this procedure would demand more time and would possibly fall out of the scope of this study. An important difference

on our study is that it focused on diagnosing the risk of the disease rather than the disease alone. It means that our model can opportunistically identify patients with osteoporosis or osteopenia. Those could be forwarded for DXA exams. This approach leads us to a less precise tool but with a higher recall.

5.6 Conclusions

We presented a carefully and detailed PAN-image mandible-focused analysis for assessing the risk of osteoporosis using artificial intelligence. This study showed that the mandible segmentation improves the accuracy of deep learning models against the approach that uses the entire image. No previous study in the literature carried out an investigation considering the entire mandible bone ROI on PAN images. Further, this mandible segmentation was obtained automatically and contains the same mandible rectangular ROIs used in previous studies with the addition of other important regions for bone health assessment, such as the oblique line and the mandibular ramus. Those facts make this approach a better alternative to the rectangular ROIs. Certainly, a future study would complement our findings by collecting more image data, performing hyper-parameter tuning for a better architecture, ensemble combination, and exploring architectures that mix clinical parameters (e.g., age, gender, weight, etc.) and the deep learning outcomes. Those improvements can potentially yield performances superior to the ones published so far for the osteoporosis diagnosis or for the osteoporosis risk assessment task.

Bibliography*

[Abdi e Kasaei 2017]ABDI, A.; KASAEI, S. Panoramic dental x-rays with segmented mandibles. *Mendeley Data*, 2017.

[Abdi, Kasaei e Mehdizadeh 2015]ABDI, A. H.; KASAEI, S.; MEHDIZADEH, M. Automatic segmentation of mandible in panoramic x-ray. *Journal of Medical Imaging*, Society of Photo-Optical Instrumentation Engineers, v. 2, p. 044003, 11 2015. ISSN 2329-4302. Disponível em: </pmc/articles/PMC4652330/ /pmc/articles/PMC4652330/?report=abstract https://www-ncbi-nlm-nih.ez67.periodicos.capes.gov.br/pmc/articles/PMC4652330/>.

[Allen et al. 2007]ALLEN, P. D. et al. Detecting reduced bone mineral density from dental radiographs using statistical shape models. *IEEE Transactions on Information Technology in Biomedicine*, IEEE Trans Inf Technol Biomed, v. 11, p. 601–610, 11 2007. ISSN 10897771. Disponível em: <https://pubmed-ncbi-nlm-nih.ez67.periodicos.capes.gov.br/18046935/>.

[Alonso et al. 2011]ALONSO, M. B. C. C. et al. Assessment of panoramic radiomorphometric indices of the mandible in a brazilian population. *ISRN Rheumatology*, v. 2011, p. 1–5, 9 2011. ISSN 2090-5467. Disponível em: <https://www.hindawi.com/journals/isrn/2011/854287/>.

[Aziziyeh et al. 2019]AZIZIYEH, R. et al. The burden of osteoporosis in four latin american countries: Brazil, mexico, colombia, and argentina. *Journal of Medical Economics*, Taylor and Francis

*De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

Ltd, v. 22, p. 638–644, 7 2019. ISSN 1941837X. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/13696998.2019.1590843>.

[Briganti e Moine 2020]BRIGANTI, G.; MOINE, O. L. Artificial intelligence in medicine: Today and tomorrow. *Frontiers in Medicine*, v. 7, 2 2020. ISSN 2296-858X. Disponível em: <https://www.frontiersin.org/article/10.3389/fmed.2020.00027/full>.

[Burge et al. 2007]BURGE, R. et al. Incidence and economic burden of osteoporosis-related fractures in the united states, 2005-2025. *Journal of Bone and Mineral Research*, v. 22, p. 465–475, 3 2007. ISSN 08840431.

[Cha et al. 2021]CHA, J. Y. et al. Panoptic segmentation on panoramic radiographs: Deep learning-based segmentation of various structures including maxillary sinus and mandibular canal. *Journal of Clinical Medicine 2021, Vol. 10, Page 2577*, Multidisciplinary Digital Publishing Institute, v. 10, p. 2577, 6 2021. ISSN 20770383. Disponível em: <https://www.mdpi.com/2077-0383/10/12/2577/htm https://www.mdpi.com/2077-0383/10/12/2577>.

[Chan et al. 2018]CHAN, Y. K. et al. *Artificial Intelligence in Medical Applications*. [S.l.]: Hindawi Limited, 2018.

[Christodoulou e Cooper 2003]CHRISTODOULOU, C.; COOPER, C. *What is osteoporosis?* The Fellowship of Postgraduate Medicine, 3 2003. 133-138 p. Disponível em: <http://pmj.bmj.com/>.

