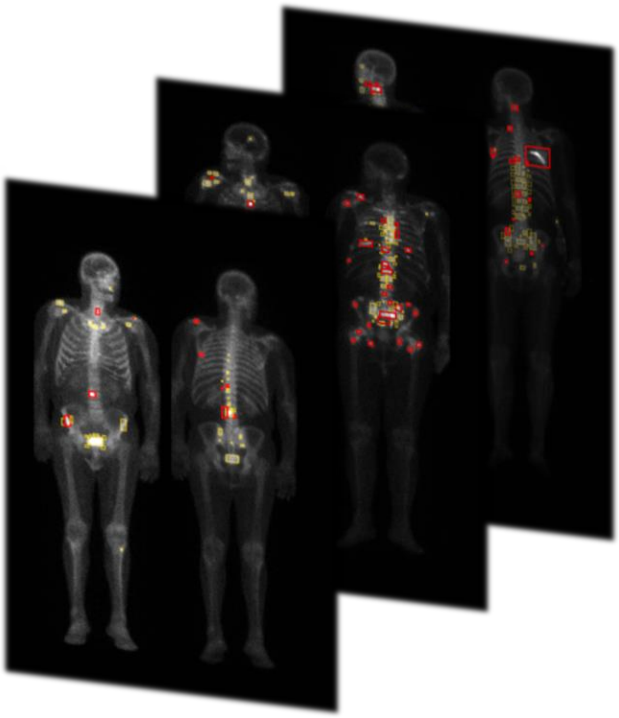


Bone scan lesions uptake quantification for therapy response in metastatic prostate cancer

Laura Providência

Mestrado em Física Médica

Departamento de Física e Astronomia da Faculdade de Ciências da Universidade do Porto

2020/21

Orientadora

Inês Domingues, Medical Physics, Radiobiology and Radiation Protection Group, IPO Porto Research Centre (CI-IPOP)

Coorientador

João Santos, Medical Physics, Radiobiology and Radiation Protection Group, IPO Porto Research Centre (CI-IPOP); Instituto de Ciência Biomédicas Abel Salazar

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

UNIVERSIDADE DO PORTO

MASTERS THESIS

Bone scan lesions uptake quantification for therapy response in metastatic prostate cancer

Author:
Laura PROVIDÊNCIA

Supervisor:
Inês DOMINGUES

Co-supervisor:
João SANTOS

A thesis submitted in fulfilment of the requirements
for the degree of MSc. Medical Physics

at the

Faculdade de Ciências da Universidade do Porto
Departamento de Física e Astronomia

December 17, 2021

Acknowledgements

The writing of this dissertation would not have been possible without all the support and assistance I have received from several people.

I would like to thank my supervisor Professor Inês Domingues and co-superviser Professor João Santos for all their help. In particular, I would like to thank Professor Inês Domingues for her incessant support and feedback through out these months. I am extremely grateful for having her as a supervisor and I know that without her insight and constant guidance this dissertation would not have been possible. To Professor João Santos I want to thank his contribution with his knowledge, namely in clarifying specific concepts that were very relevant for this work.

A special thanks goes to Inês and Sara not only for this past year but for the whole master. Their friendship and support made a master, 3/4 of which took place during a pandemic, much easier. Our numerous chats have helped me more than they can ever imagine and made me always realise that we were all on the same (sinking ☺) boat.

Finally, I would like to thank my parents, without whose support this master would not have been possible. A special thanks goes to my dad for always being ready to help no matter how little time he had, and for always saying my work was very good even when it was not ☺.

Thank you all!

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Física e Astronomia

MSc. Medical Physics

Bone scan lesions uptake quantification for therapy response in metastatic prostate cancer

by Laura PROVIDÊNCIA

The aim of this work is to develop an algorithm capable of automatically quantifying bone scintigraphy images from patients with metastatic prostate cancer. Currently, the assessment of bone scans relies solely on the interpretation of an expert physician who visually evaluates the scan. Besides this being a time consuming task, it is also subjective, as there are no absolute criteria either to identify bone metastases or to quantify them with a straightforward and universally accepted procedure. Here, an algorithm for the detection of hotspots, followed by the the attenuation of false-positive detections, is developed. This algorithm is mainly motivated by the call for a method for bone scintigraphy quantification whose development does not require a fully labelled database, which is very rare and most likely unavailable for most researchers. The detection algorithm was able to detect 100% of the metastases, though with the downside of presenting a high false-positive rate of 73%, corresponding to an average of 32 false-positive detections per image. The number of false-positive detections was reduced through two different methods: one using image analysis techniques and other using machine learning. The image analysis method was able to correctly classify 30% of the non-malign hotspots from the test set as false-positives, leading to a decrease in the the number of false-positive detections per image from 32 to 22. For the machine learning method, an iterative semi-supervised classification algorithm was specially created for the purpose of hotspot classification, only requiring knowledge about the type of bone scan the hotspots were extracted from, and not about the hotspots themselves. The best machine learning algorithm achieved a sensitivity, specificity, false-negative rate and false-positive detections per image of 63%, 58%,

37% and 14 respectively, outperforming state-of-the-art classification algorithms. The final assessment of the scintigraphic exam was accomplished by calculating the Bone Scan Index, a quantitative biomarker specially developed to improve the interpretation and clinical relevance of bone scans from patients with metastatic prostate cancer.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Física e Astronomia

Mestrado Física Médica

Quantificação de cintigrafias ósseas para avaliação da resposta a terapia em cancro da próstata metástico

por Laura PROVIDÊNCIA

O objetivo deste trabalho é desenvolver um algoritmo capaz de quantificar de uma forma automática cintigrafias ósseas de pacientes com cancro da próstata. Até à data, a avaliação deste tipo de exame está totalmente dependente da interpretação de um médico especialista que analisa visualmente a cintigrafia. Para além de ser um processo moroso, é também subjetivo, uma vez que não há um critério absoluto nem para identificar as metástases ósseas nem para as quantificar de uma forma simples e universalmente aceite. Aqui é desenvolvido um algoritmo para a detecção de *hotspots*, ao qual se segue a atenuação de detecções falsas-positivas. Este trabalho é particularmente motivado pela necessidade de um método para a quantificação de cintigrafias ósseas cuja elaboração não requeira uma base de dados "etiquetada", que é bastante rara e de difícil acesso à maioria dos investigadores. O algoritmo de detecção conseguiu detectar 100% das mestástases, com a contra-partida de apresentar uma elevada taxa de falsos-positivos de 73%, o que corresponde a uma média de 32 deteções falsas-positivas por imagem. O número de falsos-positivos foi reduzido através de dois métodos diferentes: um que usa técnicas de análise de imagem e outro que usa *machine learning*. O algoritmo de análise de imagem conseguiu correctamente identificar 30% dos *hotspots* não-malignos do conjunto de teste como falsos-positivos, reduzindo o número de detecções falsas-positivas por imagem de 32 para 22. Para o método de *machine learning*, um algoritmo iterativo semi-supervisionado foi especialmente desenvolvido para a classificação de *hotspots*, que necessitava apenas de informação sobre o tipo de cintigrafia óssea ao qual um *hotspot* pertencia (e não sobre os *hotspots* em si). Este algoritmo obteve uma sensibilidade, especificidade, taxa de falsos-negativos e deteções falsas-positivas por imagem de 63%, 58%, 37% e 14, respectivamente,

mostrando ser superior aos algoritmos de classificação de estado da arte usados para comparação. A avaliação final dos exames foi feita através do cálculo do *Bone Scan Index*, um biomarcador quantitativo especialmente desenvolvido para melhorar a interpretação e relevância clínica de cintigrafias ósseas de pacientes com cancro da próstata.

Contents

List of Figures

List of Tables

Glossary

AE	Autoencoder
ANN	Artificial Neural Network
API	Application Programming Interface
BLS	Bone Lesion Scoring
BS	Bone Scintigraphy
BSI	Bone Scan Index
BSLA	Bone Scan Lesion Area
BSLC	Bone Scan Lesion Count
BSLI	Bone Scan Lesion Intensity
CAD	Computer Aided Diagnosis
CNN	Convolutional Neural Network
CRPC	castration-Resistant Prostate Cancer
EOD	Extent Of Disease
FN	False Negatives
FP	False Positives
ISC	Identified Skeletal Collection
OCC	One-Class Classification
OC-SVM	One-Class Support Vector Machine
PAB	Positive Area on Bone scans
PCa	Prostate Cancer
PSA	Prostate-Specific Antigen

ReLu	Rectified Linear Unit
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
SVM	Support Vector Machine
TN	True Negatives
TP	True Positives
VAE	Variational Autoencoder

Chapter 1

Introduction

Machine learning is one of the major branches of artificial intelligence and has been successfully applied in a variety of medical domains. Combined with the image processing and analysis field, it can be used for the development of artificial systems capable of constructing explicit and meaningful descriptions of medical images, which is extremely useful for disease diagnosis, disease progression assessment, treatment planning and for the overall patient management. These algorithms not only reduce the dependency on manual analysis, but they can actually outperform humans in a variety of tasks when it comes to speed, objectivity and efficiency. With these developments in mind, machine learning algorithms combined with image analysis techniques are used to build a computer-aided diagnosis system for the quantification of hotspots in bone scans from patients with prostate cancer.

1.1 Contextualization

Prostate Cancer (PCa) is the second most commonly diagnosed cancer in men (World Health Organization, a). Patients with PCa often develop metastases, with more than 80% of these metastases appearing in the bones (Bubendorf et al., 2000, Gandaglia et al., 2014). Even though bone metastases are seldom the cause of death, they are the leading cause of morbidity and a major challenge in the management of patients with this disease, leading to a diminished quality of life. Furthermore, the presence of bone metastases is an indicator of progression of the disease and typically correlates with a poor prognosis (Soloway et al., 1988, Norgaard et al., 2010). Given the high occurrence of bone

metastases and the associated medical implications, a frequent imaging follow-up of PCa patients is commonly needed. The most common diagnostic procedure used for screening, diagnosis, assessment of treatment efficiency and monitoring of disease evolution is bone scintigraphy, due to its high sensitivity and widespread availability at relatively low cost. This imaging modality has, however, some limitations, such as its low specificity: because it evaluates the distribution of active bone formation in the skeleton and identifies the sites where metabolic reaction is occurring, it shows several suspicious uptake of non-metastatic origin such as micro-fractures, inflammation or physiological processes, which are not related to bone metastases. To date, the assessment of bone scans relies solely on the interpretation of an expert physician who visually assesses the scan. Besides this being a tedious and time consuming task, it is also quite subjective, as there are no criteria to differentiate bone metastases from benign bone lesions, nor to quantity them. This means that, to this day, the disease stage, as well as the response to treatment is immeasurable, rendering the process of determining whether or not the patient condition is regressing too subjective.

Given the high occurrence of metastatic PCa, there should be by now more practical and, most importantly, more objective criteria to evaluate a bone scintigraphy. The literature found on this topic shows that there has been some effort to develop a computer-aided diagnosis system capable of automatically detecting and quantifying bone metastases in bone scintigraphy images. Such system would give the physicians a fast, precise and reliable tool to quantify bone scans and evaluate disease progression and response to treatment. To this date, however, only one diagnosis system has been approved by Medical Device Directive, being available for clinical use only in a limited number of countries. Furthermore, it resorts to supervised machine learning algorithms which, despite the promising results, require access to a large database of labelled bone lesions, which is very difficult to find in the medical context. An automatic system for bone scan assessment and metastatic PCa diagnosis and follow-up accepted by and available to the medical community is thus very much needed.

1.2 Objectives and Contributions

This work aims to develop a system that uses image analysis techniques, as well as semi-supervised machine learning algorithms to quantify bone scintigraphy images, and thus

assist physicians during the diagnosis and follow-up of PCa patients. No literature of an algorithm for hotspot classification that does not require a labelled data set could be found to this date, which makes this work a pioneer on the topic. Such a system is very much needed in the medical community and will benefit not only the physicians, by massively reducing the time and effort needed to evaluate a bone scan, but also the patients themselves, as a much more objective and precise assessment of the disease stage and evolution can be given. The main contributions of this dissertation include:

- The development of an automatic algorithm for hotspots detection;
- The proposal of an algorithm based on image analysis techniques to reduce the number of false-positives detections;
- The development of a new, iterative, semi-supervised algorithm for attenuation of false-positive metastases, that does not require a fully labelled data set;
- Extensive experiments on a real data set of 198 scintigraphy images from 102 patients with prostate cancer, used to test out the here proposed algorithm.

The work developed during this dissertation has given rise to:

- An oral presentation in the 14^{th} edition of *Encontro Investigação Jovem da Universidade do Porto* with the title "Quantification of Whole-Body Bone Scans with Imaging Processing and Machine Learning Algorithms " (IJUP, 2021);
- An oral presentation in a Medical Physics workshop held as part of a collaboration between the University of Porto, the Ludwig Maximilian University of Munich and EUGLOH (European University Alliance for Global Health), wit the title "Bone scan lesions uptake quantification for therapy response in prostate cancer";
- An oral presentation for CI-IPOP scientific meeting with the title "An automatic algorithm for the assessment of bone lesions in bone scintigraphy images";
- An e-Poster at ENJIO ("Encontro Nacional de Joves Investigadores em Oncologia") with the title "Algoritmo para a quantificação automática de cintigrafias ósseas de pacientes com cancro da próstata" (ENJIO, 2021);

- A published article in MPDI's Journal of Imaging under the title "An iterative algorithm for semi-supervised classification of automatically detected hotspots on bone scintigraphy images" (Providência et al., 2021a);
- A paper for the REDCAP 2021 (Portuguese Conference on Pattern Recognition) with the title "False-positives attenuation of automatically detected hotspots on bone scintigraphy images using image analysis techniques" (Providência et al., 2021b).

1.3 Document structure

This thesis is organised as follows: Chapter 2 gives an overview of the concepts that will be necessary for the understanding of the work developed; Chapter 3 reviews the state of the art; Chapter 4 gives a detailed description of the methodology, including the materials and methods used, with proper justification for their choice; Chapter 5 presents and discusses the results and Chapter 6 finishes this dissertation with the main conclusions and possible directions for future work.

Chapter 2

Background knowledge

In this chapter, an overview of the concepts that are most relevant to this dissertation is made. It starts with an introduction to prostate cancer, bone scintigraphy, and biomarkers currently used for assessment and follow-up of patients with bone metastases. It then covers several machine learning techniques that can be useful for the current work, along with their connection to computer vision, where the topics of image classification and feature extraction will be addressed. Finally, a summary on the performance metrics that can be suitable to evaluate the algorithm that will be developed through this thesis is given.

2.1 Prostate Cancer

According to the World Health Organization, prostate cancer (PCa) is the second most commonly diagnosed cancer in man, accounting for more than 1.4 million new cases and more than 375000 deaths worldwide in 2020 (World Health Organization, b). A surge in the number of PCa diagnoses was observed in the early 1980s due to the introduction of the prostate-specific antigen (PSA) test, which was reported as a potential biomarker and a screening tool for PCa during the following years (Klein and Jones, 2013). Since then, the incidence rates of prostate cancer have largely increased, partially because of the increased availability of screening for PSA but also due to an increased population awareness and a longer life expectancy. Prostate cancer is caused by an abnormal growth of the cells in the prostate gland and, in most cases, is relatively slow-growing, which means it can take years for it to be detectable or to spread beyond the prostate. Due to its slow growth, a

man to whom a tumour was detected does not have to rush into treatment right away, as it could cause more complications than the disease itself, and an active surveillance or watchful waiting may instead be recommended. If, however, the benefits of a treatment outweigh the risks, several treatment options are available and the most appropriate one will be chosen having into account the patient health history and quality of life, the presence of other medical conditions, the growth rate of PSA levels and other symptoms, and the disease stage (including whether or not the tumour has spread to other parts of the body). Treatment options include surgery to remove the prostate, radiation therapy (e.g., external-beam radiation or brachytherapy), systemic treatments (e.g., chemotherapy or hormonal therapy), and focal therapies (e.g., cryosurgery and high-intensity focused ultrasound) (Cancer Net, 2020). Focal therapies aim to destroy directly the tumour inside the prostate while leaving the remaining gland intact; because they spare most of the healthy tissue, this kind of therapy is known for minimising the side effects that are generally associated with more aggressive treatments like radiation and systemic treatments, which use high-energy rays and medication to kill cancer cells, respectively. Nonetheless, they are not approved as a standard of care treatment in most national and international guidelines, and are still subject to clinical trials (Worthington). Surgery, radiation therapy, and systemic treatments are, therefore, the most common and consensually approved treatments for prostate cancer.

2.2 Bone Metastases

Patients with advanced prostate cancer often develop metastases, which are caused by primary tumour cells that escape from the prostate gland and spread through the lymphatic system or the bloodstream to other areas of the body. The most frequent site for metastatic growth of prostate cancer is the bone, and almost all patients with advanced prostate cancer show histological skeletal involvement (Msaouel et al., 2008). An autopsy study carried out by Bubendorf et al. (2000) revealed that 90.1% of the patients who had developed metastases had bone metastases. A later study by Gandaglia et al. (2014), with the aim to report the most common sites of metastases in PCa patients, reached a similar conclusion, with 84% of the patients included in the study having bone involvement. Even though the bone metastases are seldom the cause of death, they are the leading cause of morbidity and a major challenge in the management of patients with this disease, leading to a diminished quality of life. The presence of bone metastases, specially in

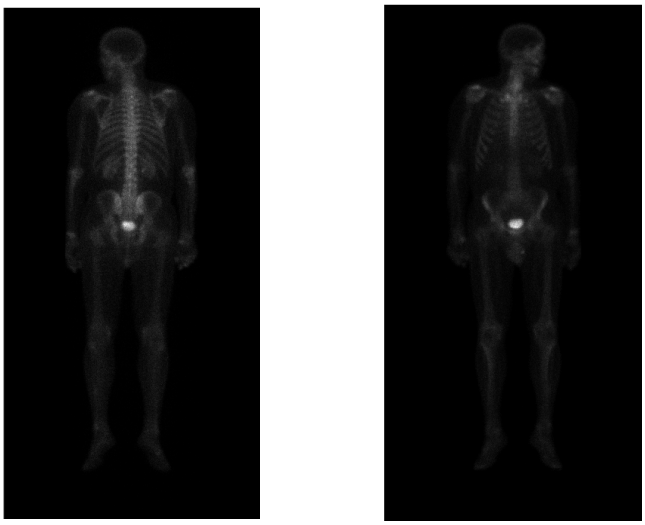
higher extents, is an indicator of progression of the disease and typically correlates with a poor prognosis (Norgaard et al., 2010, Soloway et al., 1988). Currently there is no cure for metastatic prostate cancer, but it can often still be treated to slow down its growth. A precise detection of bone metastases is essential to provide the doctors and physicians the accurate staging they require to choose the appropriate treatment for the patient in question, to monitor the evolution of the disease, and to evaluate the treatment efficiency.

2.3 Bone Scintigraphy

The most common diagnostic procedure used for screening, assessment of treatment and follow-up of patients with bone metastases is whole-body bone scintigraphy (BS) (Brenner et al., 2012), due to its relatively high sensitivity, ranging from 70% to 78% (Ohta et al., 2001, Even-Sapir et al., 2006a, O'Sullivan, 2015), and widespread availability at relatively low cost. Bone scintigraphy, also known as bone scan, is a nuclear medicine imaging technique used in screening for several skeleton related pathological conditions, including bone metastases. To understand how a BS works, it is necessary to introduce the concepts of bone remodelling and bone turnover.

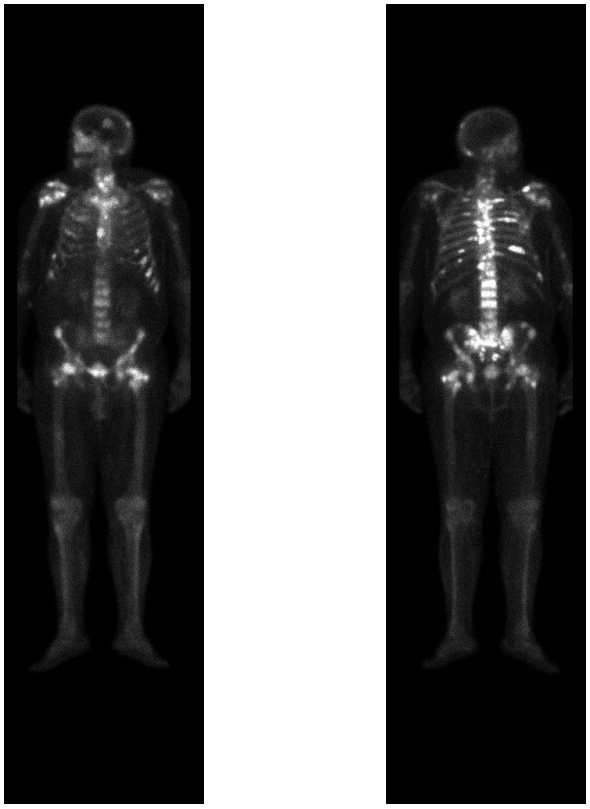
Bone is a dynamic tissue which is constantly being remodelled throughout a person's life. The bone remodelling is a continuous process that is characterised by the removal of bone tissue from the skeleton – a process called resorption – followed by the formation of new bone tissue, which comprises both bone matrix syntheses and mineralization. The processes of bone destruction and bone formation is carried out by two groups of cells called osteoclasts and osteoblasts, respectively. This constant cycle of bone synthesis and destruction is essential for adult bone homeostasis and maintenance of the shape of bone: it adjust bone architecture to meet changing mechanical needs, helps to repair micro-damages in bone matrix preventing the accumulation of old bone and maintains normal calcium levels in the body (Hadjidakis and Androulakis, 2007). Bone remodelling is the cellular mechanism behind the bone turnover, which is in turn defined as the total volume of bone that is both resorbed and formed over a period of time, usually expressed as percent per year (Chun-Yi, 2020). In a bone scintigraphy, a bone-seek radioisotope, that is, a substance that collects in the bones following the normal physiological processes, is injected intravenously into the patient. The most used radiotracer in BS in Technetium 99m-methyl diphosphonate (^{99m}Tc MPD) since it is absorbed into hydroxyapatite, a naturally

occurring form of calcium phosphate that constitutes 60% of the bone structure (Feng, 2009, Jeong et al., 2019). This radiotracer has the advantage of being highly absorbed into the skeleton and rapidly removed from the soft tissues once in the patients body. The radioactive isotope will flow through the body and the osteoclasts and osteoblasts will incorporate it directly into the bone, that is, it will have a tendency to accumulate in areas of active bone formation. A period of time after the injection, that generally goes from 2 to 4 hours (Mettler and Guiberteau, 2019), the patient is placed inside a device with a double headed digital whole-body gamma camera which will detect and locate the radiation emitted by the radiopharmaceutical. The reason behind the use of two detectors is so that a simultaneous image of the anterior (AP) and posterior (PA) views can be acquired (see Figure 2.1). Because the radioisotope has accumulated in the regions of bone remodelling, the final scan will reveal brighter areas, which indicate an increased rate of bone production (see Figure 2.2). These areas are referred to as hotspots, and may indicate not only the presence of bone metastases, but also other conditions such as trauma, micro-arthritis, benign degeneration, or bone infections (Purden, 2019). The biggest disadvantage in the use of bone scintigraphy to detect bone metastases is, therefore, its low specificity. Because this exam evaluates the distribution of active bone formation in the skeleton and identifies the sites where metabolic reactions are occurring, it detects several suspicious uptakes of non-metastatic origin, which lead to a high false-positive rate to detect bone metastases. Furthermore, the task of differentiating malign from benign bone uptakes in a bone scan is a very challenging and subjective task, even for the most experienced physicians.



AP PA

FIGURE 2.1: Illustrative picture of the AP and PA views, obtained during a bone scintigraphy



AP PA

FIGURE 2.2: Illustrative picture of a bone scintigraphy of a patient with bone metastases. The brighter areas indicate an increased rate of bone production and are called hotspots

Other imaging modalities besides bone scans for the detection of skeleton metastases have emerged over the last years. These include Single-Photon Emission Computed Tomography (SPECT), Positron Emission Tomography (PET), a PET scan combined with Computer Tomography (PET/CT), and Magnetic Resonance Imaging (MRI). The technique that showed the most promising results was 18F-NaF PET/CT. A study developed by Even-Sapir et al. (2006b) revealed that the bone scan had a sensitivity and specificity of 70% and 57%, respectively, whereas the 18F-NaF PET/CT achieved a result of 100% for both quantities. Another study by O'Sullivan (2015) obtained similar results, with the bone scan presenting a sensitivity and specificity of 78% and 48%, respectively, and the 18F-NaF PET/CT presenting 100% and 97% for the same quantities. Despite the low false negative rates compared to other imaging modalities, the bone scintigraphy is still the most used technique for diagnosing bone metastases, due to its low price, easier implementation and widespread availability.

After the bone scintigraphy is obtained, it is up to the physician to analyse the obtained image and assess the patient condition. Although this can be accomplished in a subjective

manner, the preferred choice would be to use an imaging biomarker.

2.4 Biomarkers

A biomarker is any any substance, structure or process that can be detected and measured in the body and that can be used as an indicator of normal or pathological processes (World Health Organization and International Programme on Chemical Safety, 2001). The most used biomarker for PCa assessment is prostate-specific antigen (PSA), a protein produced by both normal and malignant cells of the prostate gland which can often be found in high quantities in the blood of men with prostate cancer (National Cancer Institute, a). The Gleason grading is another biomarker used for clinical staging in patients with PCa. This score is obtained through an evaluation of samples from a prostate biopsy: cancer cells that resemble normal prostate tissue when viewed under a microscope are associated with less advanced tumours and have lower scores, while cancer cells that have mutated so much that barely resemble healthy cells are associated with more aggressive tumours and receive higher scores (National Cancer Institute, b). Although the staging of a patient with PCa cannot be determined by these measurements alone, these quantitative biomarkers play an important role as clinical and diagnostic tools in disease assessment.

When it comes to the evaluation of whole-body bone scintigraphy of PCa patients with bone metastases, however, there is yet no quantitative measurement to assess clinical or pathological staging that is approved by the medical community. At most, a bone scan allows for a qualitative description of the extent of the disease, performed by a physician who interprets and evaluates the patient's condition and its response to treatment. Besides being a very complex and time consuming process that requires the intervention of an experienced doctor as well as a previous analysis of the patient medical history, this evaluation is to a certain degree quite subjective, which leads to a divergence between health professionals. There is thus a need to develop a quantitative imaging biomarker for the assessment of whole-body scintigraphy in patients with PCa.

An imaging biomarker is a particular type of a biomarker which can be defined as any biological feature that is detectable in an image (in this case, a bone scan) and that works as an indicator of normal or pathological processes. It must also be quantitative, that is, it must rely on the extraction of quantifiable features from the bone scan that can objectively assess the severity and degree of change of a metastatic PCa patient's condition.

Ideally, this biomarker should be able to determine the staging of disease, monitor clinical response to a treatment and serve as biomarker for disease prognosis. Although there is no formally approved biomarker to perform a quantitative assessment of bone scans, there are some parameters, both qualitative and quantitative, that have been developed for this purpose. An overview of these parameters is presented next.

Extent Of Disease

The Extent Of Disease (EOD) is a semi-quantitative grading system used for a subjective evaluation of bone metastases in bone scintigraphy. It is calculated by assessing the number and size of lesions classified as malign and assigning a number on a scale according to the obtained value. A scan can be classified five EOD grades, from 0 to 4, as described in Table 2.1. In grades 1 and 2, to be counted as one metastasis, the lesions must be less than 50% of the size of a vertebral body, otherwise they may be counted as two (Soloway et al., 1988). Although EOD has proved to have some prognostic information (Soloway et al., 1988) as well as the ability to evaluate disease progression or regression (Mustansar, 2018), it is still based on a visual assessment of the whole-body images, thus lacking quantitative information about the patient's condition.

TABLE 2.1: Grades of the Extent Of Disease and how they are reflected in a bone scan.

Grade	Description
0	No evidence of bone metastases
1	$N < 6$
2	$6 \leq N \leq 20$
3	$N > 20$
4	superscan*

*A scan is called a *superscan* when more than 75% of ribs, vertebrae and pelvic bones have metastases

Positive Area on Bone Scan

The Positive Area on Bone Scan (PABS) is a quantitative measurement that can be used to quantify bone scintigraphy, which is calculated by the formula:

$$\text{PABS (\%)} = \frac{\text{Positive Area}}{\text{Square Area}} \cdot 100 \tag{2.1}$$

where the positive area is considered to be the sum of the areas of all metastases, both from anterior and posterior views, and the square area is defined as the area of a rectangle with width and height equal to the length of the gluteal region and the vertical height of the entire skeleton, respectively (Noguchi et al., 2003). This gives us an estimate of the fraction of involvement of bone metastases over the whole body.

Bone Scan Index

The Bone Scan Index (BSI), developed to improve the interpretation and clinical relevance of bone scans of patients with PCa (Dennis et al., 2012), is another quantitative biomarker used for estimating metastatic burden on bone scans, and one of the most used. This method has reported to be a reproducible biomarker for staging, disease progression and treatment response, as well as a reliable prognostic tool. The BSI represents the tumour burden expressed as the percentage of bone mass affected by tumour relative to the entire skeleton mass (Li et al., 2017), and can be calculated in three steps. First, the area of a hotspot classified as a metastatic lesion is divided by the area of the anatomical region where the lesion is found. This assumes that the skeleton has been previously segmented into several regions of interest whose areas are known, just as shown in Figure 2.3. Secondly, this ratio is multiplied by a coefficient reflecting the fractional mass of the present skeletal region with respect to the total skeleton mass; this gives an estimate of the volumetric fraction of the skeletal region occupied by the metastasis (Kaboteh et al., 2013). For a certain skeleton region, this coefficient is obtained by calculating the fraction between mass of the bone where the metastasis is found and the mass of the entire skeleton. The mass of the bones can be obtained, for instance, from the ICRP Publication No.23 (Snyder, 1981), where the mass of 158 bones is expressed as a fraction of the mass of the entire skeleton, derived from the reference man representing the average adult (Mustansar, 2018). Finally, this procedure is performed for all hotspots classified as possible metastases and the BSI is calculated by adding-up all bone tumour involvement values, i.e.:

$$\text{BSI}\ (\%) = \sum_{i=1}^{N} \frac{A_{HS_i}}{A_{R_i}} \cdot C_R \cdot 100 \tag{2.2}$$

where N is the number of hotspots classified as metastases, A_{HS_n} is the area of the n^{th} hotspot, A_{R_n} is the area of the skeleton region where the n^{th} is located and C_R is a coefficient that reflects the fractional mass of the skeletal region R_n with respect to the total skeleton mass.

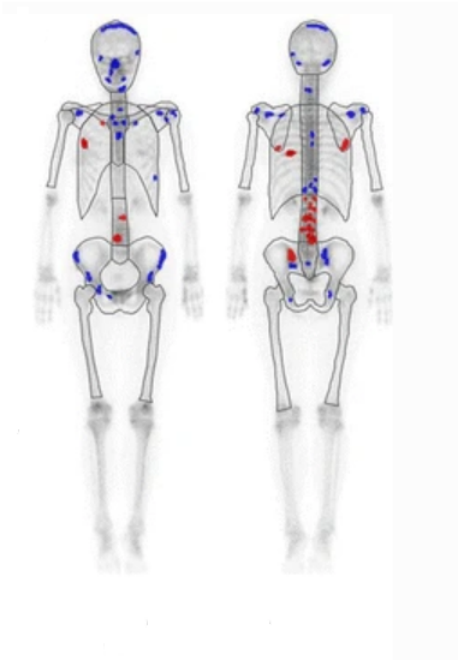


FIGURE 2.3: Schematic picture to illustrate how BSI is obtained. The red and blue regions represent metastatic and benign hotspots, respectively. The regions outlined by a black line represent the anatomical regions of interest (Ito et al., 2016).

2.5 Machine Learning

Machine learning is a branch of artificial intelligence which explores the construction of algorithms that make computers capable of learning how to perform specific tasks based on training data. The first developments in the area bring us back to the XX century, when Arthur Samuel, know as the man who coined the term "machine learning", developed a computer program for playing checkers in the 1950s (Samuel, 1959). A typical learning machine analyses a collection of examples of some phenomenon and, by learning the underlying structure in that data set, finds a mathematical algorithm that, when applied to another set of similar inputs, will predict the desired outputs (Burkov, 2019). The data set from which the computer will learn to predict these outputs is called the *training data*, since it it used to train the machine to perform a specific task.

Another particularity of machine learning algorithms is that the model built by the machine from the analysis of the training data will adapt and improve when higher amounts of data are available, that is, the machine improves its performance at some task through experience. According to Mitchell (1997), the type of tasks that a machine is capable of learning fall under the the class of problems that improve through experience:

> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

These three features (class of tasks, performance, and experience) can be identified in Arthur Samuel's first machine learning algorithm: the **performance** of the algorithm could be measured as its ability to win a game of checkers, which in this case would be defined as the **task**; it was also expected that the computer performance would improve with **experience**, that is, the more games the machine played against itself the higher the chances it would win a game (Mitchell, 1997).

Just like Arthur Samuel programmed a computer that could learn how to play checkers better than the person who wrote the program (that is, Arthur Samuel himself), several machine learning algorithms that have been implemented in our daily practice, and that already perform better than humans, have since then been developed. These include search engines, fraud detection, email intelligence, social networking, financial forecasting, computer vision, medical diagnosis and healthcare.

Machine learning algorithms are commonly dived into four categories: supervised, unsupervised, semi-supervised and reinforcement learning. Here, after a brief introduction to supervised learning, we will focus more on unsupervised and semi-supervised learning techniques.

2.5.1 Supervised Learning

In supervised learning, a data set of labelled examples is used to train the algorithm. This set consists of a vector of N data points $\{(\mathbf{x}_i, y_i)\}^N$, where each element $\mathbf{x}_i$ is known as a feature vector and y_i is the label that corresponds to the i^{th} feature vector. A feature vector is a set of m independent variables belonging to the $\mathbb{R}^m$ space that is used to numerically describe each object in the data set; combining the features vector from all data points, an

$m \times N$ feature space is obtained. Depending on the type of problem, the label y_i can either be an element belonging to a finite set of classes (classification problem), a real number (regression problem), or a more complex structure like vector or a matrix (Burkov, 2019). In a supervised learning problem we thus have a set of input vectors $\{\mathbf{x}_i\}^N$ belonging to the input space $\mathcal{X}$ which are mapped into a set of labels $\{y_i\}^N$, belonging to the output space $\mathcal{Y}$. There is an unknown target function $f\colon \mathcal{X} \to \mathcal{Y}$ such that $f(\mathbf{x}_i) = y_i$, that is, f is a function that fully maps the relationship between the features (or input variables) to the the labels (or output variables), and can thus predict with 100% accuracy on future data. The goal of a supervised learning algorithm is to use the available labelled data set to find a function g that approximates f, that is, to find a function that is able to predict the correct label for any given feature vector with the highest possible accuracy.

Supervised machine learning algorithms can be divided into two categories: regression and classification algorithms. Although they both work with labelled data sets, they differ in the type of machine learning problems they are used in, specifically in the type of variables they predict.

Regression Algorithms

Regression algorithms are used to predict a continuous variable. They use the available labelled examples to find a mapping function that translates the relationship between the independent (input) variables and the dependent (output) variables, which must be continuous. The obtained model is then used to predict a continuous quantity output of an unlabelled example. Some popular regression algorithms include Simple and Multiple Linear Regression, Polynomial Regression, Lasso Regression, Support Vector Regression, Decision Trees Regression or Random Forest Regression. These algorithms have a wide range of applications, such as weather and financial forecasting, house values prediction in real estate business, drug response modeling, social science research and even in behavioural analysis (Bhatia, 2017).

Classification Algorithms

Unlike regression algorithms, classification algorithms (also known as classifiers) are used to predict non-continuous variables. In this case, the labelled data points are analysed by the computer to find a function that can group the data into different classes, also referred to as categories or labels, based on the features of the phenomenon being studied.

The goal of the final classifier is to predict which of the predetermined classes the new input data will fall into. Depending on the number of outcomes, we can have a binary classification problem, where only two outcomes are possible, or a multi-class classification problem, where more than two outcomes are possible. This last one should not be confused with the multi-label classification problem, where multiple labels might be assigned to each sample. Some examples of problems where this type of classifiers are used are spam detection, speech and handwriting recognition, biometric identification, image and video recognition and risk of disease. The most famous classification algorithms are Support Vector Machines (SVM) and Kernel SVM, K-Nearest Neighbours, Logistic Regression, Random Forests, Decision Trees and Naïve Bayes.

In this thesis in particular we are dealing with a classification problem, as we aim to classify bone hotspots as malign or benign, although it does not fall into the supervised learning category.

2.5.2 Unsupervised Learning

In unsupervised learning, we work with a collection of unlabelled data points $\{\mathbf{x}_i\}^N$, with $\mathbf{x}_i$ being the feature vector of the i^{th} sample. As no labels are provided, the algorithm analyzes patterns of similar attributes across the data and tries do find some structure within. Unsupervised models can usually be broken down into two categories: clustering and association (Hodeghatta and Nayak, 2017). Here we will focus on the clustering methods as they are the most relevant for the current work: clustering has been proven to be an effective tool for discovering structure in unlabelled data sets, and therefore, in the particular case of this investigation, it could be particularly helpful to find patterns in the data that would allow to distinguish between healthy and unhealthy samples.

Clustering

Clustering is a technique used to group the objects in a data set into different clusters, such that objects that are similar to each other are grouped into the same cluster. This type of algorithm can uncover previously undetected patterns and relationships within a set of unlabelled data. Clustering can be further divided into exclusive clustering, fuzzy clustering, agglomerative clustering and density-based clustering, according to the method the algorithm uses to form the clusters. All these types are now briefly described.

In exclusive clustering, also known as hard clustering, a data point that belongs to a certain cluster can not be included in any other cluster, that is, it is exclusive to it. An example of this type of algorithm is the K-means clustering. This unsupervised learning algorithm aims to divide the data into K disjoint subsets (or clusters) $C_1, ..., C_K$ such that the final clusters are optimised, i.e., objects within a certain cluster are similar between them and different from the objects belonging to other clusters. In this method, the number of clusters K that the data set will be partitioned into is previously established; the algorithm will then place the clusters centres $\{\mu_k\}^K$, known as centroids, at arbitrary positions in the feature space*. Next, each instance is associated to one of the centroids according to the chosen criterion. The most widely used clustering criterion is the sum of the squared Euclidean distances between each data point $\mathbf{x}_i$ and each centroid m_k. In this method, the sum of the squared distances between each of the N data points and the K clusters' centroids is computed, and each data point is assigned to the closest cluster k according to the formula:

$$k = \arg\min_j \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \tag{2.3}$$

that is, each data point $\mathbf{x}_i$ is assigned to the closest cluster k, such that the square of the Euclidean distance between the point $\mathbf{x}_i$ in the feature space and the cluster centroid $\boldsymbol{\mu}_k$ is minimum. Before assigning all the points, the clusters' centroids are recalculated by taking the average of all the points in each cluster. The points are then reassigned to the closest cluster and the centroids are recalculated, and so on. The algorithm keeps repeating these two steps until it converges, that is, until the cost function reaches a minimum. The cost function associated to K-means clustering is the sum of the squared Euclidean distance between each point and the centroid of the cluster it belongs to, that is:

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N} \omega_{ij} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \tag{2.4}$$

where $\omega_{ij} = 1$ if the point $\mathbf{x}_i$ belongs to the cluster with centroid $\boldsymbol{\mu}_j$ and $\omega_{ij} = 0$ otherwise. Here, N is the number of points in the data set and K is the number of clusters. The

*It is important to refer that the algorithm is sensitive to the initial position of the centroids, as different initial locations will result in different outcomes. There are several methods that can be applied to choose the initial position of the centroids, e.g., through a manual or random selection, or using more sophisticated algorithms such as the popular *K-means++*, by Arthur and Vassilvitskii (2007). Several runs with different initialisation points should be performed to obtain optimal results.

function J will reach its minimum value when the intra-cluster distances are minimised and the inter-cluster distance are minimised.

In fuzzy (or soft) clustering, a data point can belong to more than one cluster. Instead of disjoint subsets, this algorithm uses fuzzy subsets, each of them being characterized by a membership function which assigns to each object a grade of membership ranging between 0 and 1 (Zadeh, 1965). The membership function quantifies the degree of membership of the elements to each cluster. Whereas in hard clustering a data point can either belong or not belong to a certain cluster, which in soft clustering would correspond to a data point having a membership value of 1 or 0, respectively, in this type of algorithm an instance can belong to more than one cluster. The most popular soft clustering algorithm is the fuzzy c-means (FCM). It follows a similar approach to the K-means clustering algorithm, except that in this case the parameter ω_{ij} from equation 2.4 now represents the membership of $\mathbf{x}_i$ to the cluster with centre $\boldsymbol{\mu}_j$, and can have any value between 0 to 1. The cost function for the FCM algorithms is therefore (Choudhry and Kapoor, 2016):

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N} \omega_{ij}^{m} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \tag{2.5}$$

where the parameter m, known as the fuzzifier or fuzzy weighting exponent (Huang et al., 2012), controls how much clusters may overlap (Klawonn and Höppner, 2003). Under the soft clustering category, one can also use probabilistic models for data clustering. The most commonly used probabilistic clustering method is the Gaussian Mixture Model (GMM). This model assumes that there are a certain number of Gaussian distributions within the data set, and each of these distributions represents a cluster. It is used to automatically find these normally distributed subsets within the original set, thus grouping the data points belonging to a single distribution together. If we have an $m \times N$ feature space, where N is the number of data points and m their dimension, the algorithm will create K clusters defined by a distribution with a $1 \times m$ mean vector $\boldsymbol{\mu}$, an $m \times m$ covariance matrix $\boldsymbol{\Sigma}$, and a mixture weight π_i, representing the probability that a randomly selected sample $\mathbf{x}_i$ belongs to the k^{th} mixture component (or cluster). As π_i is a probability, the following conditions must be met:

$$\begin{cases} 0 \leq \pi_{ik} \leq 1 \\ \sum_{k=1}^{K} \pi_{ik} = 1 \end{cases}$$

Gaussian Mixture Models use a iterative algorithm called Expectation-Maximisation (EM) to determine the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and π_k.

In agglomerative clustering, objects are grouped in clusters based on their similarity. The most common algorithm of this type is hierarchical clustering. In this method, each point in the data set begins as being one individual cluster, that is, we have N groups with one single element each. The algorithm goes then through an iterative process in which merges the two most similar groups until a single cluster containing all the data points is formed. The rules of clustering are based on the method chosen to measure similarity between points and groups of points, and are usually called "linkage methods". Different rules may be applied, such as measuring the longest distance between two points in each cluster (complete-linkage), measuring the shortest distance between two points in each cluster (single-linkage), measuring the average distance between all pairs of points from the different clusters (average-linkage) or finding the centroids of each cluster and measuring the distance between them (centroid-linkage) (Chauhan, 2019). These distances are commonly calculated with the Euclidean distance, but other metrics, such as the squared Euclidean, the Manhattan or the Mahalanobis distance, might also be used. The hierarchical clustering is represented by a dendrogram, a type of tree diagram that will show the hierarchical relationships between the objects in the data. A dendrogram stores all the information about the process of clusters arrangement from the very beginning.

Finally, we have density-based clustering. In this method, the algorithm analyses the feature space to find regions with a high density of data points, separated by regions with a low-density of points; the dense regions are defined as clusters. While in K-means clustering all points in a data set are assigned a cluster, even if they do not belong to any, in density-based algorithms data points in the separating regions of low point density (that is, that do not belong to any high density cluster) are usually labelled as noise. This can be an advantage when one is working with problems that fall under the anomaly detection domain or noise reduction. The algorithm takes two parameters: a distance ϵ and minimum number of points, $\mathrm{N_{min}}$. The parameter ϵ defines the maximum distance that a

point can be from another point for them to be considered neighbours; N_{min} defines the minimum number of points a region must have to be considered a high-density region. The algorithm starts by randomly selecting a data point, which can be either a core point or an outlier. If there are less than N_{min} points within a distance ϵ from the chosen point, then it is labelled as an outlier; otherwise, it is classified as a core point, and all the points that are within a distance ϵ, called the neighbours of the core point, are assigne to the same cluster. The process is repeated for all the new data added to the cluster; the borders of a cluster are defined by finding points that are neighbours of a core point but do not have a minimum of N_{min} points within a distance ϵ. The algorithm ends when all points are assigned to a cluster or labelled as outliers.

2.5.3 Semi-Supervised Learning

Unlike supervised learning, where the labels from the training set are known, and unsupervised learning, where the labels are unknown, the semi-supervised paradigm deals with situations in which a data set is composed of both labelled and unlabelled data. This is of great interest when one wants to build a classification algorithm but has only access to a database with a small portion of labelled samples and a large number of unlabelled samples. One can also use an unlabelled data set, which is much more common and easy to find, to improve supervised learning tasks when labelled data is scarce or expensive (Zhu and Goldberg, 2009). In these cases, the label of a sample in the training set is either completely known or unknown. However, other situations where there is incomplete knowledge about the training set can occur, for example, when one only possess partial information about the label (Domingues and Cardoso, 2014a).

We will now focus on a type of semi-supervised technique that will be of special interest for the current work: one-class learning.

2.5.3.1 One-Class Learning

Sometimes it may happen that we do not have access to a fully labelled data set, but we do have a set of samples that we know to belong to a certain class. By using a training set containing only examples from that class, known as the target class, we can build an algorithm that learns how to identify data belonging to that specific class. This type of learning problems are known as One-Class Classification (OCC) or One-Class Learning.

They differ from the conventional classification problems in the sense that the aim of an OCC algorithm is not to classify objects into one of several predefined categories (even if they do not belong to any), but rather to be able to decide whether an object belongs to a particular class or not. One-class learning algorithms are famous for their application in anomaly detection problems, which seek to identify examples that do not fit the characteristics of the *inlier* observations, that is, that are inconsistent with the remainder of a data set (Johnson and Wichern, 2007). Although the terms anomaly, outlier and novelty detection are often used interchangeably, there is an underlying difference between them. Outlier detection is usually referred to as an unsupervised method where both (unlabelled) normal and abnormal samples are present in the data set, and the algorithm tries to find observations that are deviant or inconsistent with the rest of the data, which are classified as outliers. On the other hand, the term novelty detection is used when one has access to a training data containing only normal data and is interested in determining whether a new observation fits within the current data set or is an outlier. Novelty detection algorithms fall under the category of semi-supervised learning. A detailed description about the underlying differences between each term and the context in which they are applied can be found in Carreño et al. (2020).

OCC is particularly useful when a representative set of labelled examples is either very difficult to obtain or not available at all. This is a problem that is very common, for example, in the medical context, where large labelled databases are very scarce. By using, for instance, a training set composed solely of data from healthy patients (which is easier to obtain), an OCC algorithm might be able to detect anomalies such as nodules and malign tumours in medical images such as mammograms, CT, PETs or MRI images (Wei et al., 2018), and thus being useful in medical diagnosis. According to Mazhelis (2006), three types one-class classifiers can be distinguished: (i) density methods, such as one-class Gaussian, mixture of Gaussian and one-class K-nearest neighbours, (ii) reconstruction methods, like one-class K-means and autoencoders, and (iii) boundary methods, as it is the case for one-class support vector machines (OC-SVM). A more detailed description about boundary OCC methods, in particular of OC-SVM, is now given, as it will be later referred to in the Chapter 4.

Khan and Madden (2014) approached the problem of OCC by developing a method called Support Vector Data Description (SVDD). This is a popular boundary based one-class

classification algorithm in which a hyper-sphere is constructed around the target data, so that almost all points are enclosed by that boundary with the minimum radius possible. A new point will be classified has an outlier if it falls outside the boundary defined by the hyper-sphere. Although this algorithm was later improved by the introduction of kernel functions, it still shows some limitations when the data set has higher dimensions or when there are large variations in density within the target objects (Khan and Madden, 2014).

An alternative approach was proposed by Schölkopf et al. (2000), who developed an OC-SVM algorithm for novelty detection. They begin by considering a training set of N data points $\mathbf{S} = \{\mathbf{x}_i \mid i \in \mathbb{N} \wedge 1 \leq i \leq \mathrm{N}\}$ belonging to the input space $\mathcal{X}$. They then define a feature map ϕ that maps $\mathcal{X}$ into a feature space $\mathcal{F}$. The goal with this transformation is to project the original data points into a higher dimension space where it is possible to define a hyper-plane that separates points from different classes. In the particular case of OCC, we want to find a hyper-plane that can separate normal points from outliers. Remapping will allow data that was not linearly separable in its original space to become separable by a hyper-plane in a higher dimensional space, which can be very useful for a classification algorithm. A drawback to this method is that, because we usually work with a high number of features, the computational costs of applying those transformations can be extremely high and impractical. Furthermore, the algorithm would then have to work with the higher dimensional vectors in the transformed features space. Fortunately, certain machine learning algorithms do not need to have access to the full transformed feature vectors in the higher dimensional space; instead, they only require certain measurement of these vectors, which is defined by the inner product. The way this is done in practice is by applying the so called kernel trick. A kernel is a function that takes the vectors from the original space as inputs, and returns the inner product of the vectors in the transformed space, that is, for two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ from the original features space, the kernel function is given by:

$$k(\mathbf{x}_1, \mathbf{x}_2) = \langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle \tag{2.6}$$

This means that we are able to work in the original features space without the need to compute the coordinates of the data in the higher dimensional space, which turns out to

be a much more efficient and less expensive option. (Zhang, 2018).

Although other kernel functions such as the polynomial or sigmoidal can be used, the method of Schölkopf et al. (2000) has proven to perform best when the Gaussian kernel is used (Khan and Madden, 2014), where the value of the inner product of two vectors from the original space in the transformed higher dimensional space can be obtain by evaluating the expression:

$$k(\mathbf{x}_1, \mathbf{x}_2) = e^{\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2}} \tag{2.7}$$

In Schölkopf et al. method, a hyper-plane is constructed which that separates the region in the $\mathcal{F}$ space that contains the target data points (in this case, the normal data) from the region that contains no data, and whose distance to the origin is maximised. The developed algorithm returns a binary function f that returns $+1$ in the region capturing the training data points and -1 elsewhere. Given a new data point $\mathbf{x}_i$, if it is lying on the side of the hyper-plane that is opposite to the origin than it is classified as normal, otherwise it is classified as an outlier.

2.5.3.2 Deep One-Class Classification

To overcome some drawbacks of the above-mentioned methods, such as the need to perform explicit feature engineering, the poor computational scaling associated with kernel methods and the expensive computational cost they required to store support vectors, deep learning approaches have been proposed for anomaly detection problems. They can be categorized into *mixed* approaches, if a deep representation of the data is previously learned* and later fed into a shallow anomaly detection algorithm such as OC-SVM, or *fully deep* approaches, where the algorithm can automatically construct a representation of the data that will be useful for the task of detection anomalies (Ruff et al., 2018).

A common approach for deep anomaly detection are autoencoders, where a multi-layer symmetric network is used to learn an intermediate representation of reduced dimension of the input data, known as latent space representation, which is then used to reconstruct the output data. Because the main goal of an autoencoder is to minimise the construction error, that is, the error between the input and the output, if during the training phase only

*How to obtain such deep representation of the data will be later explained, in section 2.6.2.2.

normal samples are fed to the decoder, it will learn a latent representation of the normal input data an therefore it will also learn to reconstruct them accurately. Given a data set with normal samples as well as outliers, the trained autoencoder should be able to reconstruct normal samples accurately while having difficulty to reconstruct anomalous examples. The outliers can therefore be detected by checking which samples present a high reconstruction error. Several applications of this approach can be found in the literature, such as in Japkowicz et al. (1999), Sakurada and Yairi (2014) and Chaurasia et al. (2020).

Another interesting approach for deep anomaly detection is proposed by Schlegl et al. (2017), where a deep convolutional Generative Adversarial Network (GAN) is used to identify anomalies in an unsupervised manner. A GAN is composed by two models: a generator G, that learns to generate plausible data by capturing real data distribution, and a discriminator D, that distinguish the generator's fake data from real data by estimating the probability that a sample came from the training data rather from G (Goodfellow et al., 2014). While the goal of the generator is to produce fake data that is as close as possible to the real data, the goal of the discriminator is to learn to correctly label all the fake data as *fake* and all the real data as *real*. If the discriminator classifies as fake the data produced by the generator, this is seen as a penalisation for the generator and an error signal is fed to the generator so that it can update its parameters to perform better. In the same way, if the data produced by G is classified as real by the D, or if real data classified as fake by the D, an error signal is also generated and is fed to the discriminator to improve its performance. The generator is therefore trained to fool the discriminator, that is, to maximise the probability that it makes a classification mistake, while the discriminator is trained to progressively become better at distinguishing real and generated images. The term *adversarial* comes from the fact that the two models are being trained and optimised to perform better at the same time, as if they were competing against each other. Given a 1D vector $\mathbf{z}$ of uniformly distributed input noise sampled from latent space $\mathcal{Z}$, the generator will map $\mathbf{z}$ into the space $\mathcal{X}$, populated by 2D images patches $\mathbf{x}$ of normal/healthy samples. After adversarial training, the generator G will have learnt a distribution p_g of the normal data and will therefore be able to map a vector $\mathbf{z}$ from the latent space to a realistic representation of a normal/healthy image $\mathbf{x}$, that is:

$$G: \begin{cases} \mathcal{Z} \to \mathcal{X} \\ \mathbf{z} \mapsto \mathbf{x} \end{cases}$$

Given a test sample $\mathbf{x_t}$, the model will try to find the $\mathbf{z}$ point in the generator's latent space that generates an image $G(\mathbf{z})$ that is as similar as possible to $\mathbf{x_t}$ If $\mathbf{x_t}$ follows the distribution p_g learned during the training phase (that is, the distribution of the real data which was captured by G), then the similarity between $G(\mathbf{z})$ and $\mathbf{x_t}$ will be high; on the other hand, if $\mathbf{x_t}$ does not follow the distribution of the normal data, it won't have a good representation $\mathbf{z}$ in the latent space: the similarity to $G(\mathbf{z})$ will be low and it will be classified as an anomaly. In order to find the point $\mathbf{z}$ which will have the image $G(\mathbf{z})$ with the highest degree of similarity to $\mathbf{x_t}$, a gradient descendent in the latent space $\mathcal{Z}$ is performed to optimise the location of $\mathbf{z}$. Finally a cost function comprised of a residual and a discrimination loss components is defined, which will output an anomaly score.

Another method for deep one-class classification was introduced by Ruff et al. (2018), called Deep Support Vector Data Description (Deep SVDD). Just like Khan and Madden (2014) approach, this method finds a data-enclosing hypershpere with the minimum radius possible. Contrary to Khan and Madden (2014) technique, where the sphere is constructed around features previously extracted from normal data, Ruff et al. (2018) use a deep neural network to learn a useful feature representation of the data so that this can be mapped into a hypersphere of minimum volume. This way, the model is able to learn weights from the network layers W while minimising the volume of a data-enclosing hypersphere R in the output space. A stochastic gradient descent algorithm is used to find the optimal W and R values in an alternating way, that is, during a certain number of epochs the network parameters W are trained while R is kept fixed, after which the minimum R value is found through a search line approach, and the process is repeated.

2.6 Computer Vision

Humans, just like many other animals, are lucky enough to be born with the sense of sight. From the moment we are born, we use our visual perception to explore the world, interact with our surroundings and learn from it. We are therefore constantly training and adapting our visual system so that we can use it to perform a vast amount of tasks, in a

way that for us is natural and automatic. But what would vision be for a computer? This is exactly what the field of computer vision (CV) intends to explore. According to Sebe (2005), "the goal of computer vision research is to provide computers with human-like perception capabilities so that they can sense the environment, understand the sensed data, take appropriate actions, and learn from this experience in order to enhance future performance". Developments in this field, which emerged during the 1980's, were only possible due to advances in the areas of image processing and analysis, that had took place a couple of decades before (Sebe, 2005).

Digital image processing refers to the process of capturing and translating visual signal into a digital image (Fernandes et al., 2020). It focuses on the development of a computer system capable of processing a digital image and transforming it in order to make it interpretable and manipulable. It encompasses a variety of techniques used for image enhancement and also to prepare images for future analyse, such as noise reduction, image equalization, image filtering and the application of affine transformations.

On the other hand, image analysis refers to the process of extracting meaningful information from an image, such as colour and brightness histograms, analysing and blocking regions based on intensity, mean and variance, or computing integral images. These can later be used to perform tasks of statistical pattern recognition, as well as inputs for image processing techniques such as image sharpening, thresholding or edge detection.

Finally, there is computer vision, which can be thought of as a field that combines together image analysis and processing with artificial intelligence. It aims for the development of artificial systems capable of constructing explicit and meaningful descriptions of physical objects from images (Ballard and Brown, 1982), which in turn will allow to interpret and understand the visual world. Complex algorithms are used to analyse digital images that allow computers to handle several visual problems of interest in fields such as the healthcare industry (Gao et al., 2018), automobile industry (Bala and Loce, 2017) or agriculture (Patrício and Rieder, 2018). These algorithms not only reduce the dependency on manual analysis, but they can actually outperform the humans in a variety of tasks when it comes to speed and competence. In the most recent years, computer vision has been shifting towards artificial neural networks and deep learning approaches as the choice for learning method during the development of visual recognition systems (Bohr and Memarzadeh, 2020).

The five main computer vision techniques are image classification, object detection, object tracking, semantic segmentation and instance segmentation (Le, 2020). In this thesis, we are particularly interested in the use of computer vision techniques for image classification.

2.6.1 Image classification

The problem of image classification in computer vision is concerned with the development of an algorithm that is capable of predicting the category a certain image belongs to, even though that image is unfamiliar to the computer. This is achieved by providing the computer with many examples of images belonging to different classes of interest, and then developing an algorithm that processes the data and learns how to distinguish objects from the different classes. This is clearly a description of the basis of a supervised machine learning algorithm: we provide the computer with a labelled training data, which we use to train a classifier capable of differentiate objects from different categories; we then ask the classifier to predict the labels of a new set of images that it had never seen before, and evaluate the algorithm by comparing the predicted classes with the true labels. But how exactly does a computer represent the visual information that is given to it? This is done through the extraction of measurable information from the images, which is presented in the form of a feature vector.

2.6.2 Feature Extraction

A feature vector is an $m-$dimensional vector composed by a set of numeric or symbolic characteristics, called features, that represent an object in a mathematical an easily analysable way (Reddy and Chatterjee, 2019). In image classification, one is thus focused on discovering informative and discriminating features that can be extracted from raw data and later be processed and analysed by a learning algorithm. Feature extraction can be accomplished manually or automatically.

2.6.2.1 Manual Feature Extraction

Manually extracting features from data requires, as the name implies, that we manually transform the raw data into a form that can be processed by the training algorithm. The way this transformation is done is through the use of feature descriptors. A feature descriptor is an algorithm which takes an image and outputs a vector of numerical features

(i.e., a feature vector) that encodes the information of interest. One can talk about two types of features descriptors, according to the features that are being extracted: global descriptors and local descriptors. Global features describe the visual content of the entire image by a single vector and include contour representation, shape descriptors, color and texture features. They describe the image as a whole and are usually used for object detection and classification. Local features, on the other hand, describe image patches (interest points or key points in the image) by a set of vectors, and are usually used for object recognition and identification. Some classical examples of local descriptors are the scale-invariant feature transform (SIFT), the speeded up robust feature (SURF) and the local binary pattern (LBP) (Kabbai et al., 2019).

2.6.2.2 Automated Feature Extraction

Instead of manually extracting the features, we can use a specialised algorithms to automatically extract useful features from an image without the need for human intervention. This procedure is known as as automated feature extraction, and the obtained features are referred to as *learned features*. Convolutional Neural Networks (CNNs) are a type of deep neural networks that can be used to extract learned features. Deep convolutional features are, however, obtained by training an algorithm to perform a certain task by providing it with a labelled image data set. Nevertheless, there are some techniques that can be used to extract these type of features when one has only access to an unsupervised data set. This work will focus on pre-trained models.