[Compston, McClung e Leslie 2019]COMPSTON, J. E.; MCCLUNG, M. R.; LESLIE, W. D. *Osteoporosis*. Lancet Publishing Group, 1 2019. 364-376 p. Disponível em: <www.thelancet.com>.

[Cruz et al. 2018]CRUZ, A. S. et al. *Artificial intelligence on the identification of risk groups for osteoporosis, a general review*. BioMed Central Ltd., 1 2018. 12 p. Disponível em: <https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-018-0436-1>.

[Economic e Affairs 2017]ECONOMIC, N. Y. N. U. N. U. D. of; AFFAIRS, S. World population ageing 2017: Highlights. 2017.

[Fedorov et al. 2012]FEDOROV, A. et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, v. 30, p. 1323–1341, 2012. ISSN 0730725X.

[Ferizi et al. 2019]FERIZI, U. et al. Artificial intelligence applied to osteoporosis: A performance comparison of machine learning algorithms in predicting fragility fractures from mri data. *Journal of Magnetic Resonance Imaging*, John Wiley and Sons Inc., v. 49, p. 1029–1038, 4 2019. ISSN 1053-1807. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.26280>.

[Franciotti et al. 2021]FRANCIOTTI, R. et al. *Use of fractal analysis in dental images for osteoporosis detection: a systematic review and meta-analysis*. [S.l.]: Springer Science and Business Media Deutschland GmbH, 2021.

[Goodfellow, Bengio e Courville 2016]GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. 800 p. ISBN 0262035618.

[Hameed et al. 2021]HAMEED, B. Z. et al. Engineering and clinical use of artificial intelligence (ai) with machine learning and data science advancements: radiology leading the way for future. *Therapeutic Advances in Urology*, SAGE PublicationsSage UK: London, England, v. 13, p. 1–12, 1 2021. ISSN 1756-2872. Disponível em: <https://journals-sagepub-com.ez67.periodicos.capes.gov.br/doi/full/10.1177/17562872211044880 http://journals.sagepub.com/doi/10.1177/17562872211044880>.

[Harrar et al. 2012]HARRAR, K. et al. Osteoporosis assessment using multilayer perceptron neural networks. In: . [S.l.: s.n.], 2012. p. 217–221. ISBN 9781467325837.

[Hasan et al. 2016]HASAN, M. M. et al. Automatic segmentation of jaw from panoramic dental x-ray images using gvf snakes. *World Automation Congress Proceedings*, IEEE Computer Society, v. 2016-Octob, 10 2016. ISSN 21544832.

[Hastar, Yilmaz e Orhan 2011]HASTAR, E.; YILMAZ, H. H.; ORHAN, H. Evaluation of mental index, mandibular cortical index and panoramic mandibular index on dental panoramic radiographs in the elderly. *European Journal of Dentistry*, Dental Investigations Society, v. 5, p. 60–67, 2011. ISSN 13057464. Disponível em: </pmc/articles/PMC3019752/ /pmc/articles/PMC3019752/?report=abstract https://www-ncbi-nlm-nih.ez67.periodicos.capes.gov.br/pmc/articles/PMC3019752/>.

[Ho-Le et al. 2017]HO-LE, T. P. et al. Prediction of hip fracture in post-menopausal women using artificial neural network approach. In: . [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2017. p. 4207–4210. ISBN 9781509028092. ISSN 1557170X.

[Johnell et al. 1997]JOHNELL, O. et al. The socioeconomic burden of fractures: Today and in the 21st century. In: . Elsevier Inc., 1997. v. 103. ISSN 00029343. Disponível em: <https://pubmed-ncbi-nlm-nih.ez67.periodicos.capes.gov.br/9302894/>.

[Kanis]KANIS, J. A. *FRAX - Fracture Risk Assessment Tool*. Disponível em: <https://www.sheffield.ac.uk/FRAX/>.

[Kanis et al. 2008]KANIS, J. A. et al. European guidance for the diagnosis and management of osteoporosis in postmenopausal women. *Osteoporosis International*, v. 19, p. 399–428, 4 2008. ISSN 0937-941X. Disponível em: <http://link.springer.com/10.1007/s00198-008-0560-z>.

[Kanis 2007]KANIS, J. A. on behalf of the W. H. O. S. G. . Assessment of osteoporosis at the primary healthcare level. who scientific group technical report. *Sheffield, UK: WHO Collaborating Centre for Metabolic Bone Diseases, University of Sheffield*, 2007.

[Kavitha et al. 2012]KAVITHA, M. S. et al. Diagnosis of osteoporosis from dental panoramic radiographs using the support vector machine method in a computer-aided system. *BMC Medical Imaging*, BioMed

Central, v. 12, p. 1, 1 2012. ISSN 14712342. Disponível em: <http://bmcmedimaging.biomedcentral.com/articles/10.1186/1471-2342-12-1>.

[Kavitha et al. 2013]KAVITHA, M. S. et al. The combination of a histogram-based clustering algorithm and support vector machine for the diagnosis of osteoporosis. *Imaging Science in Dentistry*, v. 43, p. 153, 9 2013. ISSN 2233-7822. Disponível em: <https://isdent.org/DOIx.php?id=10.5624/isd.2013.43.3.153>.

[Klemetti, Kolmakov e Kröger 1994]KLEMETTI, E.; KOLMAKOV, S.; KRöGER, H. Pantomography in assessment of the osteoporosis risk group. *European Journal of Oral Sciences*, John Wiley & Sons, Ltd, v. 102, p. 68–72, 2 1994. ISSN 16000722.

[Lecun, Bengio e Hinton 2015]LECUN, Y.; BENGIO, Y.; HINTON, G. *Deep learning*. [S.l.]: Nature Publishing Group, 5 2015. 436-444 p.

[LeCun et al. 1998]LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, p. 2278–2323, 1998. ISSN 00189219.

[Lee et al.]LEE, J. S. et al. Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: A preliminary study. *Dentomaxillofacial Radiology*, British Institute of Radiology, v. 48. ISSN 1476542X.

[Lee et al. 2020]LEE, K. S. et al. Evaluation of transfer learning with deep convolutional neural networks for screening osteoporosis in dental panoramic radiographs. *Journal of clinical medicine*, J Clin Med, v. 9, 2 2020. ISSN 2077-0383. Disponível em: <https://pubmed-ncbi-nlm-nih.ez67.periodicos.capes.gov.br/32024114/>.

[Lindsey et al. 2018]LINDSEY, R. et al. Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences of the United States of America*, National Academy of Sciences, v. 115, p. 11591–11596, 11 2018. ISSN 10916490.

[LING e YANG 2020]LING, H.; YANG, J. Osteoporosis prescreening using dental panoramic radiography with deep learning. *Oral*

Surgery, Oral Medicine, Oral Pathology and Oral Radiology, Elsevier BV, v. 130, p. e72–e73, 8 2020. ISSN 22124403. Disponível em: <http://www.oooojournal.net/article/S2212440320301267/fulltext http://www.oooojournal.net/article/S2212440320301267/abstract https://www.oooojournal.net/article/S2212-4403(20)30126-7/abstract>.

[Lorentzon e Cummings 2015]LORENTZON, M.; CUMMINGS, S. R. Osteoporosis: The evolution of a diagnosis. *Journal of Internal Medicine*, Blackwell Publishing Ltd, v. 277, p. 650–661, 6 2015. ISSN 13652796. Disponível em: <https://pubmed-ncbi-nlm-nih.ez67.periodicos.capes.gov.br/25832448/>.

[Lowekamp et al. 2013]LOWEKAMP, B. C. et al. The design of simpleitk. *Frontiers in Neuroinformatics*, Frontiers, v. 7, p. 45, 12 2013. ISSN 16625196.

[Machado et al. 2023]MACHADO, L. F. et al. Deep learning for automatic mandible segmentation on dental panoramic x-ray images. *Biomedical Physics & Engineering Express*, v. 9, p. 035015, 5 2023. ISSN 2057-1976. Disponível em: <https://iopscience.iop.org/article/10.1088/2057-1976/acb7f6>.

[Naik et al. 2016]NAIK, A. et al. Automatic segmentation of lower jaw and mandibular bone in digital dental panoramic radiographs. *Indian Journal of Science and Technology*, The Indian Society of Education and Environment, v. 9, p. 1–6, 6 2016. ISSN 0974-5645.

[Odén et al. 2015]ODÉN, A. et al. Burden of high fracture probability worldwide: secular increases 2010-2040. *Osteoporosis International*, Springer London, v. 26, p. 2243–2248, 9 2015. ISSN 14332965.

[Ollivier et al. 2013]OLLIVIER, M. et al. Radiographic bone texture analysis is correlated with 3d microarchitecture in the femoral head, and improves the estimation of the femoral neck fracture risk when combined with bone mineral density. *European Journal of Radiology*, Elsevier, v. 82, p. 1494–1498, 9 2013. ISSN 0720048X.