Pre-trained Models

Convolutional Neural Networks are specially popular in the computer vision community, having shown to excel in a wide range of image recognition tasks. During the ImageNet Large Scale Visual Recognition Challenge*, several milestone CNNs, which have since then become standards in image classification tasks, have been presented. These include the famous AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2014), VGG (Simonyan and Zisserman, 2015) and ResNet (He et al., 2015). A typical CNN is composed of two parts: a convolutional base composed by a series of convolutional an pooling layers,

*The ImageNet Large Scale Visual Recognition Challenge (ImageNet) was an annual competition held between 2010 and 2017, created to promote the development of better computer vision techniques and to benchmark the state of the art algorithms in computer vision. The participants had access to approximately one million labelled images, belonging to 1000 object classes, and the developed algorithms had to successfully perform tasks such as object detection, object localisation and image classification.

whose goal is to learn and extract high-level features from the input images, and a classifier, commonly composed of fully connect layers followed by a soft-max function which outputs a value between 0 and 1 (Figure 2.4). An important aspect of these deep learning networks is that they learn a hierarchical representations of the input data, progressing from low-level features to high-level features. While the initial layers detect very basic patterns like edges, gradients and blobs, the more advanced layers detect more complex patterns like objects and larger shapes in the images, which are more specific and therefore more useful for the classification task at hand (Siddiqui et al., 2017). This is known as an hierarchical features representation because this kind of vision systems extract features in a feedfoward manner: lower layers extract the low-level features, which in turn are used to build higher level features. Because lower level features are more general, one can extract these features, obtained by a pre-trained model, and use them in different tasks. This technique is known as transfer learning, because we are using a model that was trained in a certain database to perform a specific task, and repurpose it for a different task. This is achieved by (i) removing the original classifier and fine-tuning the model by either using the architecture of the pre-trained model to train the data set (if one as access to a large data set and high computational power), (ii) by training some layers and leaving others frozen*, or (iii) by completely freezing the convolutional base, that is, keeping the convolutional architecture in its original form, and using its outputs to feed the classifier (if the data set is small). The first two situations are only possible if one as access to a labelled data set because, even though we are keeping some of the structure and weights of the pre-trained model, we still have to update the CNN by re-training it, in order to adapt it to the data set. That last case, however, can be used as a mere feature extraction mechanism: we feed the network with the input images, allow them to propagate forward, stop at the desired layer and take the outputs of that layer as features, remembering that the more advanced the layer is the more specific the features are. This means that we can use a pre-trained model as an arbitrary feature extractor, and use the obtained features as inputs of a unsupervised machine learning algorithm. Pre-trained models can therefore be extremely useful even when a labelled data set is not available.

*Freezing a layer means that the weights of that layer are not modified, that is, the weights obtained by the pre-trained model are used, instead of estimating new weights based on the new data.

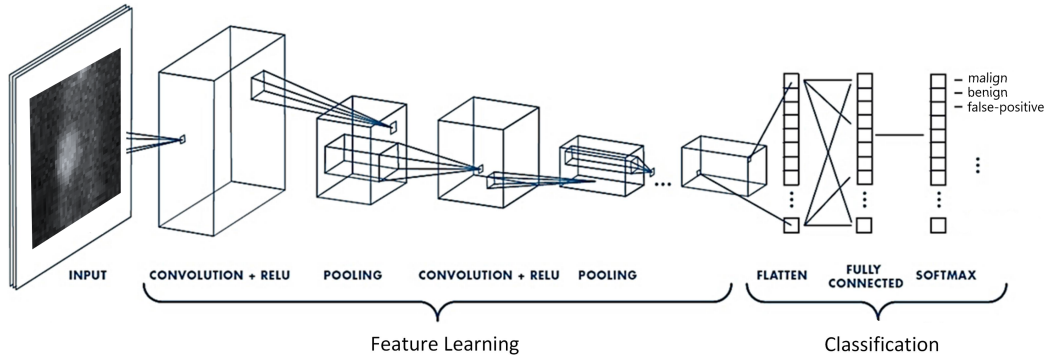


FIGURE 2.4: Schematic diagram of a Convolutional Neural Network. A typically CNN is composed by a feature learner and by a classifier. It is possible to remove the classifier and use the convolutional base as a feature extractor (adapted from Tabian et al. (2019)).

2.7 Evaluation

After the model has been trained, the only remaining step is the evaluation of its performance. This is achieved through the selection of suitable evaluation metrics (also known as performance measures), which are used to measure the quality of the model. Evaluating a model is an essential step when building a machine learning algorithm, as the feedback we get from the chosen metrics will let us know what improvements must be made and will also allow us to compare the model being developed to other existing ones.

Different performance metrics are used to evaluate different machine learning algorithms. When working with classification problems, more specifically with binary classification problems, the correctness of a classification can be evaluated by computing these four parameters that allow to compare the label predicted by the classifier with the actual class a data point belongs to :

1. **True Positives (TP)**: observations correctly predicted as belonging to a class
2. **True Negatives (TN)**: observations correctly predicted as not belonging to a class
3. **False Positives (FP)**: observations incorrectly assigned to a class they do not actually belong to
4. **False Negatives (FN)**: observation not assigned to a class they actually belong to

These four outcomes constitute the confusion matrix, a $N \times N$ matrix with N being the number of classes being predicted, the rows being the labels predicted by the classifier

and the columns the actual class an object belongs to. In Figure 2.5 a confusion matrix is shown for the case of the binary classification.

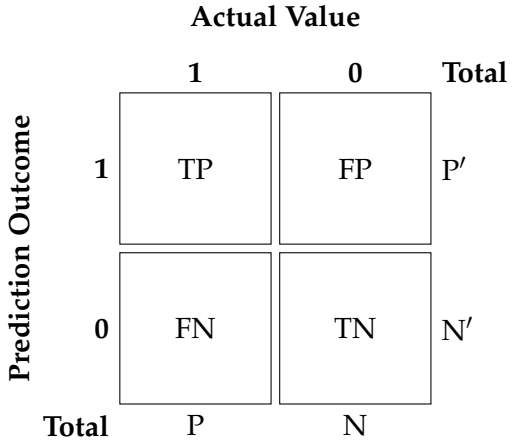


FIGURE 2.5: Confusing matrix for a binary classification problem with a positive ($y = 1$) and negative ($y = 0$) class. In this figure, P′ and N′ represent the number of objects assigned to the positive and the negative classes, respectively, while P and N represent the actual number of positives and negatives objects with the test set.

From the confusion matrix, several evaluation metrics can be computed.

Accuracy

Accuracy is the ratio of all correct predictions over the total number of predictions made by the model. It is given by:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.8}$$

This metric can be seen as the overall effectiveness of the classifier.

It is worth noting that, although the accuracy is a commonly used evaluation metric, it can lead to misleading conclusions, specially in the cases of imbalanced data (Chawla et al., 2004). An imbalanced data set is very often found in the real world: we can think for instance of the problems of finding defrauded accounts among a large number of normal accounts, or detecting a rare disease in a medical exam given a large amount of healthy images (Provost and Fawcett, 2001). The following metrics are more suitable for imbalanced classification, as they focus on one class.

Sensitivity

The sensitivity, also known as recall or True Positive Rate (TPR), is the proportion of positives correctly classified with respect to all examples that are actual positives, and is given by:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.9}$$

which will return a value between 0 and 1. This metric can be seen as a measurement of how effective the classifier is at identifying positive labels.

Specificity

In contrast with the sensitivity, the specificity, or True Negative Rate (TNR) is the proportion of negatives correctly classified with respect to all examples that are actual negatives, that is:

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.10}$$

Precision

The precision, also known as Positive Predict Value (PPV), is the proportion of positives correctly classified with respect to all instances classified as positives (correctly or incorrectly), and is given by:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.11}$$

It can be understood as a measurement of how plausible it is that an instance is positive when the model has classified it as such.

False Positive Rate

The false positive rate measures how many negatives the classifier incorrectly predicted as positives, and is obtained by calculating the ratio between the number of false-positives and all negative predictions (TN + FP):

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{2.12}$$

False Negative Rate

The false positive rate measures how many positives the classifier incorrectly predicted as negatives, and is obtained by calculating the ratio between the number of false-negatives and all positive predictions (TP + FN):

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{2.13}$$

F1-score

We can combine precision and recall into a single score called the F1-score, which is given by the harmonic mean of the two metrics:

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2.14}$$

The optimal classifier will reach a F1-score close to 1, which means it rarely classifies a positive instance as negative and vice-versa.

ROC Curve

It is importance to notice that the output of most predictive models will not be 0 or 1, but rather a number between those two values. This means that we have to choose a threshold value above which an output will be classified as 1 and below which an output will be classified as 0. The common choice for this threshold is 0.5, but there might be certain situations where one is interested in choosing a different value. For example, if we are developing an algorithm to classify a hotspot as malign or benign for bone cancer detection, we might want to choose a threshold lower than 0.5, arguing that is better to misclassify a hotspot as positive when in truth it is negative than letting a patient go with an undiagnosed cancer.

To analyse a binary classifier ability to discriminate classes, one can use the Receiver Operating Characteristic (ROC). The ROC curve is a visualization technique which summarizes the trade-off between sensitivity (or TPR) and False Positive Rate (FPR) for a predictive model using different threshold values. The false positive rate is actually the

complement of specificity (or TNR), that is:

$$\text{FPR} = 1 - \text{TNR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2.15}$$

To plot this curve, one must calculate the TPR and FPR of a model for different probability thresholds, and plot the the values in the ROC space, where the TPR is represented in the Y axis and the FPR in the X axis (Provost and Fawcett, 2001). A typical ROC curve is illustrated in Figure 2.6. A good classifier in one that presents an high TPR, while keeping the FPR low. Because it is difficult to compare ROC curves of different models, it is common to compute a measurement given by the area under the obtained curve, called Area Under the ROC Curve (AUC). It represents the probability that a randomly chosen positive instance will receive an higher score than a randomly chosen negative instance. The perfect model will have an AUC of 1, but values that fall within the range of $0.8 - 1$ are usually associated with a good classifier with good discrimination power. A high AUC means that, given a randomly chosen positive instance, the probability that the model will classify it as positive is much higher than the probability that the model will classify it as negative. A random classifier will have an AUC of 0.5, which means that, given a randomly chosen positive instance, the probability that the model will classify it as positive or negative is the same; this is an example of a poor classifier with no discriminative power.

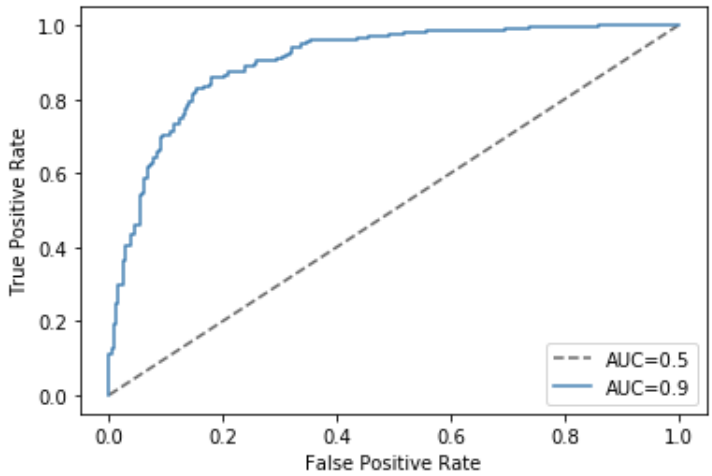


FIGURE 2.6: Graphic representation of two different ROC curves, and their respective AUC values. The blue curve is an example of a good classifier: we can get a high TPR while keeping the FPR low; the grey curve is an example of a random classifier: the probability that the model will label a positive instance as positive is the same as the probability that the model will label a positive instance as negative.

2.7.1 Performance metrics for multi-class algorithms

The above classification metrics are defined for binary classification problems, where a positive and a negative class can be defined. When extending to a multi-class problem, this division into a positive/negative class is no longer possible, and one needs to find a different strategy to evaluate the algorithm. A common approach is to follow a One-vs-Rest (OVS) technique, in which a multi-class problem with N classes $c_i \in C = \{1, ..., N\}$ is converted into N binary tasks such that the i^{th} task considers c_i as the positive class and the remaining N-1 classes c_j, with $i \neq j$, as the negative classes. Using this division to find the TP, TN, FP and FN for each class, the previous performance metrics defined for a binary problem can be applied to each task individually. This will result in N different values per metric, one for each binary task the initial problem was divided into (e.g., N recall scores, N precision scores, N F1 scores, etc.). To compare classifiers, however, one cannot have N values for each performance measure, and therefore it is necessary to find a wa to combine each of the N values obtained for each metric into just one value that characterises the classifier. There are two common averaging techniques that can be used for this end:

- Macro-average: the macro-average is obtained by calculating each metric independently for each class, and then taking the average. For example, in a 3-class problem with classes a, b and c, the recall would be calculated independently for each class using a OVR technique, and the macro-recall would be given by:

$$\text{macro-recall} = \frac{r_a + r_b + r_c}{3} \tag{2.16}$$

  where r_a, r_b and r_c are the recall values for classes a, b, and c, respectively.

- Micro-average: the micro-average uses the global number of TP, TN, FP and FN to calculate each metric. The formulas are therefore the same as the ones used for a binary classifier, but with $\text{TP} = \sum_{i=1}^{N} \text{TP}_i$, $\text{TN} = \sum_{i=1}^{N} \text{TN}_i$, $\text{FP} = \sum_{i=1}^{N} \text{FP}_i$ and $\text{FN} = \sum_{i=1}^{N} \text{FN}_i$, with N the number of classes. For the previous example, the micro-recall would be given by:

$$\text{micro-recall} = \frac{\text{TP}_a + \text{TP}_b + \text{TP}_c}{\text{TP}_a + \text{TP}_b + \text{TP}_c + \text{FP}_a + \text{FP}_b + \text{FP}_c}$$

Which of the averages should be chosen is hand to hand to the specific problem one is dealing with. The micro-average focuses on the larger class, and therefore the performance of that class will have more impact on the final result than the performance from a small class. Usually, it is easier for a classifier to identify the majority class and harder to identify the minority class, so that means that with micro-averaging one can have a high score even if the classifier is failing to identify the minority class (the classifier is still doing a lot of correct predictions in the majority class). However, the minority class is often more important than the majority class, and therefore the macro-averaging is more appropriate, as it treats the classes equally. For example, considering a classification problem aiming to diagnose cancer in a population where the prevalence of cancer is of 10% . In this case, "healthy" is the majority class and "cancer" in the minority class. If the final algorithm predicted that every patient was cancer-free, it would still get a micro-averaged precision of 90%, even though it was failing to identify patients with cancer. Despite the great precision score, using such an algorithm in the clinical practice would be very dangerous as it would fail to detect a life-threatening disease. When the macro-average is used instead, the precision score would fall to 50%, which is a more appropriate value for that classifier.

2.7.2 One-class Classification

With regard to one-class classification, one has access only to a positively labelled set, used for training, and an unlabelled set containing positive samples and outliers, meaning that practically none of the previously described metrics can be used to evaluate the algorithm's performance [*]. An alternative performance criteria for comparing OCC models was proposed by Lee and Liu (2003), by noting that there is an evaluation metric (recall) that can be estimated by applying the trained model on the examples from the validation set (which are, by definition, all positive). Because a validation set is made from the training data, all samples are positives; when applying the OCC algorithm to this set, there will be two possible outcomes: either the model classifies a samples as positive (or an *inlier*), which is considered a true positive, or as an *outlier*, in which case it is considered a false negative, because it is known a priori that all the samples belong to the positive class. The proposed criterion can be seen as an analogous to the F1-score, and is given by:

[*]Except if one has access to a data set where the samples are labelled as *inliers* or *outliers*, although in that case it would be possible to train a regular binary classifier instead of using one-class classification algorithms.

$$\frac{p \cdot r}{\Pr[Y = 1]} = \frac{r^2}{\Pr[f(X) = 1]} \tag{2.17}$$

where r is recall and p precision. Just like recall, $\Pr[f(X) = 1]$ can be obtained from the validation set, and therefore the performance measurement can be easily calculated trough the right term of equation 2.17. Similarly to the F1-score, this metric is proportional to both precision and recall and they both present a similar behaviour.

Chapter 3

State of the Art

Given the high occurrence of metastatic PCa, there should be by now a more practical and, most importantly, more objective criterion to evaluate a bone scintigraphy. Fortunately, computerized tools like computer-aided diagnosis (CAD), machine learning and convolutional neural networks (CNNs) have been developed to improve the accuracy of exams and to increase consistency in interpretation of images. These tools can support physicians during diagnostic image evaluation and in the therapeutic decision-making process, making the task of lesion detection and quantitative assessment of disease burden much more objective, consistent, and reproducible (Brown et al., 2012, Koenigkam Santos et al., 2019).

CAD systems have developed several tools that are fundamental for medical imaging processing and analysis and, in particular, for the task of bone scintigraphy quantification. The literature found on this topic shows that there has been some effort to develop a computer-aided diagnosis system capable of automatically detecting and quantifying bone metastases in bone scintigraphy images. This process involves four phases which include image pre-processing, lesion detection and segmentation, feature extraction and classification. The following sections describe the various methods used in different studies to perform each one of these tasks. All of these studies were concerned with the development of an automatic algorithm for the quantification of bone scans.

3.1 Image Pre-processing

A common step to most works that aim to develop a CAD system for bone lesions classification is to pre-process the images to be used before performing any type of segmentation, detection or classification algorithm. This intends to not only enhance the image to improve the quality of original data prior to processing, but also to attenuate heterogeneity between bone scans that arise from differences between body physiques, radiotracer dosing levels, time between tracer administration and image acquisition, scanner type, and acquisition parameters (Brown et al., 2012). This will improve our image data and allow our algorithm to have a better performance during the following tasks.

A common procedure in to perform intensity normalisation, which transforms a grey-scale image by modifying the range of intensity values, resulting in a contrast enhancement. To perform a linear normalisation of their test set, Brown et al. (2012) and Brown et al. (2018) start by extracting the 75^{th} percentile from the intensity histogram of twenty high-quality bone scans, used as reference. The median of all the acquired values was defined as being the normal bone intensity in this reference set. The ratio between the 75^{th} percentile of each new scan and the median normal bone intensity previously obtained was used to linearly rescale the pixels in the images. Because bone scans vary greatly in intensity, using a reference histogram (or a set of histograms) to perform image normalisation in all images allows to have a more consistent data set which will be important for the detection and classification tasks. Shimizu et al. (2019) also performed a grey scale normalisation by modifying the pixels values with an intensity that fell in the upper 98^{th} percentile , i.e.:

$$I_{\text{normalized}} = \begin{cases} log_e\left(\phi \cdot \frac{I_{\text{in}} - I_{98\%}}{I_{10\%} - I_{98\%}} + 1\right), & I_{\text{in}} > I_{98\%} \\ 0, & \text{elsewhere} \end{cases}$$

where I_{in} is the input pixel value, $I_{10\%}$ and $I_{98\%}$ are the upper 10^{th} and 98^{th} percentile, respectively, and ϕ is the golden ratio. Another approach is histogram equalization, which reassigns the grey-level values of the pixels in the input image to obtain an image with an uniform intensity distribution. This was the method followed by Huang et al. (2007).

Another technique used during the pre-processing stage is to remove or attenuate noise

in the original images. Huang et al. (2007) used a 5×5 convolutional mask that approximates Gaussian distribution with a σ of 1.4 to smooth the noise inside the body region of the bone scans. To eliminate the noise from the background, they applied a threshold value that corresponded to a valley of the histogram found between two peaks at the low grey area.

When it comes to the pre-processing and preparation of data that will be used in a machine learning algorithm, there are other important steps that must be followed. Papandrianos et al. (2020a), Papandrianos et al. (2020b) and Papandrianos et al. (2020c) start by normalizing their images by rescaling all pixels values to fall within the range of 0 to 1. This procedure is specially important in machine learning algorithms to ensure that all feature data is in the same scale for training and testing, and also discarding possible outliers that could interfere with the algorithm performance (Papandrianos et al., 2020a). After that, a shuffling method is applied to give a random order to the data, which is followed by the data split stage that divides the data into three groups: training, validation and test. Finally, they perform data augmentation in the training set to artificially increase the sample size by using techniques such as rescaling, rotation zoom range, flipping, cropping or padding. A similar methodology, apart from the data augmentation step, is followed by Dang (2016) and Belcher (2017).

Although other techniques can be applied during the pre-processing stage, intensity normalisation and noise reduction are usually common to every work that involves image processing, and further data processing can be done depending on the specific problem one is working with.

Following the image pre-processing stage, one is finally ready to dive into the first step directly related to the work here developed, which in this case is detection and segmentation.

3.2 Detection and Segmentation

One of the most important medical imaging tool is image segmentation, which extracts the region of interest (ROI) from the background, thus being extremely useful for segmenting body organs/tissue or separating normal from abnormal structures. This is done by analysing (i) the values of the greys levels, (ii) discontinuities and gradients for edge

detection or (iii) similarity between pixels using thresholds or region-growing algorithms. In this work, an accurate and optimized segmentation of bone scintigraphy is needed to detect and segment bone lesions from the rest of the body (Koenigkam Santos et al., 2019, Guo and Ashour, 2019). This segmentation can be done trough a manual, semiautomatic or automatic process. The ideal solution would be an algorithm capable of segmenting the bone lesions in a fully automated way, with no intervention from an expert nuclear physician. Segmentation is also performed in bone scintigraphy to outline different regions of the skeleton, which is important if we need to specify the bone where the lesion is present.

There is no segmentation algorithm that can be applied to all types of image or disease, and the most appropriate method must be chosen by taking into account the task at hand and the type of image one is dealing with (Aslantas et al., 2017). A popular approach used in bone scintigraphy images is to perform a segmentation of the skeleton into several regions, followed by the application of a threshold to each region for the identification of hotspots.

For the purpose of developing an automated method for the interpretation of bone scans regarding the presence or absence of metastases, Sadik et al. (2006) used a combination of several image processing techniques like histogram analysis, image filtering edge detection and morphological operations to automatically segment the entire body. An optimized threshold value for each region was used to detect and segment the hotspots. Hotspots that were excluded right away included the ones with an area under six pixels, the one located on the bladder and symmetrical hotspots relatively to the spine, which were considered to be due higher bone turnover or arthrosis.

The CAD system was later improved by Sadik et al. (2008). They used an active-shape model method to segment the entire skeleton into four parts: head and spine, proximal arms and clavicles, chest, and pelvis and proximal legs. A set of training images was used to manually delineate the shape of the skeleton by selecting a set of landmarks on each anatomic region. The obtained shapes were aligned to a common coordinate frame to form a point distribution model that represented the mean geometry of each body part. The resulting shape model was then used to automatically segment new images by finding the best match position between the landmarks of the model and the data of the new image through an iterative process. Finally, the hotspots were detected using a

region-specific threshold that took into account the mean and standard deviation of the pixel values from each region.

Huang et al. (2007) aimed to build a CAD system capable of locating possible lesions in whole body bone scan scintigraphy. They used the fuzzy histogram thresholding method proposed by Tobias and Seara (2002) to separate bone regions from soft tissue. The resulting images were used to find reference points in the neck, shoulder, vertebra, pelvis, and arms. These points were then used to perform segmentation of the the head, arms and shoulders, pelvis, legs, vertebra and thorax. Finally, they studied the grey-level distribution from one hundred whole body bone scan images of healthy patients to determine the most suitable threshold value for hotspot detection in each of the segmented regions. For the detection of malignant lesions, this system obtained an overall sensitivity of 92.1% and 7.58 false positive per patient.

Brown et al. (2012) developed a computer-aided system to automatically segment and quantify bone scan lesions. The bone lesion segmentation was accomplished by doing an atlas-based anatomic segmentation to divide the body into 6 different regions: sternum, spine, ribs, head, extremities, and pelvis. After an intensity normalisation, they proceeded to bone lesion segmentation by applying region specific thresholding, i.e., each of the previously segmented regions had a specific thresholding value, that would optimise the lesion detection in each body part. These values had been previously determined by a ROC analysis. This lesion segmentation algorithm was validated by measuring the tum or pixels it detected and comparing the results with the ones obtained by experienced nuclear medicine physicians by visual assessment. The method achieved a median sensitivity of 94.1%, specificity of 89.2%, and accuracy of 89.4%. With the purpose to develop a completely automated decision support system for whole body bone scans, Ohlsson et al. (2009) also performed an atlas-based segmentation to divide the skeleton into twelve anatomical regions: skull, cervical vertebrae, thoracic vertebrae, lumbar vertebrae, sacrum, pelvis, ribs, scapula, humerus, and femur, clavicle and sternum. The segmentation of the hotspots was achieved by applying a threshold to a band-pass filtered version of the image.

CADBOSS, developed by Aslantaş et al. (2016), was another CAD system develop for the assessment of bone metastases in bone scintigraphy scans. It performs hotspot segmentation, feature extraction and selection and classification of the image as a whole. The

segmentation of hotspots was carried out by a level set active contour algorithm proposed by Li et al. (2007). By the end of this stage, a binary image was obtained, with black representing the background and white the possible hotspots.

A more recent technique uses artificial neural networks for the task of skeleton and lesion segmentation. Having in view the calculation of the BSI, Shimizu et al. (2019) proposed a deep-learning based approach to perform skeleton segmentation and hotspot extraction in whole-body scintigraphy. For that they used a butterfly type network, BtrflyNet, that is able to process both anterior and posterior images simultaneously by fusing two U-Nets. Besides some evident differences regarding the size of the input and output images, as well as the number of output nodes, the structure of the networks for skeleton segmentation and lesion detection was very similar. A thorough description of both networks is given in the article. Regarding the skeleton segmentation, the BtrflyNets received as input a pair of anterior and posterior images, and had an output layer with a size equal to the number of bones the skeleton was being divided into, plus the background (thirteen and twelve layers for AP and PA images, respectively). On the other hand, the hotspot extraction network would receive a pair of anterior and posterior patch images of 64×64 pixels. The output consisted of three layers corresponding to (i) bone metastatic lesions, (ii) benign lesions such as fractures and infections, and (iii) other non-malignant hotspots like physiological renal uptake and radiotracer uptake in the place of administration.The Dice scores* for AP and PA segmentation were 0.842 and 0.882, respectively. The performance for hotspot segmentation was obtained by measuring the average number of false positive pixels (192.5 and 237.9 for AP and PA, respectively), the number of false positive regions (10.0 and 9.41 for AP and PA), and the misclassified pixels (268.8 and 320.2 for AP and PA). In addiction to having an acceptable computational time for clinical use, this algorithm proved to be effective in segmenting the skeleton and in the detection of hotspots.

A summary of the articles, methods and results obtained for the the task of skeleton an lesion segmentation is given in table 3.1.

*The Dice score measures the similarity between two samples; in this case, it was used to measure the similarity between the segmented bone region and true region

TABLE 3.1: Summary table with an overview of the articles and respective methods and results for the tasks of skeleton an lesion segmentation. As the main goal of these studies was to classify hotspots on bone scans, the results presented by the articles often refer to the classification and not to the detection, hence the few results for that task.

Article	Skeleton segmentation	Lesion detection	TP/TN	FP
Sadik et al. (2006)	Image processing	Region-specific threshold	—	—
Sadik et al. (2008)	Active-shape model	Region-specific threshold	—	—
Huang et al. (2007)	Image processing	Region-specific threshold	92.1/ —	7.58
Ohlsson et al. (2009)	Atlas-based	BPF & threshold	—	—
Brown et al. (2012)	Atlas-based	Region-specific threshold	94.1/89.2	—
Aslantaş et al. (2016)	—	Level-set active contour	—	—
Shimizu et al. (2019)	CNN	CNN	—	9.70

The True Positive (TP) and True Negative (TN) rates are given in % and the False Positives (FP) are given in detected hotspots per patient; these results refer to the lesion detection task. BPF and CNN refer to Band-Pass Filter and Convolutional neural networks, respectively.

3.3 Feature Extraction

Having detected all possible hotspots, it is necessary to identify useful characteristics and attributes of the data that will allow its distinction and classification. This technique is referred to as feature extraction. It effectively reduces the amount of data while preserving the information from the original data set, making the task of pattern classification easier. This will play an important role in the recognition and characterization of bone lesions. Image features have an hierarchical organisation which goes from low-level to high-level features. Low-level features include information about color, shape, texture and spatial location, and their extraction is based on imaging processing techniques. This type of features is used in the CAD systems developed by Sadik et al. (2006), Sadik et al. (2008) and Ohlsson et al. (2009), previously described in section 3.2. While in the first one features were extracted from the image as a whole, e.g., number and distribution of hotspots, hotspots coverage and coefficient of variation in different regions of the body, on the second and third ones the features were extracted from each detected hotspot. These included area, geometry, mean, standard deviation and maximal pixel values, skeleton region, area ratio of hotspot to region, among others. In the three cases the extracted features constituted the inputs of the ANNs used for classification.

After the hotspot detection phase described in section 3.2, which resulted in a binary image with black representing the background and white the possible hotspots, Aslantaş et al. (2016) divided each resulting image into 25 patches with size 8×28 pixels, thus obtaining 625 sub-images. There were, however, too many black pixels coming from the

background of the original bone scintigraphy images, which could negatively affect the performance of the classification process. To overcome this possible problem, they transferred the average value of each sub-image into a matrix, and used Principal Components Analysis (PCA)* to reduce its dimension by finding the principal components that retained most of the information. These were then used as inputs to an ANNs for classification.

High-level features are object-based and their extraction is mostly based on machine learning algorithms (Crommelinck et al., 2016). Using deep learning, a machine learning technique method based on artificial neural networks, the computer is capable of automatically learning features that optimally represent the data for the task at hand (Litjens et al., 2017). Because the algorithm learns to identify and extract features to perform a specific problem, it becomes extremely effective at it and has thus been gaining a wide recognition in the biomedical field for the several applications it may have in medical image analysis. One way to obtain this type of features is by using a pre-trained image classification network to extract high-level features directly from raw images and use them to perform a task different from the one the network was trained to, a technique known as transfer learning. Autoencoders are another method used for feature extraction. They play an important role in unsupervised machine learning since they are a type of ANN for which the input is the same as the output, that is, they do not require labelled data. They work by compressing the input to a latent-space representation (encoder), and then reconstructing the output from this representation, as close as possible to its original input (decoder). Both of these methods are quite popular in the field of unsupervised learning since they constitute a strategy to extract features in an unsupervised manner, which is very useful when labelled data is not accessible. There is still no published literature on the use of pre-trained networks and autoencoders for feature extraction from bone lesions in planar bone scintigraphy. Nonetheless, the interest in performing classification tasks in entirely unsupervised setting has been growing over the recent years and several papers have reported successful approaches to this problem by extracting features from pre-trained models and autoencoders. Some of this examples can be found in Kumar et al. (2015), Cohn and Holm (2020), Alaslani and Elrefaei (2018) and Khan et al. (2019), where an autoencoder was used to extract deep features from 2D CT images to build a

*Principal Components Analysis is a method used for dimension reduction, in which an initial set of features from a high-dimensional space is projected into a reduced set of features from a low-dimensional space, revealing a simplified structure of the initial data while preserving most of the information (Xu, 2018).

CAD system for lung cancer detection and pre-trained networks were used to extract features for classifying hot-rolled steel defects observed in micrographs, to develop an iris recognition system, and for bone lesion detection in CT scans in patients with multiple myeloma, respectively.

In table 3.2, a summary of the articles and respective methods and results for the task of feature extraction is given.

TABLE 3.2: Summary table with an overview of the articles and respective methods and results for the task of feature extraction.

Article	Method	Extraction from
Sadik et al. (2006)	handcrafted	whole-body
Sadik et al. (2008)	handcrafted	hotspot
Ohlsson et al. (2009)	handcrafted	hotspot
Aslantaş et al. (2016)	handcrafted	whole-body

3.4 Hotspots Classification

The next step focuses on the classification of the hotspots. This can be achieved through a manual process or through fully automated techniques such as machine and deep learning algorithms. Under the latter category, we can still have supervised and unsupervised methods.

An example of a CAD system that uses manual classification can be found in Brown et al. (2012). After the automatic detection of hotspots described in section 3.2, the resulting images were reviewed by a nuclear medicine physician who removed false-positive lesions. The hotspots classified as malignant were used to assess the severity of the disease and disease response to treatment. The final classification would later be used to evaluate the BSLA as biomarker for overall survival in PCa patients subject to drug treatments. Despite the good results, this algorithm is not fully automatic, as it requires the intervention of a physician to remove false-positives (non-metastases related bone uptakes) from the scans. This is an huge downside as the automatic differentiation between malignant and non-malignant bone uptakes is an essential requirement in a bone metastases evaluation algorithm, as it is a task that is not trivial even for the most experienced physician and thus brings a lot of subjectivity to the final assessment. A classification algorithm capable of automatically distinguish metastases from benign lesion is thus needed.

The following studies approach automatic algorithms that use machine and deep learning for whole body scans and hotspot classification. In particular, they fall under the category of supervised learning.

In 2006 and 2008, Sadik et al. developed a fully automated classification system for the detection of metastases that used artificial neural networks. Both works intended to classify the whole-body bone scan as a whole, regarding the presence or absence of bone metastases, and not the hotspots individually. The ground truth for each body scan was provided by experienced physicians who estimated the probability of bone metastases on a scale from 0 to 1, based on the images and clinical reports. Patients with an estimated probability lower than 0.5 were classified has having no metastases and the ones with a probability of 0.5 or higher were classified as pathological. In Sadik et al. (2006), the classifier consisted on an assemble of ANNs with an input layer with fourteen nodes (one for each extracted feature), a hidden layer with ten nodes and an output layer with one node that returned zero for no metastases and 1 for metastases. The classifier would then compute the mean of all the individual values predicted by each individual member of the assemble, returning a value between 0 and 1. Overall, this algorithm obtained a sensitivity of 90% and a specificity of 74%. In Sadik et al. (2008), they started by building an ANN to assess the likelihood that a specific hotspot represented a metastasis. An ensemble of ANNs similar to the ones described in the previous work was used, only differing in the number of nodes in each layer: forty-five nodes on the input layer, ten on the hidden layer and an output node to classify the lesion has being a metastasis or not. Twenty-six features extracted from the four hotspots in each scan with the highest outputs (e.g., highest probability of representing a metastasis) were then used in another assemble of ANNs to classify each scan as a whole. The classifier would again return a value between 0 and 1, that reflected the probability of the patients having metastases. This new CAD system had the same sensitivity of 90% but achieved a higher specificity of 89% compared to the one developed in 2006. Although these two studies used neural networks, other algorithms like Support Vector Machines (SVM), Decision trees of k-Nearest Neighbors (k-NN) could be applied to perform supervised classification. Literature on this topic can be found, for example, for bone lesion detection in Computed Tomography (CT) (Kumar and Suhas, 2016), but not for bone scintigraphy. This methods do require a previously manual extraction of features that are then fed into a classifier.

Instead of manually extracting the features and use them as input in a machine learning algorithm, a logical step is to let the computer learn useful features that optimally represent the data at hand (Litjens et al., 2017). This concept is the basis of deep learning, a machine learning technique based on artificial neural networks that is capable of automatically learn features from big data to solve a specific problem, like the classification of a lesion as malignat or benign. In particular, convolutional neural networks (CNNs) have proven to be a powerful tool in computer vision tasks like image segmentation, object detection and image enhancement and reconstruction (Ginneken and Summers, 2016). Because they learn to identify and extract the features that will have the highest impact on a particular classification task, they become extremely effective at it and have thus been gaining a wide recognition in the biomedical field for several applications they may have in image analysis. The use of CNNs can thus be a huge benefit to the classification and quantification of bone scintigraphy.

Papandrianos et al. (2020a, 2020b, 2020c) have published three papers describing the work they have made on this field, devoted to the development of CNN models for automatic classification of whole-body scans from patients with bone metastases. Just like Sadik et al. (2006), the authors intended to classify the body scans as a whole, and not the hotspots individually. In Papandrianos et al. (2020a) and Papandrianos et al. (2020b) they were dealing with a two-class classification problem regarding the presence (malignant scan) or absence (healthy scan) of bone metastases in patients with breast and prostate cancer, respectively. A nuclear medicine physician labelled all the images in the data set as belonging to each of one of the two categories, and this was used as ground truth. As they aimed to cope with a two-class classification problem, all scans from patients containing degenerative lesions and other non-malignant bone uptakes were removed in a manual pre-selection process. The proposed CNN architectures were very similar and consisted of a deep-layer network with an image input size of $256 \times 256 \times 3$ pixels with three convolution-pooling layers; in Papandrianos et al. (2020b) they used one dense layer followed by a dropout layer and an output layer with one node while in Papandrianos et al. (2020a) they used one dropout layer followed by a dense layer and an output layer with one node. They used the rectified linear unit (ReLU) as the activation function in the convolutional and fully connected layers and the sigmoid function in the output nodes. In Papandrianos et al. (2020a), the best CNN architecture (the one that produced the best results) had a accuracy of 92.50%, a sensitivity of 94% and a specificity of

92%. In Papandrianos et al. (2020b), the best CNN architecture had a classification accuracy of 97.38%, a sensitivity of 96.5% and a specificity of 96.8%. The major problem with these systems is that the images from patients containing degenerative lesions and other non-malignant bone uptakes were removed from the original data set and therefore the networking didn't learn to identify this type of scan. This is a major drawback as a fully automatic algorithm to assess whole body scintigraphy should also be able to classify false-positive bone uptakes as benign lesions.

In Papandrianos et al. (2020c) the same authors investigated a way to partial solve this problem. They developed a similar CNN based algorithm to classify bone scintigraphy images as healthy, malignant or degenerative, leading to a three-class classification problem. The best CNN architecture achieved a sensitivity of 92.7% and a specificity of 96.0%. Although the automatic distinction between malignant and non-malignant images is an improvement over the previous models, it does not offer a solution for the cases in which one patient has bone uptakes with both malignant and non-malignant origins, which is one of the major problems in visual bone scintigraphy assessment. In fact, neither of the papers proposed by Papandrianos et al. (2020a) and Sadik et al. (2006) propose an algorithm that is capable of quantifying the bone lesions individually, which is essential when an objective assessment of the disease staging is needed. It is not enough to build an algorithm that is able to distinguish images that present solely malignant lesions from those that present solely benign lesions. A suitable algorithm must be able to quantify and classify each lesion individually.

The following two studies will approach this problem. Nonetheless, they were only concerned with the classification of hotspots, leaving aside the task of detection and segmentation. This was possible due to a hotspots database provided by EXINI Diagnostic AB, a Sweden based company that uses artificial intelligence to develop automated analysis platforms for medical images like cardiac, brain and bone scans (EXINI Diagnostics AB, 2020). It has shown to be quite popular among researchers concerned with quantification of bone metastases. EXINI has developed the aBSI (automated Bone Scan Index), a software only medical device that provides a fully quantitative assessment of a patient's skeletal disease on a bone scan, as the fraction of the total skeleton weight (aBSI, 2019). As it is a closed-source software, little is known about its operating principles, except that it was trained to classify hotspots as lesions using a collection of more than 40000 hotspots

derived from bone scans of patients with a variety of metastatic cancers. It is able to segment the skeleton, identify hotspots, quantify their intensity and classify them as lesions (Ulmert et al., 2012). After the initial development of aBSI by EXINI, the software was further developed and validated in Japan and nowadays it is used by the Japanese nuclear medicine community to calculate the BSI in metastatic prostate cancer. The revised platform was launched in 2011 with partner FUJIFILM RI Pharma under the name BONENAVI (EXINI Diagnostics AB). The software was developed by retraining the CNN with a different database consisting of Japanese patients. A study presented by Horikoshi et al. (2012) suggests that algorithms trained with different databases will have different performance in different populations. They concluded that a CAD system based on a Japanese database showed significantly higher performance in interpreting bone metastases in Japanese patients than a CAD system trained with an European database.

It is therefore important to note that the studies that mention EXINI have access to hotspots that were collected, segmented and cropped from bone scans using programs developed at EXINI, and given to the researchers for analysis purposes. Most importantly, the data set consists of hotspot images already labelled as "high risk" or "low risk", i.e., this was a supervised machine learning algorithm.

In his Master thesis, Dang (2016) developed a CNN to classify hotspots in bone scintigraphy images for prostate cancer. The main task of this work was to determine whether hotspots had a high or low risk of being bone metastases from PCa metastatic cancer. The data consisted of 10428 labelled hotspots provided by EXINI, all of them coming from the spine, as he believed that these would be the easiest to classify. The CNN was implemented in Keras, a deep learning application programming interface (API) written in Python, the used hyper parameters being given in the thesis. The trained CNN had an accuracy of 0.890, a F1-score of 0.919, a true positive rate of 0.981, a true negative rate of 0.649 and an AUC (Area Under the ROC Curve) of 0.955. This software might be the one that comes closest to what we want to achieve with this thesis, as it is able to classify bone lesions individually as malign and non-malign. Nonetheless, it should be noted that he did not have to worry about the segmentation or the labelling of the lesions, as he was given access to a large data set of labelled hotspots.

A very similar Master thesis was developed by Belcher (2017). He also resorted to a data set provided by EXINI that contained 10427 hotspots labelled as positive (high risk of

metastases) and negative (low risk of metastases) to build a CNN for hotspot classification in bone scintigraphy. The software was developed in Python with the Tensorflow library. To measure the performance of the CNN he only used the area of the ROC curve, for which he obtained a score of 0.974.

The previously described works use supervised techniques which, despite appearing to be a promising approach to the classification problem, rely on an extensive number of labelled data. Such large scale annotated data sets are, however, very rare in the medical context. Training a CNN from scratch to perform bone lesion classification would require thousands of labelled images, a task that would not only be extremely complex and time consuming, but also dependent on the availability of experienced physicians. Furthermore, the labelling would be subject to the subjectivity inherent in the classification of lesions detected in bone scintigraphy.

To address this challenge, unsupervised algorithms are used to draw inferences on unlabelled data sets by finding natural patterns in the data to determine class labels, which in turn can be used for image segmentation, object detection and classification. One of the most popular approaches is clustering, a technique that tries to find a structure in a collection of unlabelled data by segregating it into groups based on their similarities. After a manual or deep learning based approach for feature extraction, several clustering algorithms like K-means clustering, Gaussian mixture model, hierarchical clustering and spectral clustering can be employed to attribute labels to the data (Ahn et al., 2019). Clustering can be found in several medical studies for the diagnosis of diseases like Parkinson (Polat, 2012), breast cancer (Chen, 2014) or Alzheimer (Alashwal et al., 2019), but none can be found for classification of hotspots in whole-body bone scintigraphy.

Another strategy that falls into the category of unsupervised classification is anomaly detection, also known as outlier detection, which seeks to identify examples that do not fit to the characteristics of the "inlier" observations, that is, that are inconsistent with the remainder of the data set (Johnson and Wichern, 2007). In the medical context, this can be achieved by training an algorithm with only healthy anatomical samples and later identify regions that present a significant discrepancy comparatively to the healthy observations, by either probabilistic, distance-based, reconstruction-based, domain-based or information-theoretic methods (Alaverdyan, 2019). This type of approach falls under the category of one-class classification problems, a method that has been successfully applied

in many application domains, where the training set contains examples of only one class and the aim is to classify new examples as either belonging or not belonging to that class. El Azami et al. (2016), Alaverdyan (2019), Gardner et al. (2006) and Spinosa and Carvalho (2005) used a one class support vector machine (OC-SVM) algorithm to classify images in the medical context. One can also use variational autoencoders (VAE)* to model the distribution of healthy data, by training the network exclusively with healthy images. When given any type of image, the model should be able to detect regions that reveal deviations from the norm, which are classified as lesions. This is the method followed by both Baur et al. (2018) and Chen et al. (2019).

Table 3.3 shows an overview of the articles, respective methods and results for the task of hotspot classification. As one can see, no literature of completely unsupervised algorithms for hotspot detection and classification could be found to this date, which makes this thesis a pioneer on the topic.

TABLE 3.3: Summary table with an overview of the articles and respective methods and results for the task of hotspot classification.

Article	Classification of	Method	TP (%)	TN(%)	AUC (%)
Sadik et al. (2006)	whole-body	ANN	90.0	74.0	—
Sadik et al. (2008)	whole-body	ANN	90.0	89.0	—
Papandrianos et al. (2020a)	whole-body	CNN	94.0	92.0	—
Papandrianos et al. (2020b)	whole-body	CNN	96.5	96.8	—
Papandrianos et al. (2020c)	whole-body	CNN	92.7	96.0	—
Brown et al. (2012)	hotspots	manual	—	—	—
Dang (2016)	hotspots	CNN	98.1	64.9	95.5
Belcher (2017)	hotspots	CNN	—	—	97.4

3.5 Validation of the BSI as an imaging biomarker

After the classification is completed, it is necessary to find a method that is able to quantify the tumour burden of a patient. This metric should be precise, reproducible, serve as a reliable marker of disease progression and treatment effects and have a good prognostic ability. Some of the quantitative methods most commonly found in literature are the BSI (Bone Scan Index), BLS (Bone Lesion Scoring), EOD (Extent Of Disease) and PAB (Positive Area on Bone scans). The effectiveness of these methods can be determined by implementing the parameters on individual baseline bone scans and again after the

*A VAE is a type of autoencoder in which the input is encoded as a distribution with mean μ and standard deviation σ, from which a point is sampled from the latent space which is then passed onward to the decoder.

patient has received treatment. During this follow up, an evaluation of the patient condition is made and, by comparing both values from the baseline and follow up scans, it is possible to assess how effective the different quantitative parameters are at describing disease progression or regressions and at predicting patient survival. Some studies also evaluate the relationship between the values of these parameters and other biomarkers and pathological grading systems used in the diagnosis and screening of prostate cancer pathological grading system, such as the Prostate-Specific Antigen (PSA) and the Gleason score

Mustansar (2018) performed a study which aimed to evaluate and compare the four bone scan quantitative parameters previously mentioned. A total of 141 patients with prostate cancer was initially included in this study, and a follow up was performed on 40 of those patients. The assessment of tumour burden on bone scan baseline and follow up was achieved by applying each of the methods (BSI, EOD, PAB and BSL) to each of the bone scintigraphy. They compared each of the quantification methods with PSA levels, using BSI, EOD, PAB and BSL as the dependent variables and PSA as the independent one, and evaluated the goodness of the model using the R-squared coefficient. The PSA is thus used as true indicator of disease status, since it is widely used as a marker of disease progression or regressions, higher values of PSA between treatments indicating disease progression and less probability of survival and vice-versa (Moradi et al., 2019). BSI and PAB showed the best linear correlation with PSA values, with R^2 of 0.891 and 0.929, respectively. EOD and BSL showed a weaker linear association, with R^2 values of only 0.610 and 0.518, respectively. An analysis was then made on how the variation of the values measured by each of the four methods in the baseline and follow up scans was related to diseases progression and to patient survival. It was concluded that all the parameters were good in describing disease progression, as a decrease of the value from the baseline bone scan to the follow up bone scan indicated a decreased risk of disease progression and better survival. For all of the parameters, it was possible to identify a cut-off value which indicated an increased risk of disease progression. Although all the four quantities showed to be good indicators of disease status and progression, PAB and BSI were the most accurate in calculating the tumour burden, as they were highly correlated with PSA levels. Although PAB is, according to Mustansar (2018), easier to calculate when compared to BSI, BSI is much more frequently found in literature. In particular, an automated

software for BSI calculations has been developed, and various studies have worked on validating this quantitative parameter as a biomarker for PCa.

Kaboteh et al. (2013) used the EXINI software to calculate the BSI in 130 patients recently diagnosed with high risk PCa who received primary hormonal therapy. They sought to investigate the relation between BSI and clinical stage, Gleason score, PSA and survival. They divided the total group into four subgroups according to the value of BSI: BSI = 0 for patients with no metastases, BSI$<$ 1, BSI $=$ 1 to 5 and BSI $>$ 5, according to the tumour burden. The Kaplan-Meier curves of the patient-survival probability for the patients in each of the subgroups were statistically significant ($p < 0.001$), and the 5-year survival probabilities decrease with the increase of the BSI, having values of 55%, 42%, 31% and 0%, respectively. These results showed that BSI is strongly associated with overall survival in patients with high-risk prostate cancer receiving primary hormonal therapy and can thus be considered an informative predictor of patient survival. It has shown that the BSI has prognostic information which can be used, along with other measurements like the PSA and the Gleason score, to assess the stage of the disease.

Poulsen et al. (2016) conducted a study which confirmed the reliability of the BSI as an effective imaging biomarker and a prognostic factor. They performed univariate and multivariate analyses on time to prostate cancer-specific death (PCSS) and on time to castration resistant prostate cancer* (CRPC) using the PSA level, Gleason score and BSI as explanatory factors. The BSI was a statistically significant prognostic factor in all analysis; in particular, in the multivariate analysis for the time to CRPC, only the prediction by BSI was statistically significant.

Similar studies providing evidence for effectiveness of the BSI as a prognostic factor in patients with bone metastases were developed by Reza et al. (2014) and Inaki et al. (2019).

Ulmert et al. (2012) studied the correlation between manual and automated BSI measurements and how the incorporation of each method into the base model, which included clinical stage, Gleason score, and total PSA in blood, would affect its predictive accuracy. The automated BSI was calculated using the EXINI software and the manual BSI was carried out by an experienced analyser. The correlation between manual and automated BSI

*Castrate-resistant prostate cancer (CRPC) is defined by disease progression despite androgen deprivation therapy (ADT), an hormone therapy used to reduce the levels of androgen hormones, which stimulate prostate cancer cells to grow (American Cancer Society).

calculations was high, having a Pearson's correlation coefficient of $\rho = 0.80$. This correlation was found to be even higher if only BSI less than 10 was considered ($\rho = 0.93$): there was an higher agreement between BSI values in patients with milder cases of the disease, compared to patients with more extensive bone metastases. To determine whether BSI measurement added prognostic value to the base model, they used the C-index to compare the discrimination power of the different models. They concluded that including both the manual and automated BSI measurements individually to the base model increased its predictive accuracy: the C-index of the base model was 0.768, which increased to 0.794 when adding manual BSI and to 0.825 when adding automated BSI. Just like in the previously described papers, they also concluded that both manual and automated BSI were independently associated with disease-specific death. This study shows that the BSI is an important clinical parameter which can add valuable information in the clinical context of patients with PCa. They also highlighted the advantages of using an automated method over a manual one, not only because of its rapid processing time but also for eliminating physician-dependent subjectivity, which in turn makes the automated BSI scoring 100% reproducible.

Li et al. (2017) published a meta-analysis in which they combined the results from several studies that investigated the relationship between BSI and survival in patients with mPCa. Their final selection consisted of 14 studies published from 2010 to 2017, of which 11 used the Swedish EXINI software and other three used the Japanese BONENAVI system. They analysed how the baseline BSI and BSI change during treatment (ΔBSI) could be predictive of poor overall survival and how the baseline BSI could be predictive of cancer specific survival prostate specific antigen recurrence survival. The results demonstrated that they were all significantly related, with hazard ratios of 1.29, 1.27, 1.65 and 2.26, respectively. They also presented the ΔC-index* values which corresponded to the difference between the C-indices of the models used to predicted OS and CSS with and without the BSI value. All ΔC-indices were greater than zero, which means that BSI could increase the predicting ability of OS and CSS in mPCa. They thus concluded that BSI could be an useful imaging biomarker in mPCa prognosis, as well as a complementary tool in monitoring patients during treatment.

*The concordance index (C-index) is used to evaluate the predictive ability of a survival model, by measuring its ability to correctly provide a reliable ranking of the survival times based on the individual risk scores (Fotso et al., 2019).

Although other quantitative parameters are available, the BSI seems to be the most popular choice as an imaging biomarker for PCa patients with bone metastases, and is definitely the quantity for which more literature can be found. An automated CAD system for the assessment of bone scans should therefore include the BSI, as it has proven to be a reliable and reproducible biomarker that can objectively evaluate the severity and degree of change of a metastatic PCa patient's condition.

Several methods have been proposed for the development of an automated system to assist physicians during the evaluation and follow-up of patients with bone metastases. Nonetheless, there is still no open-source software available for clinical use. This work not only intends to develop that kind of software, but it also proposes to do it resorting to methods that haven't been tried before when addressing this specific challenge, i.e., using unsupervised machine learning algorithms. This work is, therefore, a pioneer on that matter and, if proven successful, it could give an huge contribution to the clinical practice.

Chapter 4

Development

In this chapter, the adopted methods to undertake the task of bone scan quantification are covered. Information about the data that was available and how it was analysed for the current research will be given in Section 4.1. Next, the methods used to perform each task, along with a proper justification for their choice, will be discussed. These tasks include:

- Detection of the hotspots (Section 4.2)
- False positive attenuation (Section 4.3):
  - Anatomical segmentation (Section 4.3.1)
  - Attenuation of false-positives through image processing techniques (Section 4.3.2)
  - Feature extraction (Section 4.3.3)
  - Attenuation of false-positives through machine learning algorithms (Section 4.3.4)
- BSI calculation (Section 4.4)

Figure 4.1 shows the methodology overview.

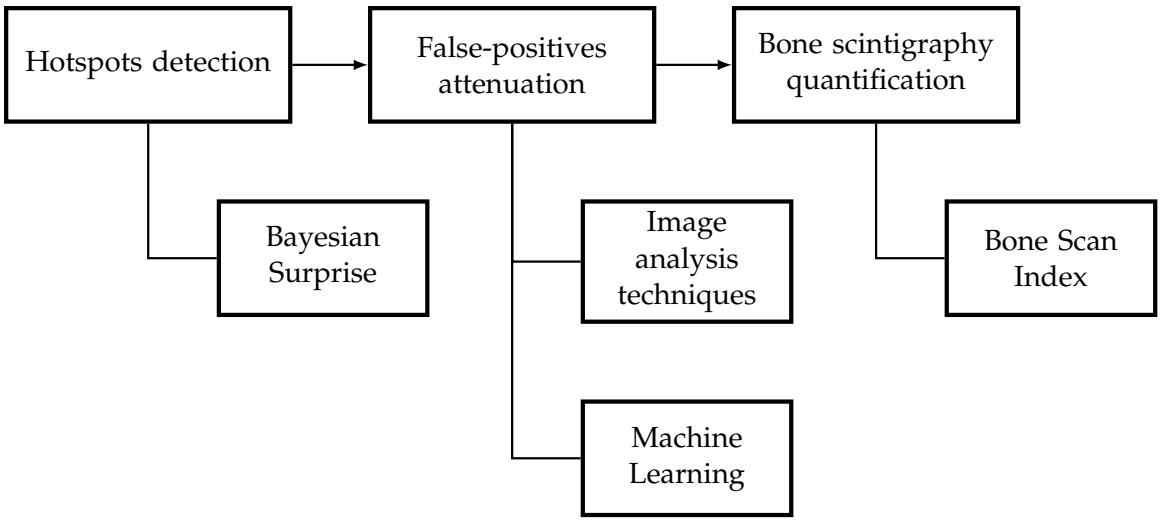


FIGURE 4.1: Methodology overview

4.1 Database

The database consists of 195 bone scintigraphy images from 102 patients with prostate cancer with suspected bone metastatic disease. The equipment used for scanning patients was either a *Millennium MG* (GE Medical Systems), which digitally records anterior and posterior scans with a resolution of 1024×256 pixels, or a *BrightView* (Philips Healthcare), which digitally records anterior and posterior scans with a resolution of 1024×512 pixels. The pixel depth (maximum number of counts which could be stored in a pixel) is 16-bits for every image. For each bone scan, a medical report written by a nuclear medicine physician describing the condition of the patient in question is available. All data was provided by Instituto Português de Oncologia do Porto Francisco Gentil (IPO Porto). The data was collected and held anonymously and the developed algorithms did not contain information concerning the patients, but rather information extracted from the data during the algorithm development. This project was authorised by IPO Porto Healthcare Ethics Committee.

Data splitting

Upon examination of the medical reports, the data set was divided into three subsets:

- The *healthy* subset, composed of 37 bone scintigraphy images from patients with no suspicious bone uptakes (neither benign nor malignant);

- The *benign* subset, composed of 72 bone scintigraphy images from patients with benign lesions, that is, with no metastatic origin. The hotspots detected in this set can therefore be benign or healthy.

- The *malignant* subset, which contained 86 bone scintigraphy images from patients with metastases. The hotspots detected in this set can therefore be malignant (metastases), benign or healthy.

Table 4.1 summarises the available database. The data set was then divided into a training and test set. The test set consisted of 30 patients randomly chosen from the healthy, benign and malignant sets. Two different divisions of the data were performed, depending if working with a 3-class or a 2-class classification problem. When dealing with a 3-class problem, the data was dived into *healthy*, *benign*, and *malignant* classes. When dealing with a 2-class problem, the *healthy* and *benign* classes were merged into one, forming a *non-malignant* class, while the *malignant* class was kept unchanged. The number of patients and hotspots detected* per category for the training and the test set for the case of the 3-class and the 2-class problems are presented in Table 4.2 and Table 4.3, respectively.

TABLE 4.1: Database summary. The database consisted of a total of 195 bone scans divided into one of three classes: healthy, if no suspicious bone uptakes were detected, benign if the patients presented bone hotspots with benign origins, or malignant , if the patient had bone metastases.

Bone scan type	No of bone scans
Healthy	37
Benign	72
malignant	86
Total	195

TABLE 4.2: Split of the data set for a 3-class problem

Bone scan category	No. of patients		No. of detections	
	Training	Test	Training	Test
Healthy	27	10	418	138
Benign	62	10	1 523	255
Malignant	76	10	5 620	918
Total	**165**	**30**	**7 561**	**1 311**

*The hotspots were detected using the method described in Section 4.2

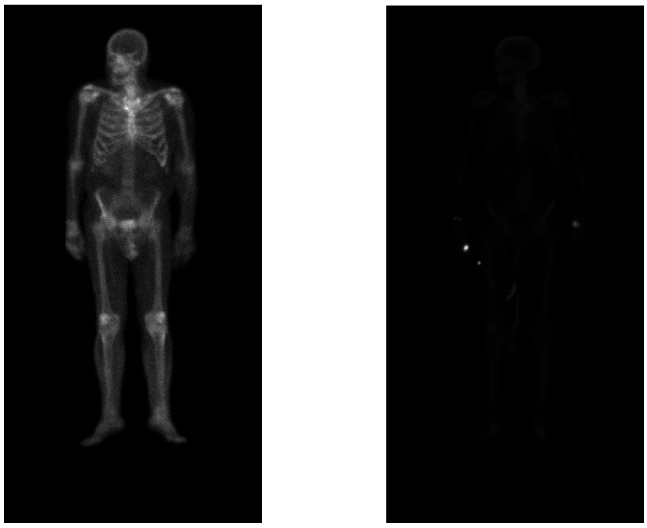
TABLE 4.3: Split of the data set for a 2-class problem

Bone scan category	No. of patients		No. of detections	
	Training	Test	Training	Test
Non-malignant	89	20	1 941	393
Malignant	76	10	5 620	918
Total	**165**	**30**	**7 561**	**1 311**

4.2 Detection

The first step required to quantify a bone scan involves detecting the regions in the skeleton where there is an increased bone uptake. As explained in section 2.3, these areas are called hotspots and will appear as brighter regions in the bone scans. An easy solution to this problem would be to simply apply a threshold to the images, as the hotspots present higher grey levels compared to the rest of the skeleton. This procedure has, however, two problems. The first one is due to the fact that the quality of the images obtained with the gamma camera will vary between bone scans. Because images with poor quality will appear a lot darker (see Figure 4.2), using the same threshold used for good quality scans would mean that a lot of hotspots would be left undetected in the darker scans. Although this could possibly be solved by performing a histogram matching so that the histogram of a bone scan of worse quality would match that of a bone scan with better quality, a second problem is left unsolved. This problem arises from the fact that different regions of the skeleton will have, following naturally occurring physiological processes, different bone turnover. As a consequence, the final image will present brighter regions that are not related to the presence of metastases but rather to a higher bone remodelling activity, as it happens with the spine. A unique threshold value would not, therefore, be suitable for every region of the body.

To overcome these problems, the detection of hotspots was made following a simpler version of the approach proposed by Domingues and Cardoso (2014b), where a technique based on Bayesian surprise is used to detect calcifications in mammogram images. The algorithm used in this thesis takes advantage of the fact that the hotspots are brighter regions (that is, regions with higher grey levels) surrounded by pixels with lower grey values. The first step of the algorithm consisted in applying a mask to the original image to exclude the background and to keep solely the body of the patient. Then, the hotspots were detected through the following steps:



(a) (b)

FIGURE 4.2: Two bone scans (AP view) from two different patients showing different image quality. As one can see, the body scan in (b) appears very dark and would require a much lower threshold for detecting hotspots than the body scan in (a)

1. Consider a square patch of the masked image with apothem r_{in};
2. Consider the region surrounding the patch described in 1, defined by an apothem $r_{\text{out}} = \sqrt{2} \cdot r_{\text{in}}$ and with centre coinciding with that of the inner patch;
3. Calculate the mean grey level of both the inner patch and the surrounding region;
4. Compare the mean grey levels: if the absolute difference of the two values is higher than a certain threshold δ, the inner patch is considered a hotspot.