[Pranata et al. 2019]PRANATA, Y. D. et al. Deep learning and surf for automated classification and detection of calcaneus fractures in ct images. *Computer Methods*

and Programs in Biomedicine, Elsevier Ireland Ltd, v. 171, p. 27–37, 4 2019. ISSN 18727565.

[Rastegar et al. 2020]RASTEGAR, S. et al. Radiomics for classification of bone mineral loss: A machine learning study. *Diagnostic and Interventional Imaging*, Elsevier Masson SAS, v. 101, p. 599–610, 9 2020. ISSN 22115684.

[Ronneberger, Fischer e Brox 2015]RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. *arXiv*, p. 16591–16603, 5 2015. Disponível em: <https://arxiv.org/abs/1505.04597v1>.

[Sambrook e Cooper 2006]SAMBROOK, P.; COOPER, C. *Osteoporosis*. [S.l.]: Elsevier Limited, 6 2006. 2010-2018 p.

[Sapthagirivasan e Anburajan 2013]SAPTHAGIRIVASAN, V.; ANBURAJAN, M. Diagnosis of osteoporosis by extraction of trabecular features from hip radiographs using support vector machine: An investigation panorama with dxa. *Computers in Biology and Medicine*, Pergamon, v. 43, p. 1910–1919, 11 2013. ISSN 00104825.

[Scanlan et al. 2018]SCANLAN, J. et al. Detection of osteoporosis from percussion responses using an electronic stethoscope and machine learning. *Bioengineering*, MDPI AG, v. 5, 12 2018. ISSN 23065354.

[Sozen, Ozisik e Basaran 2017]SOZEN, T.; OZISIK, L.; BASARAN, N. C. An overview and management of osteoporosis. *European Journal of Rheumatology*, AVES Publishing Co., v. 4, p. 46–56, 3 2017. ISSN 21479720. Disponível em: </pmc/articles/PMC5335887/ /pmc/articles/PMC5335887/?report=abstract https://www-ncbi-nlm-nih.ez67.periodicos.capes.gov.br/pmc/articles/PMC5335887/>.

[Sukegawa et al. 123]SUKEGAWA, S. et al. Identification of osteoporosis using ensemble deep learning model with panoramic radiographs and clinical covariates. 123. Disponível em: <https://doi.org/10.1038/s41598-022-10150-x>.

[Tan e Le 2021]TAN, M.; LE, Q. V. Efficientnetv2: Smaller models and faster training. 2021. Disponível em: <https://github.com/google/>.

[Valentinitsch et al. 2019]VALENTINITSCH, A. et al. Opportunistic osteoporosis screening in multi-detector ct images via local classification of textures. *Osteoporosis International*, Springer London, v. 30, p. 1275–1285, 6 2019. ISSN 14332965.

[Wang et al. 2020]WANG, J. et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Computer Society, v. 43, p. 3349–3364, 8 2020. ISSN 19393539. Disponível em: <https://arxiv.org/abs/1908.07919v2>.

[Wani e Arora 2020]WANI, I. M.; ARORA, S. Deep neural networks for diagnosis of osteoporosis: A review. In: . Springer, 2020. v. 597, p. 65–78. ISBN 9783030294069. ISSN 18761119.

[WATANABE et al. 2008]WATANABE, P. C. A. et al. Avaliação da densidade mineral óssea na maxila e mandíbula utilizando-se radiografias periapicais. *Revista da ABRO - Associação Brasileira de Radiologia Odontológica*, v. 9, p. 52–60, 2008.

[Watanabe et al. 2022]WATANABE, P. C. A. et al. Oblique line contrast: A new radiomorphometric index for assessing bone quality in dental panoramic radiographs. *Heliyon*, Elsevier, v. 8, p. e12266, 12 2022. ISSN 2405-8440.

[Watanabe, Watanabe e Tiossi 2012]WATANABE, P. C. A.; WATANABE, M. G. de C.; TIOSSI, R. *How Dentistry Can Help Fight Osteoporosis*. 2012. 821-852 p.

[Yamamoto et al. 2020]YAMAMOTO, N. et al. Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates. *Biomolecules*, MDPI AG, v. 10, p. 1–13, 11 2020. ISSN 2218273X.

[Zarch et al. 2011]ZARCH, S. H. H. et al. Evaluation of the accuracy of panoramic radiography in linear measurements of the jaws. *Iranian Journal of Radiology*, v. 8, p. 97–102, 2011.

[Zhang et al. 2020]ZHANG, B. et al. Deep learning of lumbar spine x-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study. *Bone*, Elsevier Inc., v. 140, p. 115561, 11 2020. ISSN 87563282.