This algorithm is summarised in the diagram of Figure 4.3.

Figure 4.4 illustrates the inner and outer patches for a metastasis located at the third lumbar spine vertebra.

The steps were repeated for every patch in the masked image with the following values:

- $r_{\text{in}} = 5\text{cm}$
- $\delta = 20$

The value of the threshold was manually obtained by trial and error, by visually analysing the hotspots detected by the algorithm. Higher values would detect fewer hotspots, but would left some bone metastases undetected; on the other hand, lower values would

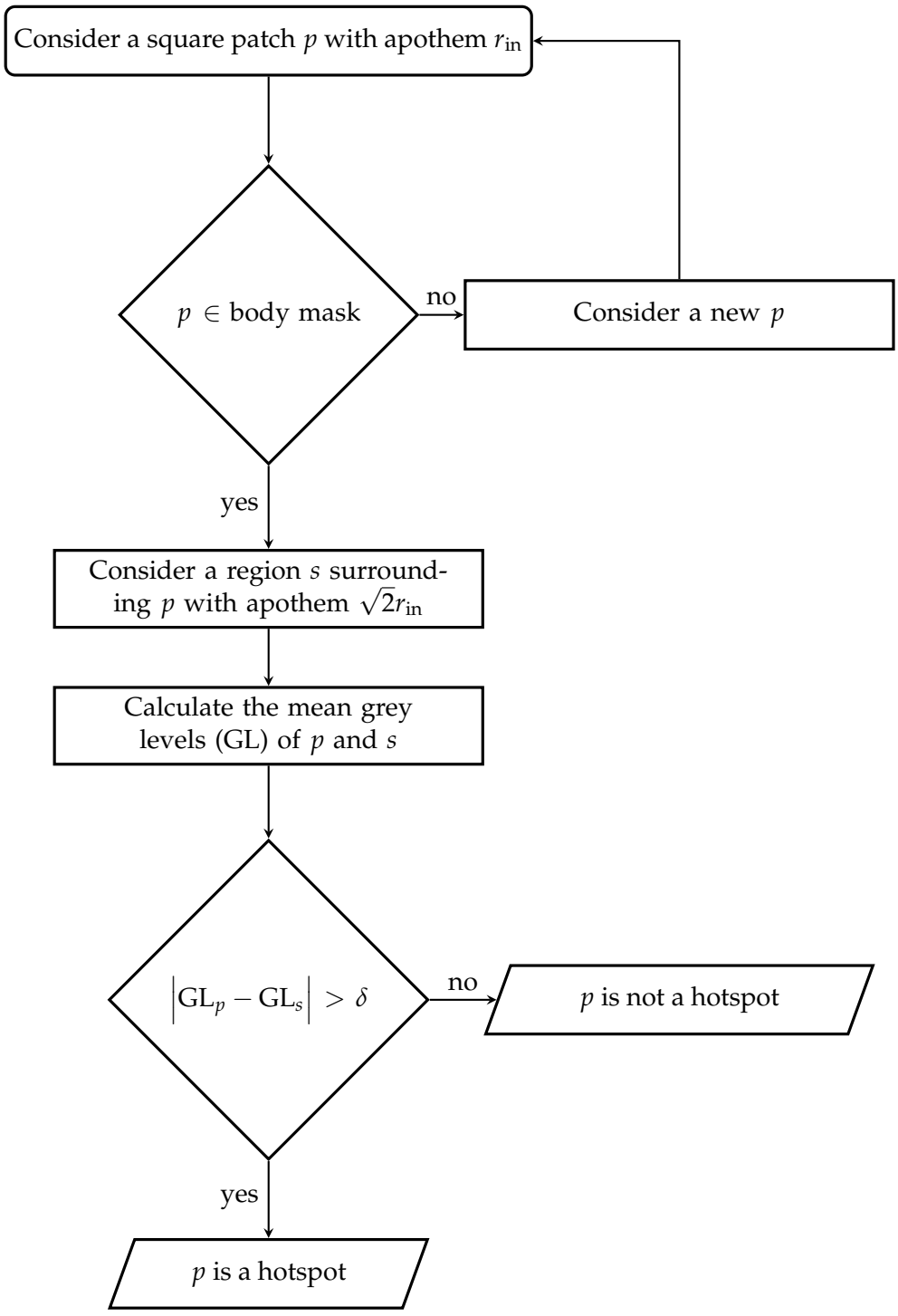


FIGURE 4.3: Flowchart of the algorithm used to detect the hotspots.

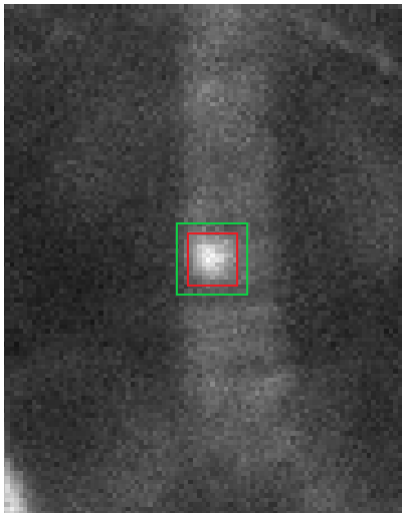


FIGURE 4.4: Illustrative image of the patches in the algorithm for hotspot detection: the red region represents the inner patch and the green region the outer patch

guarantee that no metastasis was left out, but would come with the cost of having a lot more of false-positives being detected. The final threshold was chosen trying to get as few false-positives as possible, while at the same time following the mandatory condition: all malignant hotspots from all the scintigraphy images in the database had to be detected. The algorithm returns a binary mask M with the same size as the original image, with the pixels belonging to a hotspot with a value of 1, and the pixels being part of the background with a value of 0 (see Figure 4.5). The 1-valued regions are called objects, connected components, or blobs.

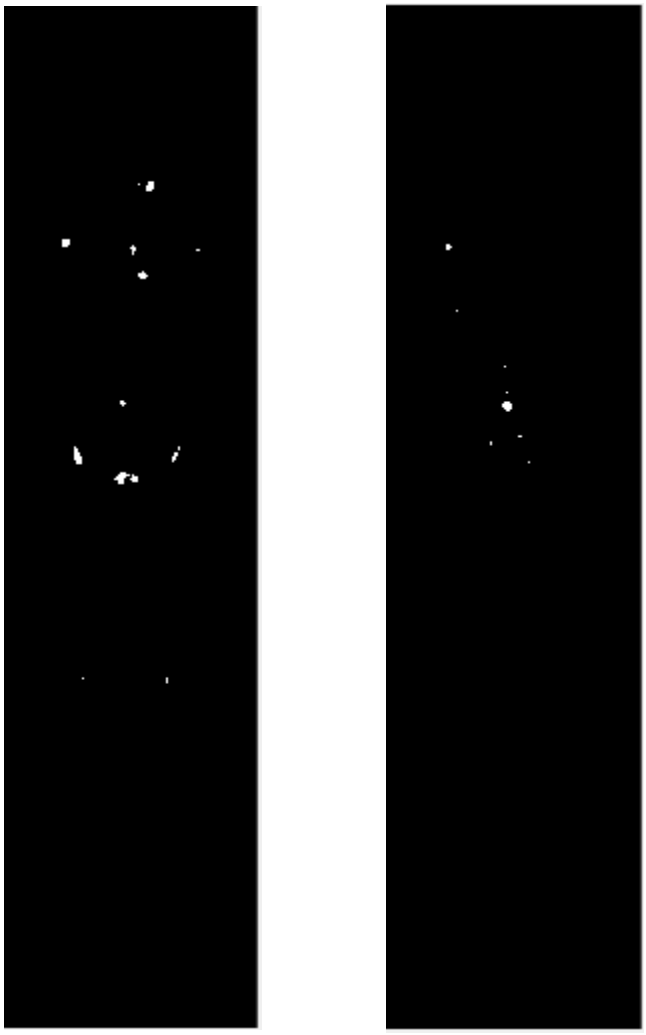


FIGURE 4.5: Example of the output of the detection algorithm: it returns a binary mask where the 1-valued pixels represent the hotspots.

4.3 False-positive attenuation

As expected, a considerable amount of hotspots not related to bone metastases was detected with previous algorithm. These hotspots could be due to some kind of benign bone condition or could just be due to normal and healthy physiological processes. Because the patient condition is determined through the assessment of the malignant bone lesions, the number of false-positive detections should be reduced. This was achieved through two methods:

1. By using image analysis techniques to eliminate regions that were known *a priori* to be non-malignant hotspots;

2. Trough the development of a classification algorithm to distinguish malignant from non-malignant hotspots.

The above steps to perform the first method are described in Sections 4.3.1 and 4.3.2, while the ones to perform the second method are described in Sections 4.3.3 and 4.3.4.

Method 1: image analysis techniques

In the first method for false-positive attenuation, image analysis techniques were used to eliminate regions that were known *a priori* to be non-malignant hotspots. This required two steps: the segmentation of the bone scans (Section 4.3.1) and the development of algorithms to identify the non-malignant hotspots (Section 4.3.2).

4.3.1 Anatomical segmentation

A method for the anatomical segmentation of the bone scans was developed. An accurate segmentation of bone scintigraphy images will allow the automatic localisation of the hotspots, which is be essential for this first method for false-positives removal (Section 4.3.2) and for calculating the final imaging biomarker, as it requires knowledge about the anatomical regions where the bone lesions are located (Section 4.4). In this work, this was achieved trough an atlas-based segmentation, by following the now described steps:

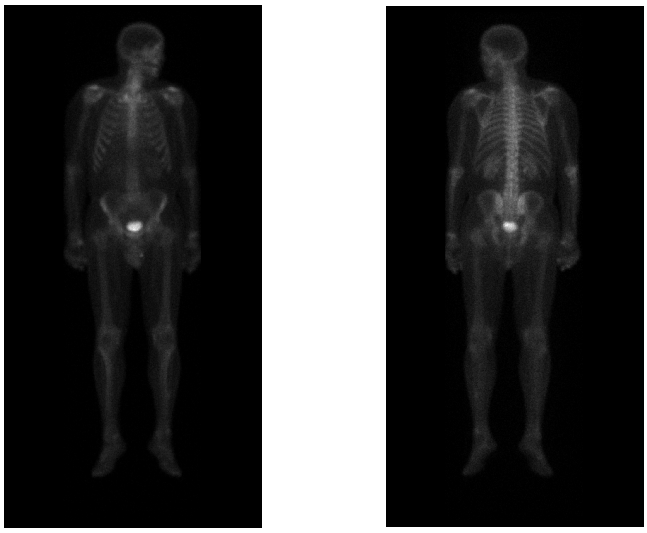
1. ***Create the atlas***. The first step to perform an atlas-based segmentation is to create the atlas, which will serve as a reference image. For this purpose, a bone scan from the database was selected (see Figure 4.6) and different anatomical regions of interest were manually drawn and labelled using MATLAB Ground Truth Labeler App. In the end, a *groundtruth* object was obtained, with the ground truth labels corresponding to the different anatomical regions. The final AP and PA atlases are shown in Figure 4.7. Notice how a 180° rotation over the vertical axis was applied to the PA view, so that the left and right side of AP and PA views would match. The ROIs into which the atlas was segmented into were based on the ones used in similar works where atlas-based segmentation was also performed in bone scintigraphy images, in particular the ones used in Huang et al. (2007) and Brown et al. (2018). These regions were:

- head
- left shoulder
- right shoulder
- left arm
- right arm
- left hand
- right hand
- sternum/spine
- rib cage
- kidneys
- pelvis
- bladder
- left femur
- right femur
- lower left leg
- lower right leg
- left foot
- right foot

2. ***Register the atlas with the target image***. Before propagating the labels from the atlas to a new bone scan (target image), it is important to register both images. The bone scans in the database present a certain degree of variability between each other due to differences in patient anatomy, imaging equipment, acquisition angle or date. The registration is therefore a key step as it will geometrically align the target bone scan with the one used as reference, so that they overlap as much as possible and can be compared. The function *imregtform* from MATLAB was first used to estimate a non-reflective similarity transformation to align the target with the atlas image. A non-reflective similarity transformation aligns the moving image (target) to the fixed image (atlas) through translation, rotation and scale operation. The scaling operation was particularly important as the images had two different sizes (1024×256 or 1024×512). Then, the function *imwarp* was used to transform the target image according to the geometric transformation output from *imregtform*. This process is illustrated in Figure 4.8.

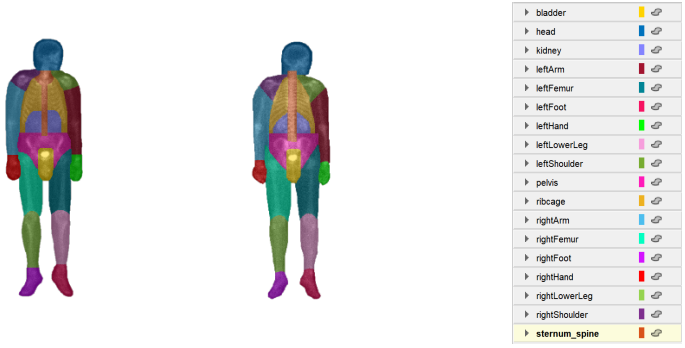
3. ***Propagate the labels.*** Having the labelled atlas and the target image aligned, it was possible to determine the anatomical region of the hotspots detected in the target image, by developing an algorithm that would analyse the area of each hotspot in each different segmented body region. To do that, and for each detected hotspot in the target image, the following steps were performed:

- A masked image of the hotspot, with the same size of the target image, was created;
- By looping trough all the anatomical regions, the percentage of hotspot pixels inside each region was calculated;
- The anatomical region that presented the highest percentage of pixels from the current hotspot was assigned as the region it belonged to.



AP PA

FIGURE 4.6: AP and PA scans used as reference to create the atlas



AP PA

FIGURE 4.7: Atlas for anatomical segmentation of the bone scans and respective labels. The reference scans of Figure 4.6 were used to manually label an atlas into the anatomical regions of interest.

4.3.2 Removal of hotspots with image analysis techniques

There are hotspots that, due to some specific characteristic that they present, can be easily identified as false-positives. Using solely image analysis techniques, they can be detected

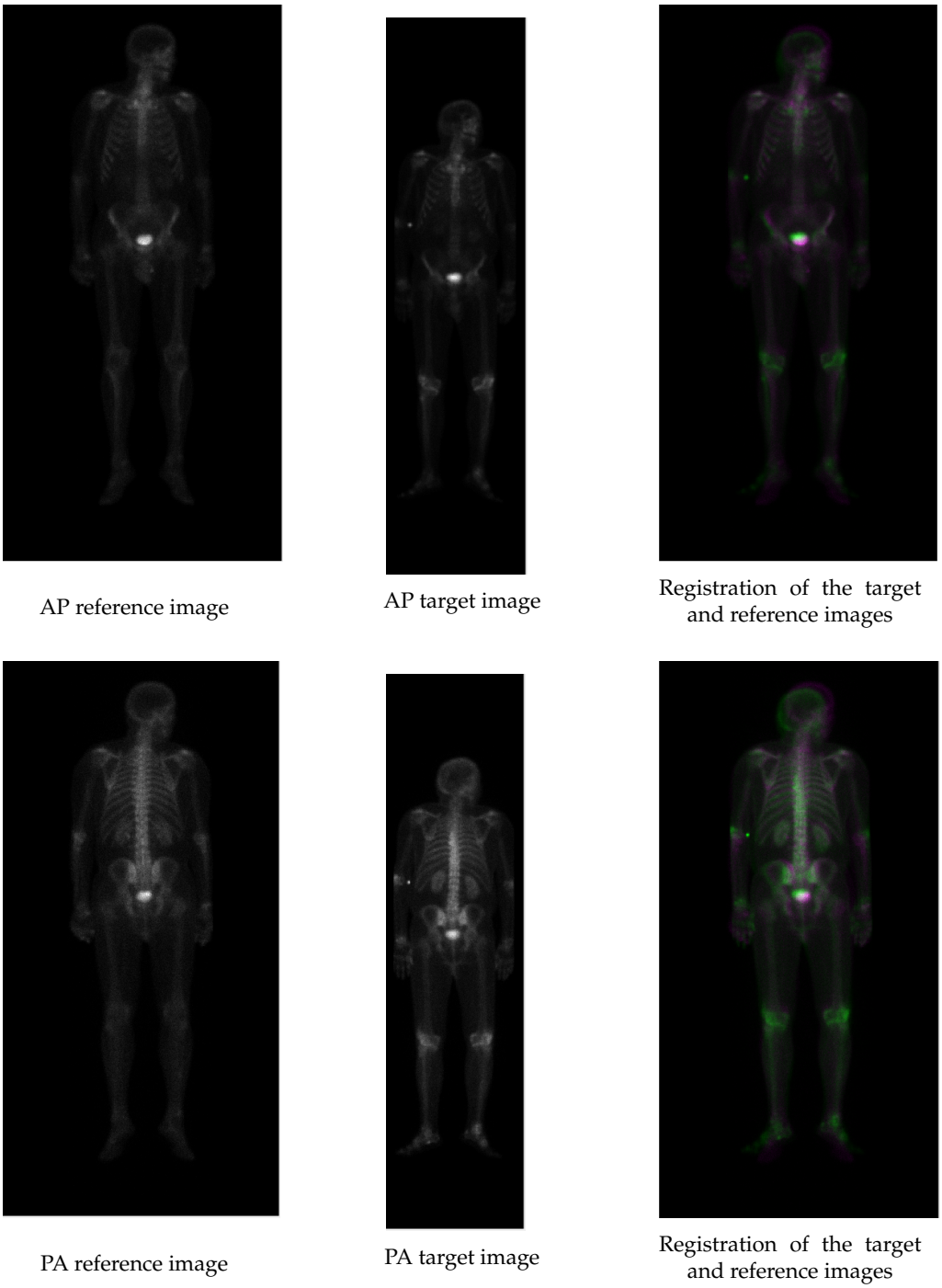


FIGURE 4.8: Illustration of the registration process for the AP (top row) and PA (bottom row) views of a patient. In each row, three images are shown (from left to right): the AP/PA image used as reference, the AP/PA image that one wants to register (target image) and the registration of both the reference and target images. In the registered image, regions where the pixels from the target image are brighter than the reference image appear as green; regions where the pixels from the target image are darker than the reference image appear as magenta; dark regions correspond to areas where the pixels from both the target and reference images are dark, and finally regions where he pixels from both the target and reference images are bright appear as grey or white.

and removed. These include:

- ***Hotspots found in certain anatomical regions***. There are certain anatomical regions where false-positive hotspots are commonly detected. For example, increased radiotracer uptake is common in urine, and therefore a noticeable hotspot in the bladder is almost always seen. Another common place for a hotspot to appear is in the hand, as this is usually the place trough which the radioisotope is injected. Hotspots that were detected outside of the body were also removed, as they corresponded to urine-collection bags. To remove these hotspots, the algorithm developed in Section 4.3.1 was applied to obtain the anatomical region the hotspots belonged to, and the ones belonging to the bladder or hands were removed (Figure 4.9a).

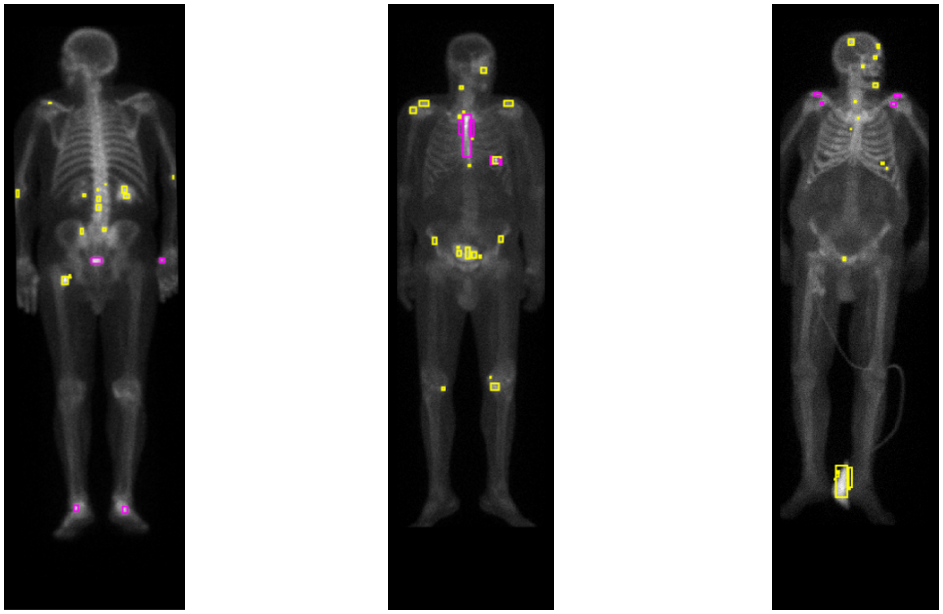
- ***Symmetrical hotspots***: The appearance of symmetrical hotspots in bone scans is very common, and is usually related to normal physiological processes. They are usually found in places like the shoulders and knees, and can also be removed (Figure 4.9c). To find them, an algorithm to detect the symmetry axis of a patient in a bone scintigraphy image was initially used. The code used for the identification of the symmetry axis was developed by Cicconet et al. (2017) and is available at GitHub. For two hotspots to be considered symmetrical, the following conditions had to be verified:

  - The absolute difference between the perpendicular distance from the hotspot centroids to the symmetry axis could not exceed a certain threshold, T_{dist};
  - The hotspots must lay on opposites side of the axis;
  - The y-coordinate of the hotspot's centroids could not exceed a certain threshold, T_y;
  - The absolute difference between the areas of the hotspots could not exceed a certain threshold, T_{area}.

The following values were used:

- $T_{dist} = 7.5$ pixels
- $T_y = 5$ pixels

- T_{area} = 30% of the area (in pixels) of one of the hotspots



(A) (B) (C)

FIGURE 4.9: Example of hotspots that can be removed using image analysis techniques. Pink regions in image A represent hotspots that can be removed as they are located in the bladder, feet and hand; pink regions in image B represent hotspots that can be removed due to their high aspect ratio; pink regions in image C represent hotspots that can be removed due to their symmetry.

Method 2: Classification algorithm

The methodology for the second method used for attenuation of false-positives is now described. The development of an algorithm for hotspots classification involves two steps: extraction of features from the detected hotspots (Section 4.3.3) and training a classifier with the extracted features (Section 4.3.4). The algorithm is then evaluated using suitable performance metrics (Section 4.3.5). Several algorithms were trained, and the one with the best performance was chosen to classify the hotspots.

4.3.3 Feature extraction

The feature extraction stage is used to obtain the features that will serve as input to a classification algorithm. Two types of features were extracted: handcrafted low-level features (shape and intensity) and learned high-level features.

4.3.3.1 Shape and Intensity features

The first type of features to be extracted were handcrafted features. For that purpose, the MATLAB function *regionprops* was used. To use this function, the masked image M (Figure 4.5) was first transformed into a label matrix L, in which the 8-connected objects were labelled with unique integer values (see Figure 4.10). In this case, each distinct object corresponded to a detected hotspot. The matrix L was used as input to the *regionprops* function, and the shape and intensity features of each hotspot were returned as a $n \times m$ table, n being the number of objects in L and m the number of properties calculated for each region. In total, 16 shape features and 4 intensity features were extracted from each hotspot. Of the shape features, 15 were built-in in *regionprops* function*, and one was manually added (the ratio between the major and the minor axis). The properties that were extracted for each hotspot are described in Table 4.4.

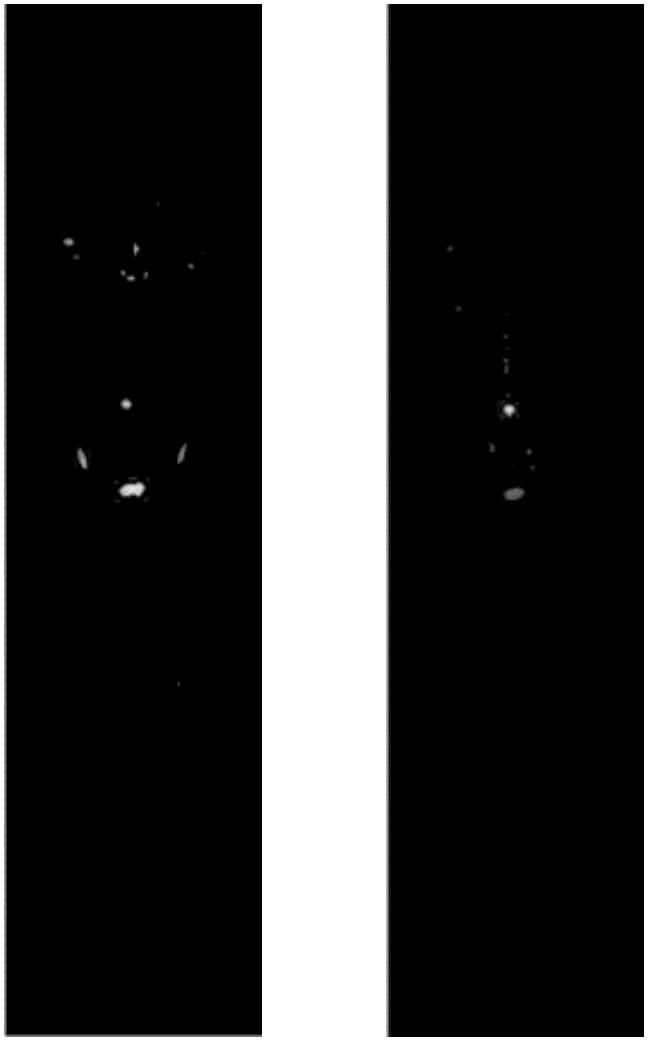
4.3.3.2 Learned features

The second type of features to be obtained were high-level features extracted with the convolutional base of a pre-trained CNN. The proposed methodology for this task is now thoroughly presented.

In image processing, histogram matching or histogram specification is the transformation of an image so that its histogram matches a specified histogram

1. ***Extraction of the hotspots***: The first step consists of creating an image datastore with all the hotspots detected with the algorithm described in Section 4.2. For the sake of homogeneity, and because different scans presented different levels of brightness, an histogram matching is applied to every image so that their histogram matches the histogram of a specific scan. After, the detection algorithm is executed and the original image is cropped according to the position and dimensions specified by the bounding box (see Table 4.4) of each detected region. Each patch was then converted to a normalised array of double values in the range [0,1], and the resultant matrix was multiplied by 255 so that each hotspot was converted into a 8-bit image. Because the CNN that is going to be used requires input images of size $\mathtt{n} \times \mathtt{n} \times 3$, each patch was also converted into RGB by replicating the grey image in

*One of these features, the *circularity*, was instead used as an inverse circularity, as hotspots with area equal to unity are considered to have null perimeter, which would result in an undefined ratio (circularity = $\frac{4*\text{Area}\cdot\pi}{\text{Perimeter}^2}$)



AP PA

FIGURE 4.10: Example of the matrices L for the AP and PA views: each 8-connected object corresponds to a detected hotspot and is labelled with a unique integer value.

each colour channel. Finally, each hotspot was saved as a png image in a folder that in the end would comprise all the hotspots detected in all the bone scans from the database.

2. ***Extraction of the features***: To extract deep features form the hotspots, the folder created in the previous step was loaded to the MATLAB workspace as an image datastore. The next step involves resizing all the images in the datastore so that they have the input size required by the network in question. This is done by creating an augmented image datastore, specifying the desired image size. In this work, the pre-trained network used was *ResNet18*, which required input images of size $224 \times 224 \times 3$. To finally obtain the features, the MATLAB function *activations* is

TABLE 4.4: Name and description of the handcrafted features. Top part of the table corresponds to Shape Measurements and bottom half to Pixel Value Measurements.

Property	Description
Area	No of pixels in the region
AxisLengthRatio	Ratio between *MajoraxisLength* and *MinoraxisLength*
BoundingBox	Position and size of the smallest box containing the region
Centroid	Center of mass of the region
ConvexArea	Number of pixels in *ConvexImage*[1]
Eccentricity	Eccentricity of the ellipse ϵ[2]
EquivDiameter	Diameter of a circle with the same area as the region
EulerNumber	No of objects in the region minus the no of holes in those objects
Extent	Ratio of pixels in the region to pixels in the total bounding box
FilledArea	Number of on pixels in *FilledImage*[3]
InvCircularity	Inverse of the circularity[4] of the object
MajoraxisLength	Length (in pixels) of the major axis of $\hat{\epsilon}$[5]
MinoraxisLength	Length (in pixels) of the minor axis of $\hat{\epsilon}$
Orientation	Angle between the x-axis and the major axis of $\hat{\epsilon}$
Perimeter	Distance around the boundary of the region
Solidity	Proportion of the pixels in the convex hull that are also in the region
MaxIntensity	Value of the pixel with the greatest intensity in the region
MeanIntensity	Mean of all the intensity values in the region
MinIntensity	Value of the pixel with the lowest intensity in the region
WeightedCentroid	Center of the region based on location and intensity value

[1] *ConvexImage*: Image that specifies the *ConvexHull*[6], with all pixels within the hull filled in (binary image)
[2] ϵ: ellipse that has the same second-moments as the region
[3] *FilledImage* Image the same size as the bounding box of the region, returned as a binary
[4] The circularity of an object is defined as $\frac{4 \cdot \text{Area} \cdot \pi}{\text{Perimeter}^2}$
[5] $\hat{\epsilon}$: ellipse that has the same normalized second central moments as the region
[6] *ConvexHull*: Smallest convex polygon that can contain the region

used, receiving as input the augmented datastore along with the chosen network and layer one wants to extract the features from. This function returns the activations, that is, the output of the specified layer, as an $n \times m$ numerical array, with n being the number of images in the datastore and m the number of output elements (features) from the layer. The layer used was *pool5*, which returned 512 features per hotspot. A schematic representation of the architecture of the ResNet18 network is given in Figure 4.11.

FIGURE 4.11: Diagram of ResNet18 with highlighted "pool5" layer.

4.3.4 Classifiers

The extracted features will serve as input variables to a classifier that should be able to learn which features are characteristic of malignant hotspots (bone metastases) and which features are more associated with healthy and benign hotspots (false-positives).

The biggest challenge that was faced was due to the fact that the extracted hotspots had no labels, which precluded us from using a supervised learning algorithm. Three different approaches were tried out: an unsupervised learning algorithm (Section 4.3.4.1), a semi-supervised learning algorithm (Section 4.3.4.2) and a semi-supervised strategy (Section 4.3.4.3).

4.3.4.1 K-means Clustering

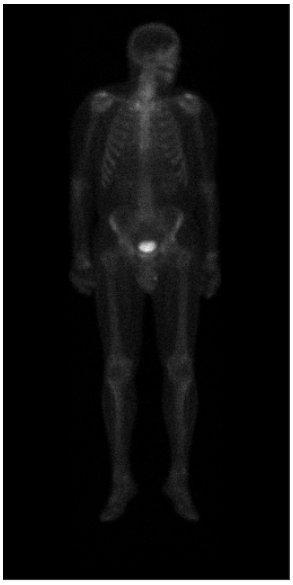
In the first approach, an unsupervised learning algorithm was used. Two separated k-means clustering algorithms, one with two clusters and another with three clusters, were initially applied to the training set. When choosing three clusters, it was hoped that the data could be partitioned into a cluster of healthy data, a cluster of benign data and a cluster of malignant data (that is, metastases). When choosing two clusters, it was hoped

that the data could be partitioned into a cluster of non-malignant data (healthy and benign hotspots) and a cluster of malignant data. For each algorithm (2-class and 3-class), a model for the classification of new data was built, by assuming that each final cluster represented a class and by assigning each hotspot from the test set to the nearest cluster centroid. The distance metric used for defining the initial clusters, as well as to assign new data to these clusters, was the square euclidean distance.

The use of such an algorithm for the classification purpose has the disadvantage that one does not know the category that each cluster represents. Here, this was chosen by analysing the number of data points that were assigned to each cluster.

4.3.4.2 One-Class Classification

Given the available data, there was a semi-supervised approach that could be applied to this problem: the one-class classification algorithm described in Section 2.5.3.1. Examining Table 4.3, one can create two distinct image data sets: one that contains non-malignant hotspots and one that contains malignant hotspots. While the non-malignant set does not contain any metastases, the malignant one contains hotspots that can or not be metastases. The idea is therefore to train a classifier on a training set containing only non-malignant hotspots, hoping that the it can learn the features that characterise them, and later identify the outliers (metastases) in a set containing both non-malignant and malignant hotspots. Figure 4.12 shows two bone scans belonging to each of the data sets. The methodology for feature extraction was performed for both data sets independently, and an algorithm for one-class classification was trained with the features extracted from the healthy data set. The final classifier was obtained with the MATLAB function *fitcsvm*, which used the OC-SVM algorithm proposed by Schölkopf et al. (2000), described in Section 2.5.3.1. The OC-SVM was trained with an outlier fraction of 5%, a Gaussian kernel function with a Kernel scale parameter of 1.81 and a Sequential Minimal Optimisation (SMO) as an optimisation routine. Ideally, the SVM classifier has learnt a boundary that can separate the non-malignant samples from the malignant samples. The features extracted from the data set containing non-malignat and malignant hotspots is fed into the trained model, which classifies each entry as non-malignant or as an outlier. If a sample is classified as an outlier it means that it does not belong to the non-malignant set of hotspots and therefore it is considered a bone metastasis.



AP healthy

PA healthy

AP malignant

PA malignant

FIGURE 4.12: Example of two scans belonging to the two different classes used in the OCC algorithm. The top images are the AP and PA views from a patient with no bone lesions: this means that every detected hotspot can be considered a non-malignant hotspot. The bottom images are the AP and PA views from a patient with bone metastases: the detected hotspots can either be non-malignant or malignant.

4.3.4.3 Iterative Algorithm

A different methodology for the classification task was developed afterwards. Two variations of this algorithm were developed: a two class classification algorithm, which assumed an initial division of the data according to Table 4.3, and a three class classification algorithm, which assumed an initial division of the data according to Table 4.2. This algorithm was named *hotBSI*.

Two class algorithm

The first step in this algorithm was to execute the detection and features extraction algorithms in all the bone scans, and label each detected hotspot as "0" (non-malignant) or "1" (malignant) according to the category of the bone scan the hotspots were extracted from (see Table 4.3). Then, an initial two class classifier, C_0, was trained to distinguish between non-malignant and malignant lesions. It should be pointed out that this classifier is trained under a lot of noise, as it was assumed that every hotspot detected in a bone scan belonging to a certain category also belonged to that same category, which is not true (for example, every hotspot detected in the patients belonging to the malignant category was labelled as "1", when in reality they could belong to the non-malignant class). The next stage involves an iterative process through the following steps:

1. The last trained classifier, C_{i-1}, is used to classify the detections on the scans belonging to the malignant class. For each detected region, the classifier returns the likelihood that the region comes from the *malignant* or *non-malignant* class;

2. For each patient in the malignant category:

   (a) The detection with the highest likelihood of being malignant is selected;

   (b) All other detections with likelihood of being malignant *higher than a pre-determined threshold* (if any) are also selected.

3. A new training data set is created, so that detections made on non-malignant scans are considered as false-positives (and labelled as 0) and the above selected regions are considered as true-positives (or malignant hotspots, labelled as 1);

4. Train a new classifier C_i with the new training data set.

A schematic description of the binary hotBSI is given in Algorithm 1 and Figure 4.13. The value of the threshold was set to 0.8.

Algorithm 1 hotBSI algorithm

Inputs:
NM - feature set from all the hotspots extracted from the non-malignant images
M - feature set from all the hotspots extracted from the malignant images
T - threshold (default as 0.8)
NrIt - number of iterations (default as 100)

Output:
C - a classifier to classify new hotspots as non-malignant or malignant

1: Train an initial classifier, C_0, with the input features (NM $\cup$ M)
2: **for** i = 1:NrIt **do**
3: Empty M
4: **for** each patient in the *malignant* set **do**
5: Use C_{i-1} to predict the probabilities of the detections to be a metastasis (P_{met})
6: Identify the hotspot with the highest likelihood of being a metastasis ($P_{\max}$)
7: **for** d = 1 : number of detected hotspots for the current patient **do**
8: **if** $P_{\text{met}}(d) == P_{\max}$ || $P_{\text{met}}(d) > T$ **then**
9: Add the hotspot to M
10: Create a new training set, NM $\cup$ M
11: Train a new classifier C_i with the new training data set
12: **return** C_{NrIt}

Three class algorithm

A similar algorithm to the just now described was developed, but with three classes instead of two. This was done by splitting the non-malignant class into a healthy and a benign class. The first step in this algorithm was to perform the detection and features extraction algorithms in all the bone scans, and label each detected hotspot as "0" (healthy), "1" (benign) or "2" (malignant), according to the category that the bone scan where the hotspots were detected belonged to (see Table 4.2). Then, an initial three class classifier, C_0, was trained to distinguish between healthy, benign and malignant lesions. Once again, this classifier is trained under a lot of noise, as it was assumed that every hotspot detected in a bone scan belonging to a certain category also belonged to that same category, which is not true (for example, every hotspot detected in the patients belonging to the malignant category were labelled as "2", when in reality they could belong to any of the three classes). The next stage involves an iterative process through the following steps:

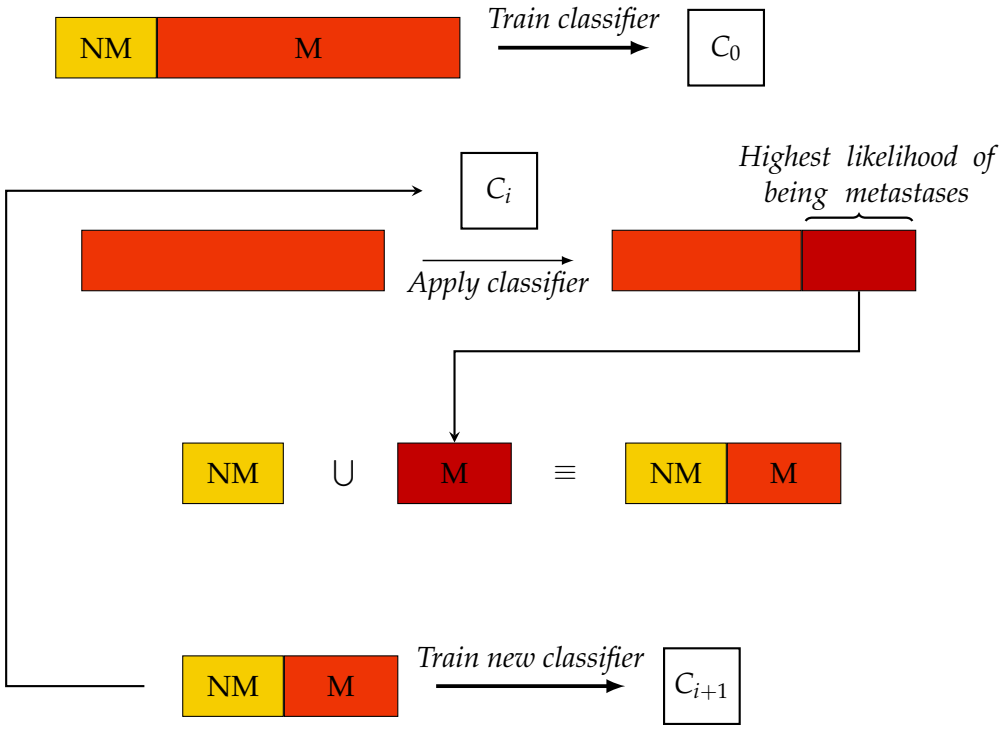


FIGURE 4.13: hotBSI algorithm for 2 classes. NM stands for detections labelled as non-malignant, while M stands for detections labelled for training in a given iteration as malignant.

1. The last trained classifier, C_{i-1}, is used to classify the detections on the scans belonging to the benign and malignant class. For each detected region, the classifier returns the likelihood that the region comes from the *healthy*, *benign* and *malignant* class;

2. For each patient in the benign/malignant categories:

   (a) The detection with the highest likelihood of being benign/malignant is selected;

   (b) All other detections with likelihood of being benign/malignant *higher than a pre-determined threshold* (if any) are also selected.

3. A new training data set is created, so that detections made on healthy scans are considered as healthy (and labelled as 0) and the above selected regions are considered as benign/malignant (labelled as 1/2);

4. Train a new classifier C_i with the new training data set.

A schematic description of the three class hotBSI is given in Algorithm 2 and Figure 4.14.

Algorithm 2 hotBSI algorithm

Inputs:
H - feature set from all the hotspots extracted from the healthy images
B - feature set from all the hotspots extracted from the benign images
M - feature set from all the hotspots extracted from the malignant images
T - threshold (default as 0.8)
NrIt - number of iterations (default as 100)

Output:
C - a classifier to classify new hotspots as healthy, benign or malignant

Train an initial classifier, C_0, with the input features (H $\cup$ B $\cup$ M)
for i = 1:NrIt **do**
    Empty B
    Empty M
    for each patient in the *benign* set **do**
        Use C_{i-1} to predict the probabilities of the detections to be benign lesion (P_{ben})
        Identify the hotspot with the highest likelihood of being benign (P_{Bmax})
        for d = 1 : number of detected hotspots for the current patient **do**
            if $P_{\text{ben}}(d) == P_{\text{Bmax}} \,||\, P_{\text{ben}}(d) > T$ **then**
                Add the hotspot to B
    for each patient in the *malignant* set **do**
        Use C_{i-1} to predict the probabilities of the detections to be a metastasis (P_{met})
        Identify the hotspot with the highest likelihood of being a metastasis (P_{Mmax})
        for d = 1 : number of detected hotspots for the current patient **do**
            if $P_{\text{met}}(d) == P_{\text{Mmax}} \,||\, P_{\text{met}}(d) > T$ **then**
                Add the hotspot to M
    Create a new training set, H $\cup$ B $\cup$ M
    Train a new classifier C_i with the new training data set
return C_{NrIt}

Learning Algorithms

The classifiers were trained using four different supervised learning algorithms:

(i) **hotBSI-SVM:** Support Vector Machine (SVM) trained with a linear kernel with scale 1, where the values obtained with the linear SVM score function (bias = 1.08) were transformed into posterior probabilities using the sigmoid function with slope -1.40 and intercept 0.06;

(ii) **hotBSI-KNN:** K-Nearest Neighbours (KNN), trained with five nearest neighbours with uniform weighting and the Euclidean distance function as the distance metric;

(iii) **hotBSI-DTs:** Decision Trees (DTs), trained with a minimum of 10 samples per branch node, a maximum number of splits equal to the number of samples minus one and

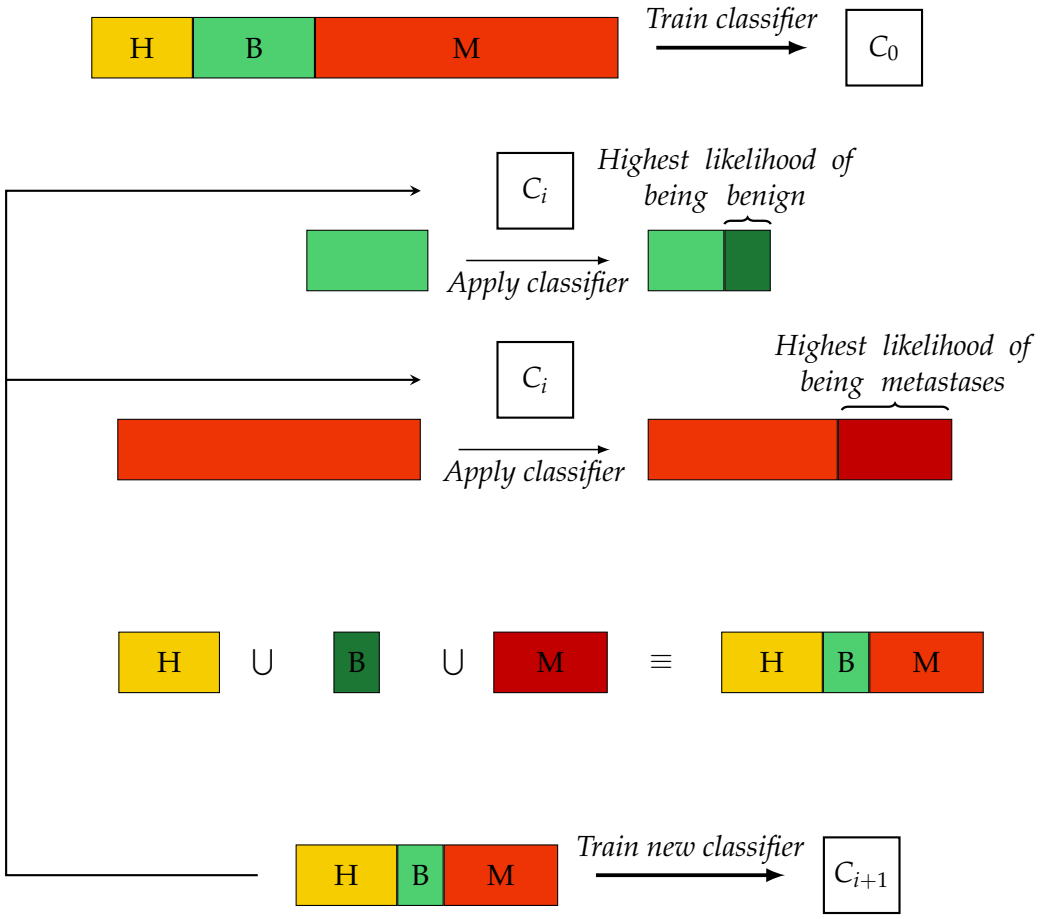


FIGURE 4.14: hotBSI algorithm for 3 classes. H stands for detections labelled as healthy, B for detections labelled as benign and M for detections labelled as malignant.

the Gini's diversity index as the split criterion;

(iv) **hotBSI-LDA:** Linear Discriminant Analysis (LDA) with both 'Delta' (linear coefficient threshold) and 'Gamma' (amount of regularisation) equal to 0.

Threshold and stopping criteria

The algorithm runs during a predetermined number of iterations (set as 100 in the current experiments). Other stopping criteria will be pursued in the future, for example, running the algorithm until the non-malignant and malignant sets (or the healthy, benign and malignant sets, in the case of the 3 class algorithm) remain unchanged between iterations.

The value of the threshold was empirically set to 0.8. A sensitivity analysis of the impact of this parameter in the results is left for future work.

4.3.5 Evaluation methodology

The the test set was manually labelled by identifying the malignant, benign and healthy hotspots. For evaluating the 3-class algorithms, this manual classification remained unchanged. When working with a two-class problem, metastases were considered true detections (malignant/positive class), while healthy and benign hotspots were considered false-positive detections (non-malignant/negative class).

The algorithms were evaluated using common performance metrics such as sensitivity, specificity, accuracy, precision, false positive rate (FPR), false-positive detections per image (FPPI),F1-score and AUC (area under the ROC curve). In addiction, the false negative rate (FNR) is also calculated, as it was considered that a low FNR was of special importance for this particular classifier. Since the goal of this algorithm is to permit its use in the clinical practise to aid physicians in the diagnosis and follow-up of patients with metastatic cancer, it is important that the final algorithm has a FNR as low as possible. A high FNR would mean that the algorithm was classifying a lot of malignant hotspots as non-malignant , which could be very dangerous to the patient, as it was failing to diagnose them with the disease and preventing them from having access to an early treatment.

4.4 BSI calculation

The main goal of this work is to develop an algorithm that can quantify a bone scan by assigning to it a number (imaging biomarker) that correlates with staging of disease, disease prognosis, and treatment efficiency. As seen in section 3.5, the most popular quantitative parameter to evaluate a bone scintigraphy image is the Bone Scan Index (BSI), and for that reason that was the imaging biomarker chosen for the final assessment of the bone scans. The calculation of the BSI requires three parameters: the area of the metastasis, the area of the skeletal region where the metastasis is found, and a coefficient C_R, which reflects the fractional weight of that same skeletal region when compared to the weight of the entire skeleton. Both the areas of the metastasis and the skeleton region were easily calculated using built-in MATLAB functions. The coefficients C_R required knowledge about the weight of the different skeleton regions as well as the total weight of the skeleton. These values were obtained using the reference values for total skeletal weight and weight of different skeletal bones found in Silva et al. (2009), which are based in a Portuguese Identified Skeletal Collection (ISC). Although data for both the female and male

sexes with ages above 28 was available, the C_R coefficients in this thesis were calculated using the weight values obtained from a sample of 10 men with over 60 years old. The paper provides the mass of 18 skeleton regions (see Table 4.5). The mass of different skeleton regions was added up to obtain the approximated mass of the ROIs into which the atlas in Section 4.3.1 was divided into (Table 4.6).

Finally, the coefficients C_R for each region of the atlas were obtained by dividing the mass of each region by the total mass of the skeleton. These values are presented in Table 4.7.

TABLE 4.5: Mass of different skeleton regions obtained from an ISC sample of 10 men over 60 years old

Region	Mass (kg)
Skull	0.644
Mandible	0.060
Humerus	0.262
Radius	0.079
Ulna	0.106
Femur	0.754
Tibiae	0.445
Fibula	0.085
Hand	0.104
Foot	0.228
Clavicle	0.038
Scapula	0.108
Coxae	0.307
Patella	0.025
Ribs	0.204
Vertebrae	0.316
Sacrum	0.079
Sternum	0.019
Total	**3.863**

TABLE 4.6: Skeleton regions (from Table 4.5) used to obtain the mass of the ROIs of the atlas

ROI from the atlas of Section 4.3.1	Skeleton regions from Table 4.5
Right/left arm	humerus + radius+ ulna
Right/left femur	femur
Right/left foot	foot
Right/left hand	hand
Head	skull + mandible
Right/left lower leg	tibiae + fibula
Pelvis	coxae + sacrum
Ribcage	ribs
Sternum/spine	vertebrae + sternum
Right/left shoulder	scapula + clavicle

TABLE 4.7: Mass and fractional mass of the atlas regions. The fraction mass correspond to the coefficient C_R, which will be used in the BSI calculation

Atlas Region	Mass (kg)	Fractional Mass (C_R) (%)
Right/left arm	0.447	11.57
Right/left femur	0.754	19.52
Right/left foot	0.228	5.90
Right/left hand	0.104	2.69
Head	0.704	18.22
Right/left lower leg	0.555	14.37
Pelvis	0.386	9.99
Ribs	0.204	5.28
Right/left shoulder	0.146	3.78
Sternum_spine	0.335	8.67
Total	**3.863**	**100.00**

Chapter 5

Results

In this chapter, the results of the methods described in Chapter 4 are presented. The performance of the detection algorithm when applied to real bone scintigraphy images is firstly shown in Section 5.1, followed by the results of the two methods developed for false-positives attenuation in Section 5.2. In section 5.3, a qualitative evaluation of the Bone Scan Index is made using two patients from the database.

5.1 Detection

The algorithm described in Section 4.2 successfully detected all the hotspots corresponding to metastases (see Table 5.1). Figures 5.1 and 5.2 show the detection algorithm results when it is applied to bone scintigraphy images from the non-malignant and the malignant set, respectively. Comparing the results with the respective patient's medical reports, it can be concluded that the algorithm successfully detected the hotspots corresponding to metastases. On the other hand, it presents a very high rate of false-positive detections: approximately 73% of the detected hotspots were not metastases, corresponding to an average of 32 false-positive detections per image (see Table 5.1). Observing the figures, it can be seen that most of the detected hotspots are healthy or benign (that is, non-malignant), while only a small percentage are actually metastases.

TABLE 5.1: Results of the detection phase.

Sensitivity	Specificity	FPR	FNR	Precision	F1	FPPI
1.00	0.00	0.73	0.00	0.58	0.73	32

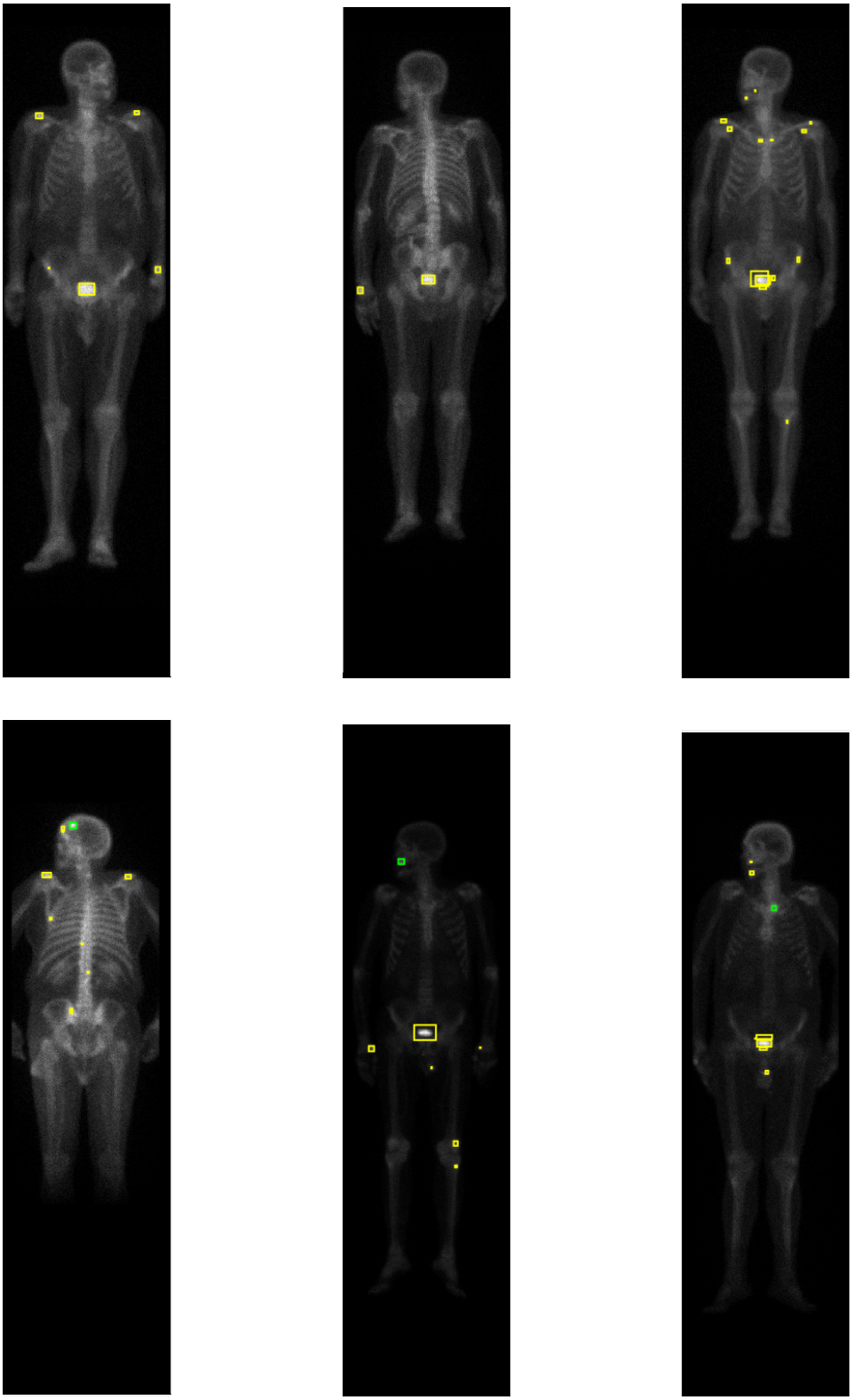


FIGURE 5.1: Results of the detection algorithm in bone scintigraphy images from the **non-malignant set**. The colours of the bounding boxes were manually chosen for the purpose of illustration, according to the medical report of the patient: red represents metastases, green represents benign bone lesions and yellow represents healthy hotspots (i.e., neither malignant nor benign lesions).

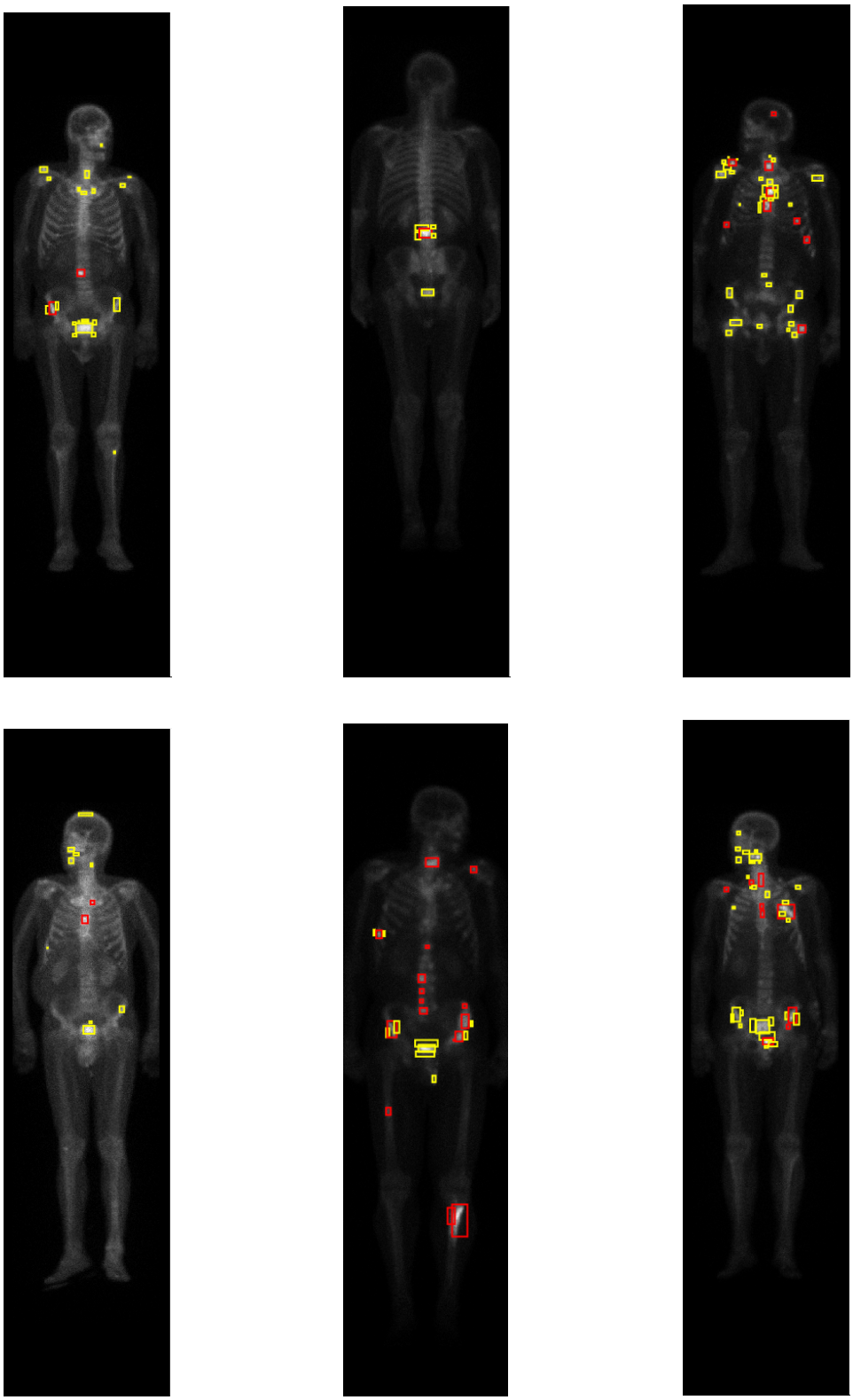


FIGURE 5.2: Results of the detection algorithm in bone scintigraphy images from the **malignant set**. The colours of the bounding boxes were manually chosen for the purpose of illustration, according to the respective medical report of the patient: red represents metastases, green represents benign bone lesions and yellow represents healthy hotspots (hotspots that are neither malignant nor benign lesions).

5.2 False-positive reduction

The results for the methods developed for false-positive reduction are now shown. Section 5.2.1 presents the results for the atlas segmentation, Section 5.2.2 for the removal of false-positives with image analysis techniques, and Section 5.2.3 for the classification algorithms.

5.2.1 Atlas Segmentation

The algorithm described in Section 4.3.1 was applied to each bone scan to identify the anatomical region of the detected hotspots. Figures 5.3, 5.4 and 5.5 show three examples of the anatomical labelling of hotspots detected in 3 different different patients.

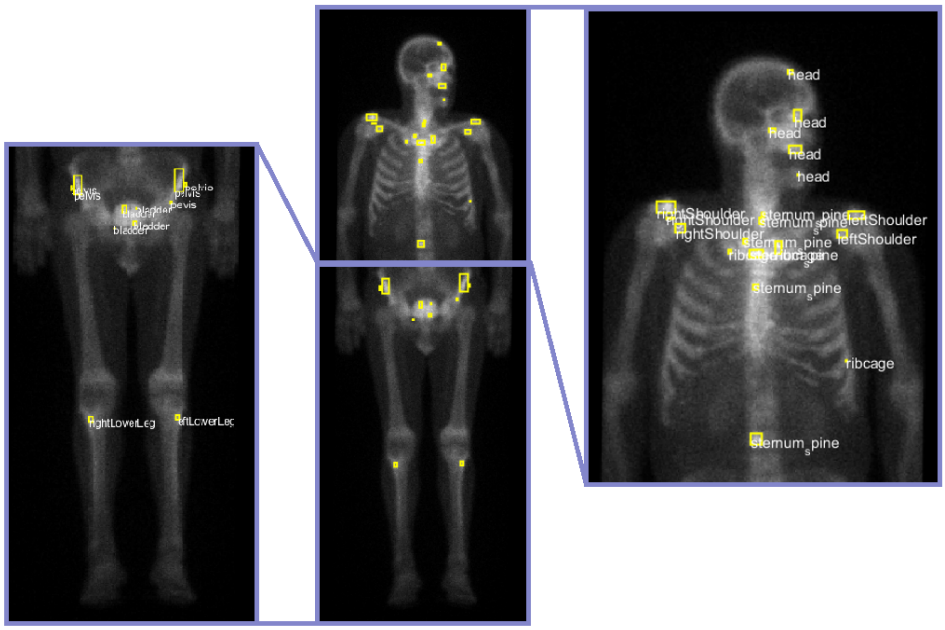


FIGURE 5.3: Results of the anatomical labelling (Example 1)

FIGURE 5.4: Results of the anatomical labelling (Example 2)

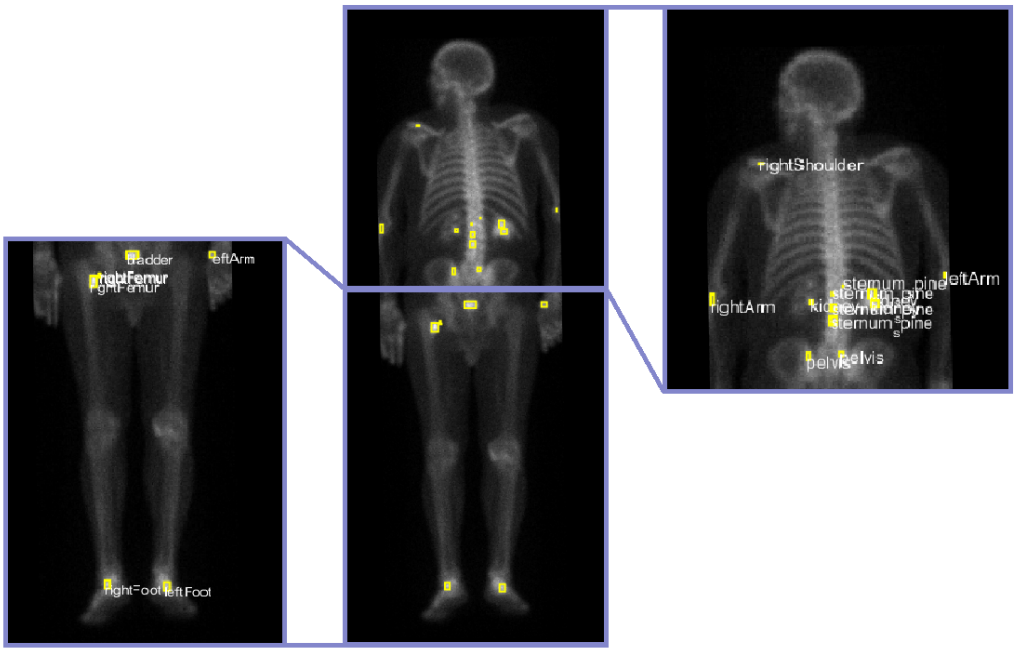


FIGURE 5.5: Results of the anatomical labelling (Example 3)

5.2.2 Attenuation of false-positives hotspots with image analysis techniques

Here, the results obtained following the methodology described in Section 4.3.2 are shown. The goal was to remove some false-positive hotspots using solely image analysis techniques. Two types of hotspots were removed: hotspots found in certain anatomical regions (Section 5.2.2.1) and symmetrical hotspots (Section 5.2.2.2).

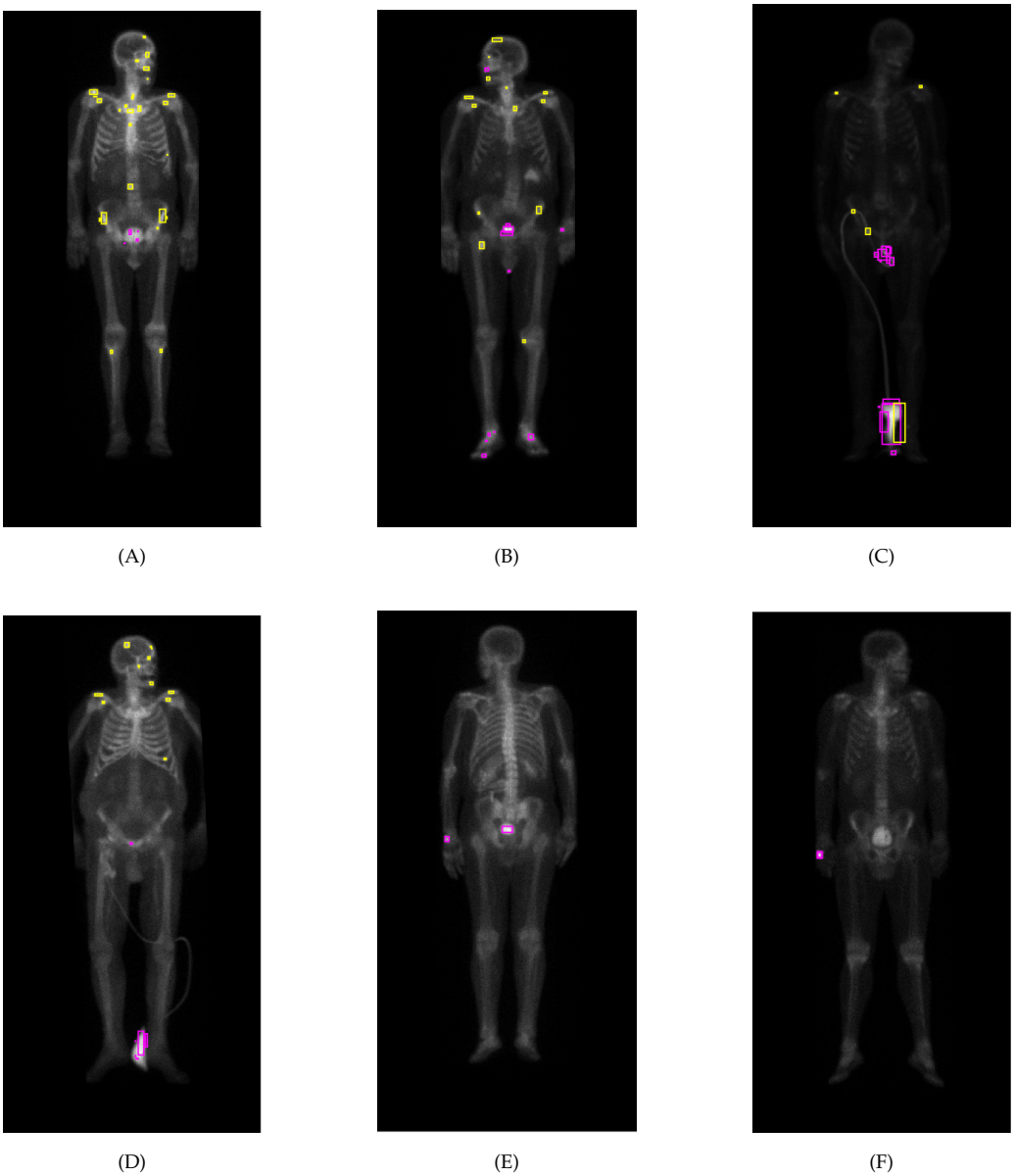
5.2.2.1 Hotspots found in certain anatomical regions

After applying the algorithm described in 4.3.1 to find the anatomical region of each detection, the hotspots located in the bladder, hands, feet or outside the body were removed. Figure 5.6 shows some results of this algorithm, when applied to bone scans from the data set.

5.2.2.2 Symmetrical hotspots

The algorithm developed for finding symmetrical detections was applied to the bone scans. Figure 5.7 shows some results of this application.

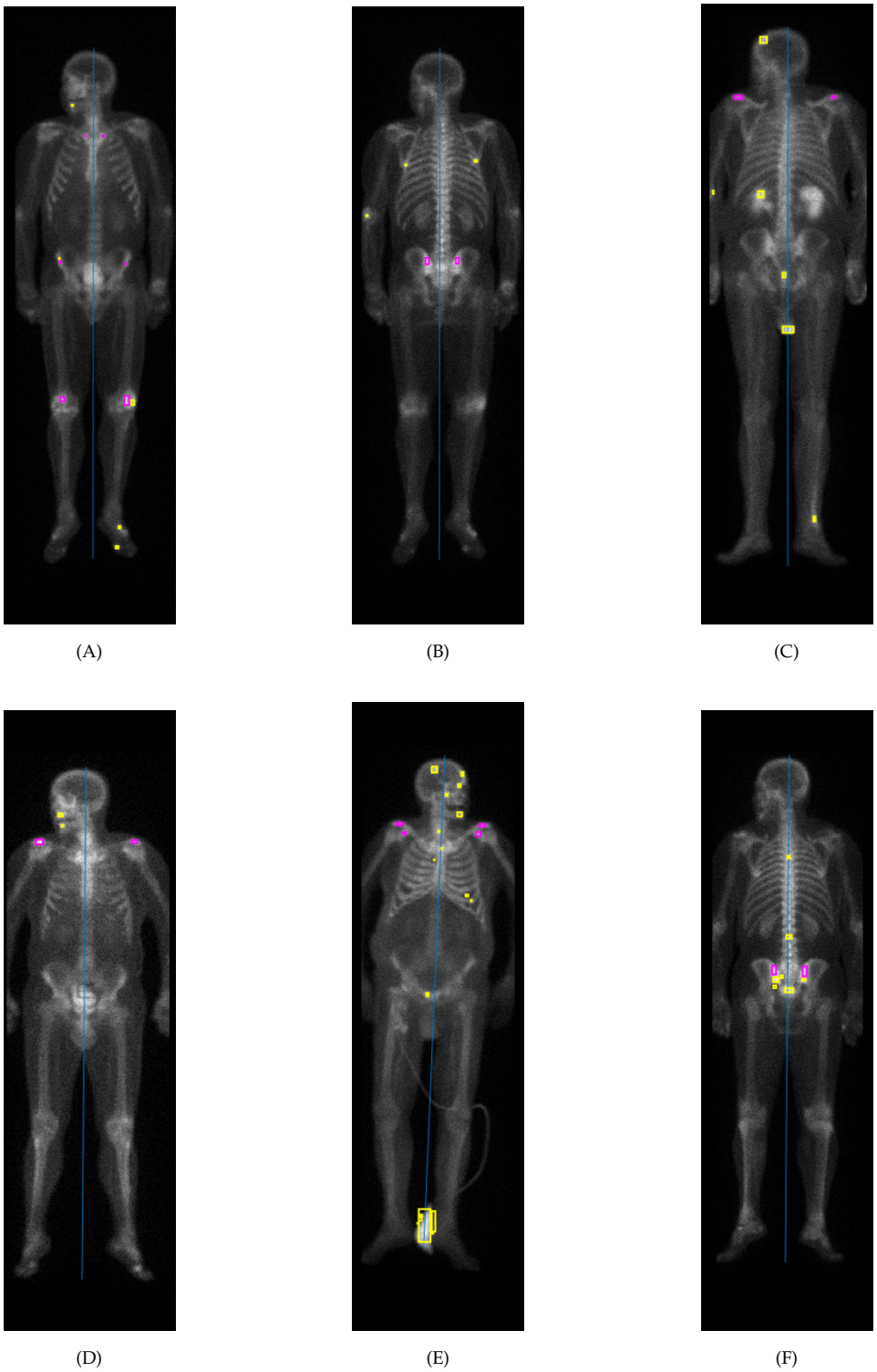
The two algorithms were at last combined into one that automatically removed hotspots found in certain anatomical regions and symmetrical hotspots at the same time (see Figure 5.8).



(A) (B) (C)

(D) (E) (F)

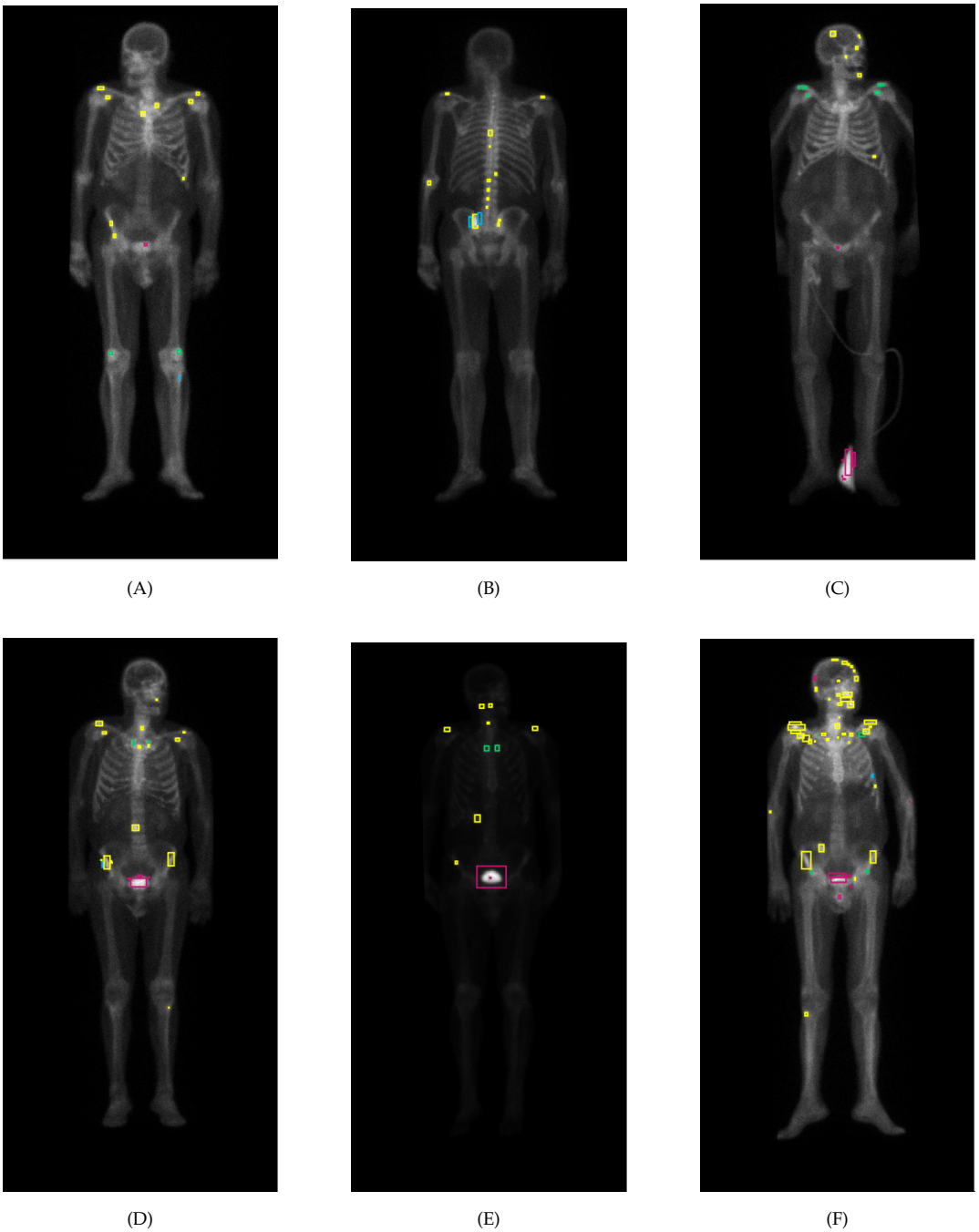
FIGURE 5.6: Results for the removal of hotspots found in certain anatomical regions. The detections with a pink bounding box are considered to be false-positives due to the anatomical region they are in. These regions included the bladder (A, B, C, D and E), the hands (B, E and F), the feet (B) and urine-collection bags (C and D).



(A) (B) (C)

(D) (E) (F)

FIGURE 5.7: Results for the removal of symmetric hotspots. The detections with a pink bounding box are considered false-positives due to their symmetry.



(A) (B) (C)

(D) (E) (F)

FIGURE 5.8: Results for the removal of hotspots using image analysis techniques. Hotspots found in unwanted anatomical regions are in pink and symmetrical hotspots are in the green.

5.2.2.3 Algorithm evaluation

To evaluate this method for false-positive attenuation, all hotspots from the test set were labelled as 0 (non-malignant) if they were considered to be false-positives and 1 (malignant) otherwise. The predicted labels were compared to the true labels and the sensitivity, specificity, accuracy, false negative rate (FNR), false positive rate (FPR) and false positive detections per image (FPPI) were calculated (see Table 5.2)

TABLE 5.2: Results of the algorithm when removing (i) only symmetrical hotspots, (ii) only hotspots found in certain anatomical regions and (iii) symmetrical hotspots and hotspots found in certain anatomical regions.

	Symmetry	Region	Total
Sensitivity	0.93	0.93	0.87
Specificity	0.10	0.23	0.30
Accuracy	0.32	0.42	0.45
FNR	0.07	0.06	0.13
FFPI	29	25	22

Discussion

With this algorithm, a high sensitivity was achieved (87%). This was expected, as the goal was to remove false-positive detections without losing any of the true-positive ones (metastases). Even so, the fact that this value was not 100% shows that a few metastases were lost; analysing the figures from the data set, this mainly happened in patients with a high density of metastases in the spine and pelvic area, where some of the hotspots fall under the symmetry conditions. The specificity score shows that with this algorithm 30% of the non-malignant hotspots were correctly identified as false-positives. The number of false-positive detections per image (FPPI) goes from 32 to 22, corresponding to a decrease of 30%. These values (sensitivity and FPPI) could be improved if a broader range of values for the symmetry conditions were considered; the reason why did was not done is because it would come at the cost of more malignant hotspots being wrongly considered false-positives, specially in patients with high density of metastases.

5.2.3 Classifiers

The results from the different algorithms described in Section 4.3.4 are now presented and discussed. The section is divided in two parts: (i) 3-class classifiers, which includes the results for the 3-class k-means clustering and 3-class iterative algorithm (hotBSI) and

(ii) binary classifiers, which include the results for the for the 2-class k-means clustering, one-class classification, and 2-class iterative algorithm (hotBSI).

5.2.3.1 Three-class classifiers

Here, the results for the three-class classification algorithms are shown. Section 5.2.3.1 presents the results for the 3-class k-means clustering algorithm and Section 5.2.3.1 for the iterative algorithm (hotBSI). This problem is an example of an imbalanced classification problem where the smaller classes (malignant hotspots, 27% of the test set and benign hotspots, 0.8% of the test set) is more important than the majority class (healthy hotspots, 72% of the test set). Having a classifier that would predict every hotspot to be healthy would do no good because it was failing to diagnose patients with bone metastases. Based on the analysis made in Section 2.7.1, and because the minority classes are more valuable, the macro-average metrics were considered to be more suitable to this problem. For the purpose of comparison with the binary algorithms, the "malignant metrics" were also obtained, that is, the metrics calculated assuming the malignant class as the positive class and the healthy and benign classes as the negative class.

Three-class k-means clustering

After applying the k-means algorithm to the **training data**, the final three clusters ended up with the following number of samples:

cluster 1: 580 samples
cluster 2: 272 samples
cluster 3: 231 samples

when using the handcrafted features and:

cluster 1: 712 samples
cluster 2: 291 samples
cluster 3: 206 samples

when used high-level features extracted from a pre-trained ResNet18 network. For both features, it was assumed cluster 1 represented the benign hotspots, cluster 2 malignant hotspots and clusters 3 the healthy hotspots. When applied to the **test set**, the model

trained with handcrafted features assigned 173 hotspots to cluster 1, 236 hotspots to clusters 2 and 902 to cluster 3; the model trained with ResNet18 features assigned 566 hotspots to cluster 1, 90 hotspots to clusters 2 and 655 to cluster 3 (see Table 5.3). The results obtained for the k-means model are gathered in Tables A.1 (macro-averages) and A.2 (assuming the malignant class as the positive class) and Figures A.1a (confusion matrix for handcrafted features) and A.1b (confusion matrix for ResNet18 features).

TABLE 5.3: Number of hotspots from the test set assigned to clusters 1, 2 and 3, for the handcrafted and ResNet18 features

	Cluster 1	Cluster 2	Cluster 3
Handcrafted	173	236	902
ResNet18	566	90	655

Three-class iterative algorithm (hotBSI)

The results from the three-class iterative algorithm, hotBSI, are presented in:

- the confusion matrices from Figures A.2a, A.2b and Tables A.3 and A.4 for hotBSI-SVM;
- the confusion matrices from Figures A.3a, A.3b and Tables A.5 and A.6 for hotBSI-KNN;
- the confusion matrices from Figures A.4a, A.4b and Tables A.7 and A.8 for hotBSI-DTs;
- the confusion matrices from Figures A.5a, A.5b and Tables A.9 and A.10 for hotBSI-LDA.

Discussion

The 3-class algorithms present very similar macro-average results: a sensitivity in the range of [0.33-0.44] and a specificity in the range of [0.65 - 70]. The algorithm that yells the best results is the k-means with ResNet18 features, which achieved a sensitivity, specificity and accuracy of 0.44, 0.70 and 0.61, respectively. One has, however, to be specially careful when using macro metrics for the problem of identifying bone metastases. Comparing, for instance, the macro-average metrics from the best model (Table A.1) with the "malign" metrics for the same model (Table A.2), one can see the sensitivity drops from 0.44 to 0.08,

meaning that this algorithm has almost no ability to classify malignant hotspots as such. This happens because the algorithm is assigning most of the hotspots to the healthy class, and because the majority of the hotspots from the test set are from this class, the high sensitivity obtained to detect healthy hotspots will compensate the very low sensitivity to detect metastases. The hot-BSI alorithms trained with KNN (Table A.6), DTs (Table A.8) and LDA (Table A.10), on the other hand, present a vert high sensitivity for the "malign" metrics, with scores ranging from 0.87 to 0.97. Although this may seem a very good result at first, looking at the respective confusion matrices, and noticing the low specificity results, one can see that this high sensitivity scores are happening because the algorithm is assigning almost every hotspot to the malignant class, which means that it has a lower discriminatory power.

Although the results for the three-class algorithm are here presented, a deeper discussion of the results will be given to the binary classifiers, which are presented in the next section. Two main reasons for this decision are given. First, it is important to remember that the main goal of this work was to build an algorithm that could quantify bone scan lesions, which is done by the assessment of bone metastases. The most important requirement is, therefore, to have a classifier that can identify if a hotspot is malignant or not. As it could be seen with the results from the 3-class algorithms, sometimes behind a reasonable macro-average score is a very low ability for the algorithm to identify bone metastases. Building an algorithm that is also capable of distinguishing between healthy and benign hotspots can, of course, bring several benefits for the medical community, but it is a challenge that goes beyond the purpose of this work. The second reason is that with a 3-class classifier a proper comparison with the state-of-the-art one-class classifier would not be possible, as it is a binary algorithm.

5.2.3.2 Binary classifiers

The results for the binary classification algorithms are now presented and discussed. Section 5.2.3.2 gathers the results for the 2-class k-means clustering algorithm, Section 5.2.3.2 for the one-class classification algorithm and Section 5.2.3.2 for the iterative algorithm (hotBSI).

Two-class k-means clustering

After applying the k-means algorithm to the training data, the final two clusters ended up with the following number of samples:

cluster 1: 6 619 samples

cluster 2: 942 samples

when using the handcrafted features and:

cluster 1: 7 017 samples

cluster 2: 544 samples

when used high-level features extracted from a pre-trained ResNet18 network. For both features, it was assumed theat cluster 1 represented the non-malignant hotspots (negative class) and cluster 2 represented the metastases (positive class). When applied to the test set, the model trained with handcrafted features assigned 1 113 hotspots to cluster 1 and 198 hotspots to clusters 2; the model trained with ResNet18 features assigned 1 205 hotspots to cluster 1 and 106 hotspots to cluster 2 (see Table 5.4). The results obtained for the k-means model are gathered in Table A.11 and Figures A.6a and A.6b.

TABLE 5.4: Number of hotspots from the test set assigned to cluster 1 and 2, for the handcrafted and ResNet18 features

	Cluster 1	Cluster 2
Handcrafted	1 113	198
ResNet18	1 205	106

Discussion

Analysing first the results from the algorithm using the handcrafted features, it can be observed that this model is assigning almost every hotspot (more precisely, 88%) from the test to cluster 1, which was assumed to be the cluster of the non-malignant hotspots. As a consequence, it will present a high specificity (86%), as the majority of the samples are being assigned to the negative class, and therefore there is a higher probability that the algorithm classifies a non-malignant hotspot as such. As a consequence, this model presents a very low sensitivity (17%), since the majority of the malignant hotspots are being wrongly classified as non-malignant. This also results in a very high false negative rate (83%). This is a big downside in this specific problem of hotspots classification, as if

this were to be the final model used in the medical context, it would mean that a lot of metastases would falsely be classified as healthy hotspot, which would have significant impact on the patient's health.

The algorithm that used the high-level features has a very similar behaviour. Analysing its results, it can be observed that this model is also assigning almost every hotspot of the test set to cluster 1 (93%). As a consequence, it will present a high specificity (92%) as well as a high false negative rate (92%), while keeping the sensitivity very low (8%). For the reason previously mentioned, such an algorithm would have to be improved before it could be used in the clinical practice.

One-class classification

The results obtained with the one-class classifier are gathered in table A.12. The confusion matrices for the algorithms trained with handcrafted and ResNet18 features are shown in Figures A.7a and A.7b. Once again, class one in the confusion matrix represented non-malignant hotspots (negative class) whilst class two represented malignant hotspots (positive class). Both the handcrafted and deep features models presented similar results to the k-means algorithm, showing a high specificity while keeping the sensitivity quite low. The consequences of using these models in the clinical practice are therefore identical to the ones previously mentioned. Both models got an AUC equal of close to 0.5, which means they have no discrimination power and that they perform no better than a random classifier.

Two-class iterative algorithm (hotBSI)

The results from the binary iterative algorithm are presented in:

- the confusion matrices from Figures A.8a, A.8b Table A.13 for hotBSI-SVM;
- the confusion matrices from Figures A.9a, A.9b and Table A.14 for hotBSI-KNN;
- the confusion matrices from Figures A.10a, A.10b and Table A.15 for hotBSI-DTs;
- the confusion matrices from Figures A.11a, A.11b and Table A.16 for hotBSI-LDA.

Discussion and choice of the best algorithm

The results obtained for the binary hotBSI algorithms are now discussed and compared to the results obtained with the two other state-of-the-art binary algorithms (k-means and one-class classification). The discussion will focus on the metrics considered to be the most relevant for choosing the best classifier.

Area Under the ROC Curve (AUC)

The AUC values, usually very close or equal to 0.50, translate the very low to none capacity of most classifiers to distinguish between non-malignant and malignant hotspots. Comparing the AUC obtained with handcrafted and ResNet18 features, it can be concluded that the latter always performs better than the former. The highest AUC score was obtained with the hotBSI trained with SVM and ResNet18 features (AUC = 0.66).

Sensitivity and Specificity

Very high values of sensitivity and specificity were only obtained when the classifier was biased towards one class: high sensitivity scores (> 0.85) were always hand with hand with a very low specificity score, which meant that it was considering almost every hotspot to belong to the positive (malignant) class; on the other hand, high specificity scores (> 0.85) were always hand with hand with a very low sensitivity score, meaning that it was assigning the majority of hotspots to the negative (non-malignant) class. The k-means and OCC algorithms fall under the latter situation. Neither situation is desirable for the final algorithm. The classifiers with more balanced scores in terms of sensitivity and specificity were (i) the hotBSI trained with SVM and ResNet18 features (sensitivity = 0.63, specificity = 0.58) and (ii) the hotBSI trained with KNN and ResNet18 features (sensitivity = 0.67, specificity = 0.51).

False Negative Rate (FNR)

An important evaluation metric for an algorithm whose goal is to classify hotspots in patients who might have bone metastases is the false negative rate. It is desirable that this value is as low as possible, as a high FNR would mean that the classifier was incorrectly labelling a lot of malignant hotspots (metastases) as non-malign; this would result in an algorithm that would label patients with metastatic cancer as healthy, which would be

very dangerous is the clinical context. Very low FNR only happened with classifiers that were assigning almost every hotspot to the malignant class: taking a look at the hotBSI trained with decision trees it can be observed that a FNR rate of 0.08 was obtained. Although at first glance this may seem like an almost perfect result, further analysis on the remaining metrics lead us to conclude that this FNR only happens because the classifier is assigning almost every hotspot to the malignant class and, therefore, it had a very low probability of missing metastases (sensitivity = 0.92, specificity = 0.14). Such a classifier is obviously not acceptable, as it has no discriminatory power. On the opposite side of the spectrum, the k-means algorithm, the OCC or the hotBSI-SVM with handcrafted features are assigning almost every hotspot to the non-malignat class and therefore have false positive rates greater than 0.82.

Classifiers that obtained lower FNR while keeping more acceptable values for the other metrics include (i) the hotBSI trained with discriminant analysis and ResNet18 features (FNR = 0.30), (ii) the hotBSI trained with KNN and ResNet18 features (FNR = 0.33) and (iii) the hotBSI trained with SVM and ResNet18 features (FNR = 0.37).

False positive rate reduction

As mentioned in section 5.1, the detection algorithm presented a false-positive rate of 73%. By applying the classification algorithms, it was hoped that this rate would lower, so that non-malignant hotspots were discarded. The lowest FPR scores were obtained with (i) the hotBSI-SVM trained with handcrafted features (FPR = 0.18), (ii) the OCC trained with handcrafted features (FPR = 0.10) and (iii) k-means with handcrafted and ResNet18 features (FPR = 0.14 and FPR= 0.10, respectively). This low values are, however, only due the fact that these algorithms were classifying most of the metastases as non-malign, which is not desirable, as it will lead to a very high FNR. The classifier that presented the lowest FPR while keeping an acceptable value for the FNR was the hotBSI-SVM with ResNet18 features (FPR = 0.42). This represents a decrease of 30.59% compared to the FPR score obtained with initial detection algorithm, when no classifiers had been yet applied. Looking at the number of false-positive detections per image, one can also see that this value drops from 32 to 14.

Comparison with the state-of-the-art algorithms

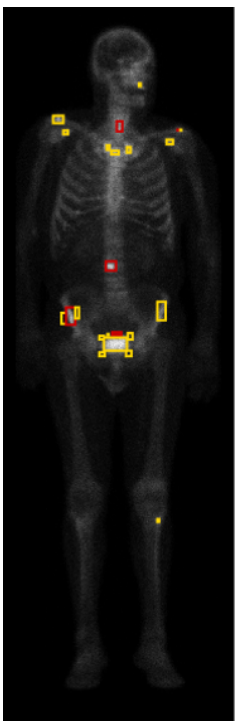
The best algorithm was considered to be the binary hotBSI trained with SVM and ResNet18 features. As the best algorithm turned out to be the one that was original proposed in this thesis, it is interesting to compare its results with the ones obtained with the best state-of-the-art models (2-class k-means and OCC). The best k-means and one-class algorithms were considered to be the ones trained with handcrafted and ResNet18 features, respectively. Table 5.5 gathers the best results obtained with these three models: the binary hotBSI-SVM, the k-means algorithm, and the one-class classifier.

The hotBSI algorithm shows superiority in almost every metric, in particular in the AUC (0.66 compared to 0.50 from the OCC classifier), sensitivity (0.63 compared to 0.17 and 0.26 from the k-means and OCC classifiers, respectively) and the false negative rate (0.37 compared to 0.83 and 0.74 from the k-means and OCC classifiers, respectively). It should be noted that the only two metrics in which the state-of-the-art algorithms performed better were accuracy and specificity. This is clearly explained by noting that this happens since these algorithms are classifying most of the hotspots as non-malignant (note the very low sensitivity from the same classifiers); as a consequence, they will present a high specificity, as if most of the hotspots are being classified as non-malignant there is a better chance that the algorithm will correctly classify non-malignant hotspots as non-malignant. Besides the low specificity, this comes with a cost of a very high false negative rate, as a lot of malignant hotspots are being incorrectly classified as non-malignant. The better scores in accuracy are also easily explained by looking at the percentage of non-malignant and malignant hotspots present in the test set: 73% of these hotspots were from the non-malignant category, while only 27% were from the malignant category. Because the k-means and OCC classifiers are manly assigning hotspots to the negative (non-malignant) class, and because most of the test set is composed by hotspots from this class, they will get a high accuracy score, even if most of the malignant hotspots are wrongly classified. Having all of this into account, it can be concluded that the proposed algorithm performs better than the state-of-the-art algorithms at the task of hotspots classification and, therefore, at the task of false-positive attenuation.

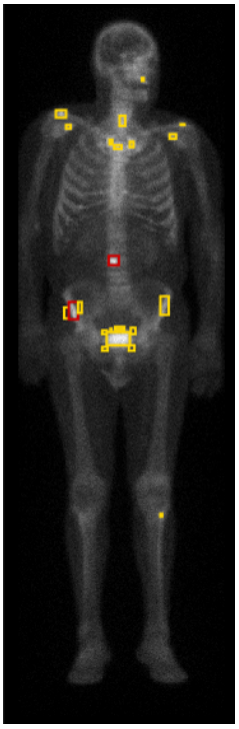
TABLE 5.5: Comparison of the best models

	hotBSI (RN18)	K-means (HC)	OCC (RN18)
Sensitivity	0.63	0.17	0.26
Specificity	0.58	0.86	0.72
Accuracy	0.59	0.67	0.60
FNR	0.37	0.83	0.74
FPR	0.42	0.14	0.28
Precision	0.35	0.30	0.26
F1	0.46	0.22	0.14
AUC	0.66	–	0.50
FPPI	14	27	9

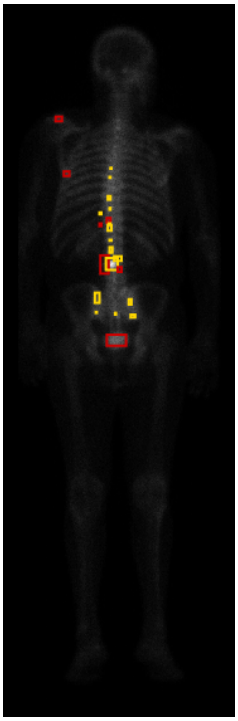
A visual representation of the model in action (hotBSI-SVM) is now presented. Figures 5.9, 5.10 and 5.11 show the AP and PA of views of a bone scintigraphy image, with the hotspots classified according to the hotBSI-SVM algorithm. For the sake of comparison, in the same image it is also shown the AP and PA views labelled according to the ground truth.



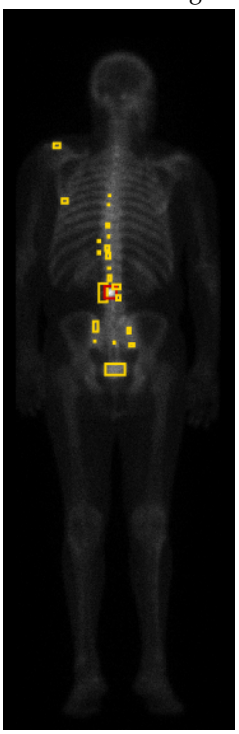
(A) AP view labelled with hotBSI



(B) AP view labelled with ground truth

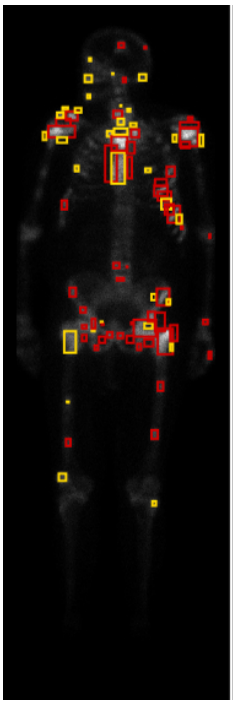


(C) PA view labelled with hotBSI

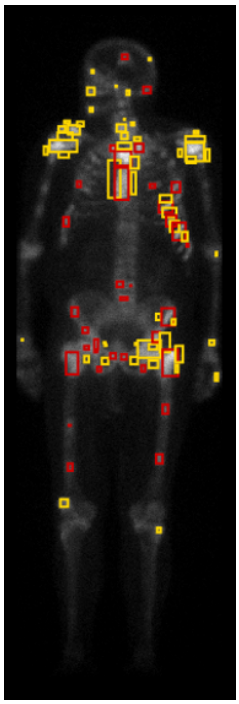


(D) PA view labelled with ground truth

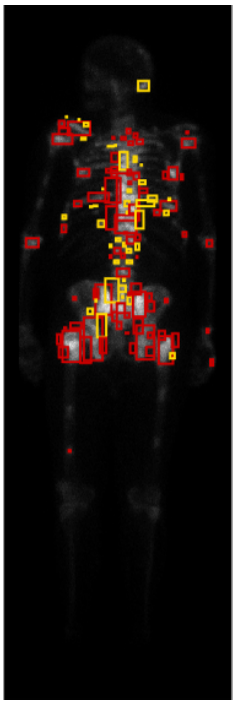
FIGURE 5.9: (Example 1) Comparison between the hotspots classification made by the best iterative algorithm (hotBSI-SVM) and the respective ground truth. Hotspots in yellow represent benign hotspots and hotspots in red represent malignant hotspots (metastases)



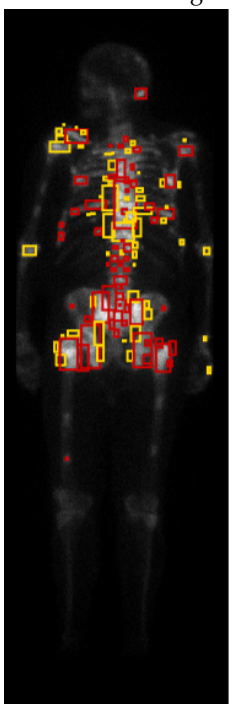
(A) AP view labelled with hotBSI



(B) AP view labelled with ground truth

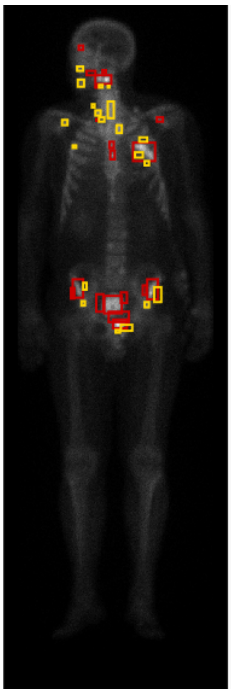


(C) PA view labelled with hotBSI

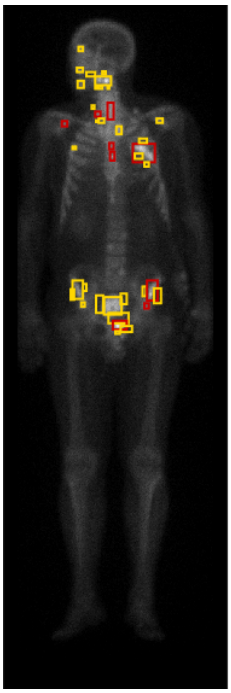


(D) PA view labelled with ground truth

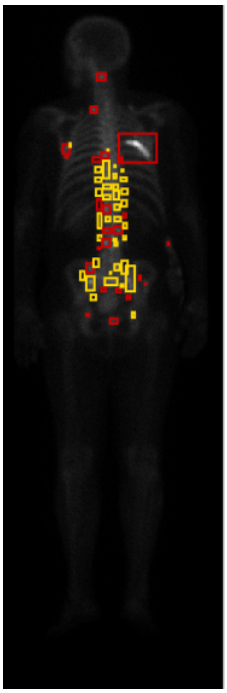
FIGURE 5.10: (Example 2) Comparison between the hotspots classification made by the best iterative algorithm (hotBSI-SVM) and the respective ground truth. Hotspots in yellow represent benign hotspots and hotspots in red represent malignant hotspots (metastases)



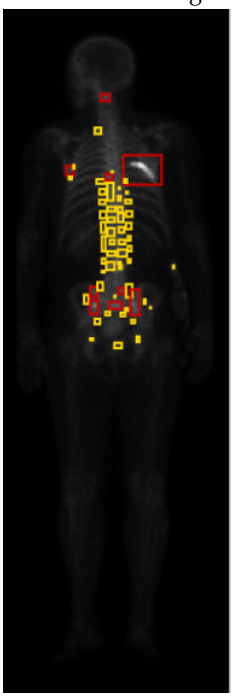
(A) AP view labelled with hotBSI



(B) AP view labelled with ground truth



(C) PA view labelled with hotBSI



(D) PA view labelled with ground truth

FIGURE 5.11: (Example 3) Comparison between the hotspots classification made by the best iterative algorithm (hotBSI-SVM) and the respective ground truth. Hotspots in yellow represent benign hotspots and hotspots in red represent malignant hotspots (metastases)

5.3 BSI Calculation

The final stage was to quantify the bone scan using the Bone Scan Index. Besides providing the physician with information about disease staging and prognosis, the increase/decrease of BSI over successive exams from the same patient will give information about disease progressions/regression. For a patient who is receiving treatment for bone metastases, this will give the physician information about the patient's response to therapy.

To calculate the BSI for a certain bone scintigraphy image, the the following steps are applied:

1. The hotspots are detected using the algorithm described in Section 4.2;
2. False-positive detections are attenuated using the two methods described in Section 4.3;
3. The hotspots that were not identified as non-malignant during step 2 are considered to be metastases, and the BSI is calculated as explained in Section 4.4.

To evaluate the overall quantification algorithm, patients from the database were selected. These were patients who were diagnosed with bone metastases, and that had therefore bone scintigraphy exams done regularly. By calculating the BSI with the here proposed algorithm and by analysing its evolution over time, it can be evaluated if the BSI variation is in agreement with the medical reports for disease evolution. It is important to point out that if the BSI variation does not reflect the condition described in the medical report, it does not mean the the BSI is not a good imaging biomarker for bone scan quantification, but rather that the here proposed algorithm is somehow failing in identifying the metastases. The BSI has already been validated as a suitable parameter for bone scan assessment as presented in Section 3.5.

Patient A

Patient A underwent 16 bone scintigraphy exams over a period of almost 5 years. The average time in between exams was of 7.6 months. The BSI for each of the exams was calculated with the algorithm developed during this work. Figure 5.12 shows a graphic with the evolution of the BSI during the course of time. Figures 5.13 (exams 1 to 8) and 5.14 (exams 9 to 12) show the bone scans obtained for Patient A throughout the months.

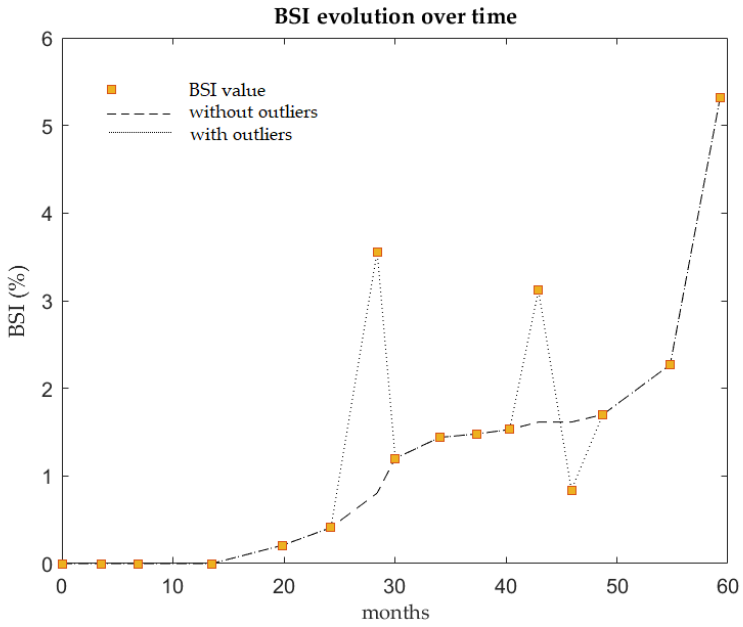


FIGURE 5.12: Evolution of the BSI obtained for Patient A bone scintigraphy exams

The evaluation of the algorithm cannot be done in a quantitative way, as the true BSI values are not available, but can be done qualitatively by comparing the evolution of the BSI with the medical reports. The terms used by the physicians in these reports include:

- *Absence of selective fixation foci of diphosphonates at bone level that could suggest metastases*, if there are no suspicious bone uptakes that may indicates metastases;
- *Diphosphonates uptake at bone level suggestive of metastases*, if there are suspicious bone uptakes that may indicates metastases;
- *Disease aggravation*, if there seems to be an increase in fixation intensity compared to the previous exam;
- *Slight improvement in scintigraphic findings*, if there seems to be a decrease in fixation intensity compared to the previous exam;
- *Similar*, if there does not seem to be an improvement or worsening of the disease since the last exam.

A comparison between the disease evolution according to the here developed algorithm and the medical reports is now made:

- **Exams 1 - 4:** According to the medical reports, no scintigraphic findings suggestive of metastases are found. The results obtained with the here proposed algorithm are in agreement with this, as a BSI = 0 is found for the four exams. The reports mention possible benign bone lesions in the 5^{th}, 6^{th} and 9^{th} costovertebral joints: these can indeed be found in the PA views from exams 1 to 3 in Figure 5.13, and are correctly classified as non-malignant. The algorithm fails to detect these in exam 4; this turns out not to be a problem since, according to the medical reports, no metastases are found, but could be problematic if it failed to detect metastatic hotspots.

- **Exam 5:** The medical report indicates scintigraphic findings suggestive of metastases. The results obtained with the here proposed algorithm are in agreement with this (BSI increases from 0 to 0.21%). However, the medical report indicates possible metastases in the spine, sternum, costal margin (bottom edge of the rib cage), pelvis, and proximal right humerus, while the current algorithm only classifies as malignant one hotspot in the costal margin and two in the spine; this means that the BSI would likely be higher than the one achieved by the algorithm.

- **Exams 6-7:** The medical reports indicate an increase of scintigraphic findings (disease aggravation). The current algorithm also translates this disease aggravation with an increase of BSI in both exams: BSI = 0.41% in exam 5 and BSI = 3.56% in exam 6. In addition to the previously detected malignant hotspots, the medical reports also refers metastases in the skull and left femoral neck, which are also found in exams 6 and 7 from Figure 5.13. In exam 6, the algorithm classifies as malignant the hotspots found in the humerus, spine and ribcage, but still classifies as benign the ones found in the pelvis, skull and femur. In exam 7, however, hotspots in the skull and humerus are already classified as metastases.

- **Exams 8-12**: The medical reports indicate no significant increase or decrease in scintigraphic findings, which should translate in a similar BSI in all these exams. The here proposed algorithm shows a sudden decrease in BSI from the 7^{th} to 8^{th}

exam (which, according to the medical report, should not happen)*. Nevertheless, it is interesting to verify that exams 8th to 11th experience very little variation in BSI scores (ΔBSI = 0.24, 0.04 and 0.05, respectively), which is in agreement with clinical reports, that say that no significant evolution of the disease has occurred. In exam 12 there is a significant increase in BSI (ΔBSI = 1.59%), even though it was expected that this value would be similar to the previous one.

- **Exam 13:** According to the medical reports, in exam 13 there is a slight improvement in scintigraphic findings; the here proposed algorithm presents a decrease of 2.28% in the BSI regarding the previous exam. Although according to the medical report a slight decrease in intensity was expected, the decrease in BSI should not be that sharp. Observing the exam 13 from Figure 5.14, it can be observed that the algorithm fails to classify as metastases hotspots that it had classified as malignant in previous exams.

- **Exam 14:** The medical report indicates no significant increase or decrease in scintigraphic findings, which should translate in a similar BSI to the one obtained in the 13th exam. With the here proposed algorithm, an increase of 0.86% is, however, obtained. This can be explained by the fact that exam 13 had a BSI lower than what it should have had, and in this exam hotspots that were not being classified as metastases in exam 13 are now (exam 14) being classified as such.

- **Exams 15 - 16:** The medical reports indicate an increase of scintigraphic findings (disease aggravation). The current algorithm also translates this disease aggravation with an increase of BSI in both exams: BSI = 0.57% in exam 15 and BSI = 3.05 in exam 16

Overall, the results obtained with the developed algorithm are in agreement with the medical reports, except from three outliers (exams 7, 12 and 13). In exams 1 to 4 the BSI is equal to zero and according to the medical reports no metastases are detected; in exams 5, 6 and 8 there is an increase in BSI, and according to the medical reports there is an aggravation of the disease; in exams 9, 10, 11, and 14 the BSI varies little, just like the patient condition does not change much according to the medical reports; in exams

*The medical reports indicate a high number of metastases, and therefore it seems more likely that the value of BSI obtained in the 7th is closer to the truth one, and exam 8th to 12th fail to classify some hotspots as malignant.

15 and 16 there is an increase in BSI, and according to the medical reports there is an aggravation of the disease. In Figure 5.15 a graphic of the variation of the BSI is shown. Each point is obtained by calculating the difference in BSI score between the current and the previous exams. A positive/negative ΔBSI indicates an increase/decrease in bone uptake from the previous exam to the current one. The colour of each point corresponds to the information given by the respective medical reports: green if the reports indicate an improvement, yellow if it remains the same and red it it indicates an aggravation of the disease.

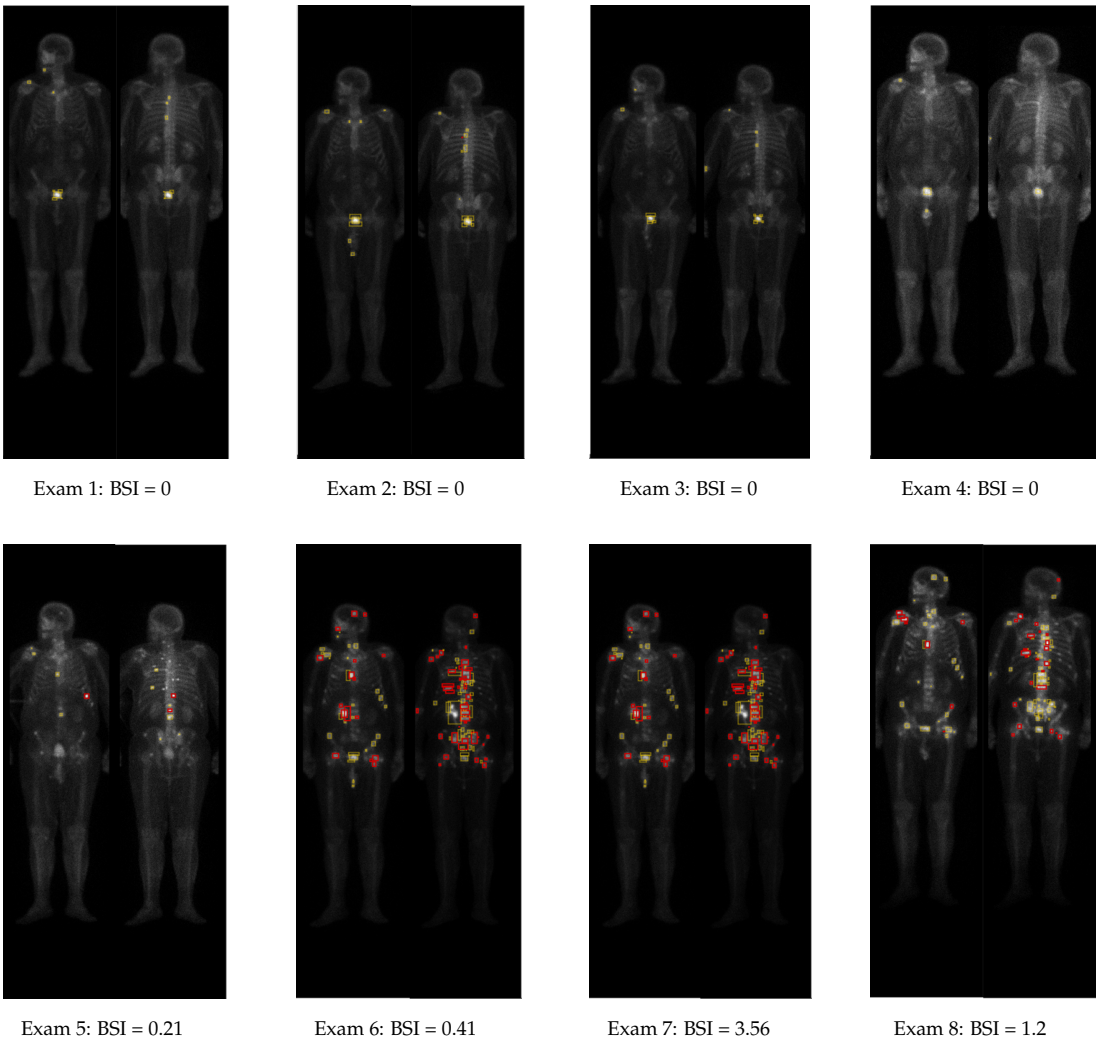
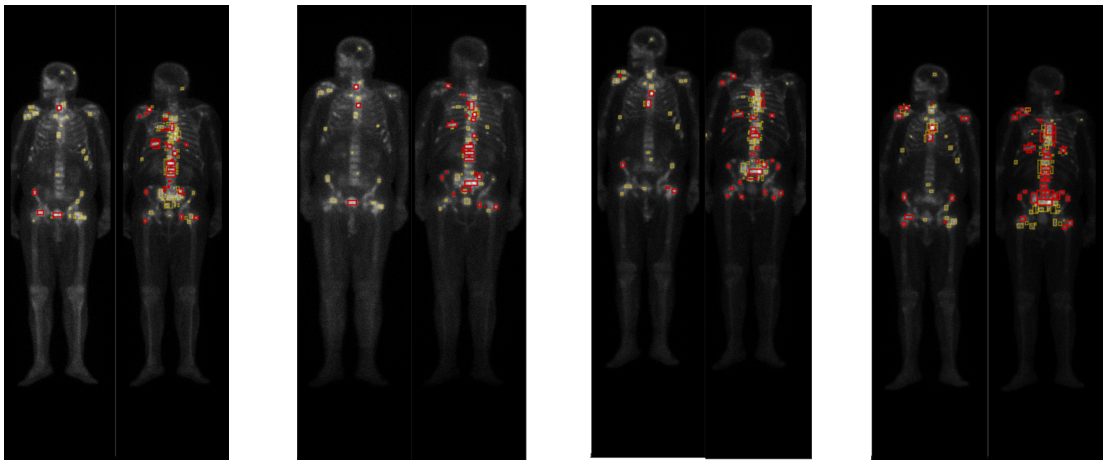
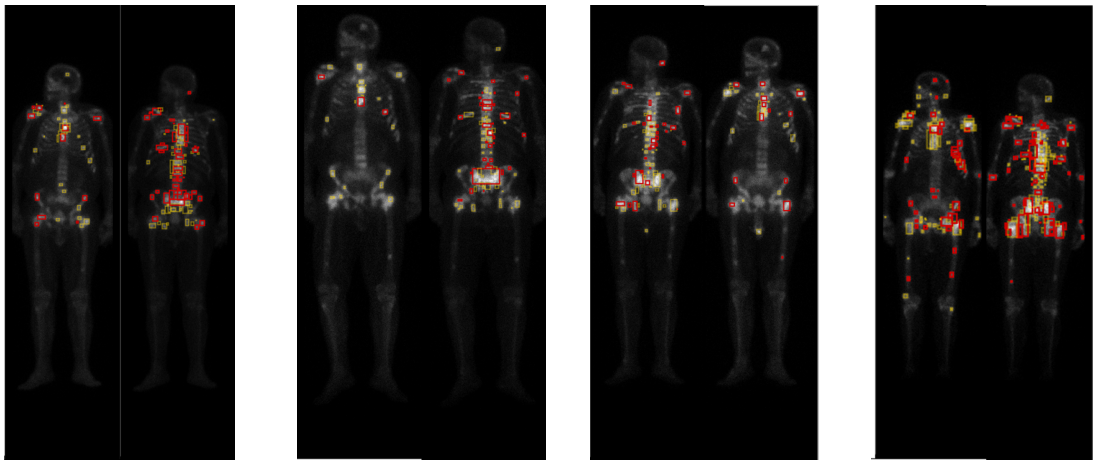


FIGURE 5.13: Exams 1-8 for Patient A, with the detection and classification of hotspots according to the proposed algorithm. In each image, the AP (left) and PA (right) views are shown, as well as the BSI score in percentage.



Exam 8: BSI = 1.44 | Exam 10: BSI = 1.48 | Exam 11: BSI = 1.53 | Exam 12: BSI = 3.12



Exam 13: BSI = 0.84 | Exam 14: BSI = 1.7 | Exam 15: BSI = 2.27 | Exam 16: BSI = 5.32

FIGURE 5.14: Exams 1-8 for Patient A, with the detection and classification of hotspots according to the proposed algorithm. In each image, the AP (left) and PA (right) views are shown, as well as the BSI score in percentage.

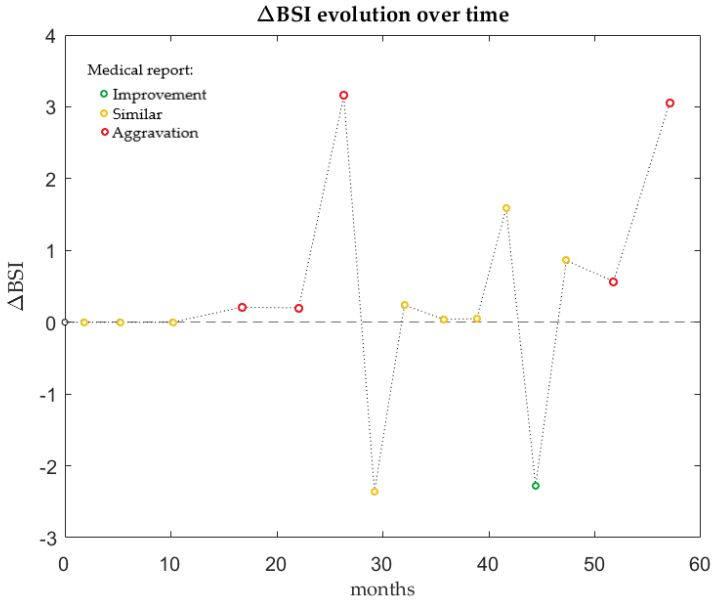


FIGURE 5.15: Variation of the BSI obtained from Patient A bone scintigraphy exams throughout the months. The ordinate of each point corresponds to the difference in BSI score between the current and the previous exams. The colour of the points translate the information given by the respective medical report.

Patient B

Patient B underwent 5 bone scintigraphy exams over a period of almost 3 years. The average time between exams was of 8.0 months. Figure 5.17 shows a graphic with the evolution of the BSI during the course of time. Figure 5.16 shows the bone scans obtained for Patient B throughout the months. In figure 5.18 a graphic of the variation of the BSI (just like the one obtained for patient A) is shown. A comparison between the disease evolution according to the here developed algorithm and the medical reports is now made:

- **Exam 1:** According to the medical reports, no scintigraphic findings suggestive of metastases are found. The results obtained with the here proposed algorithm are in agreement with this, as a BSI = 0 is found for the first exam.

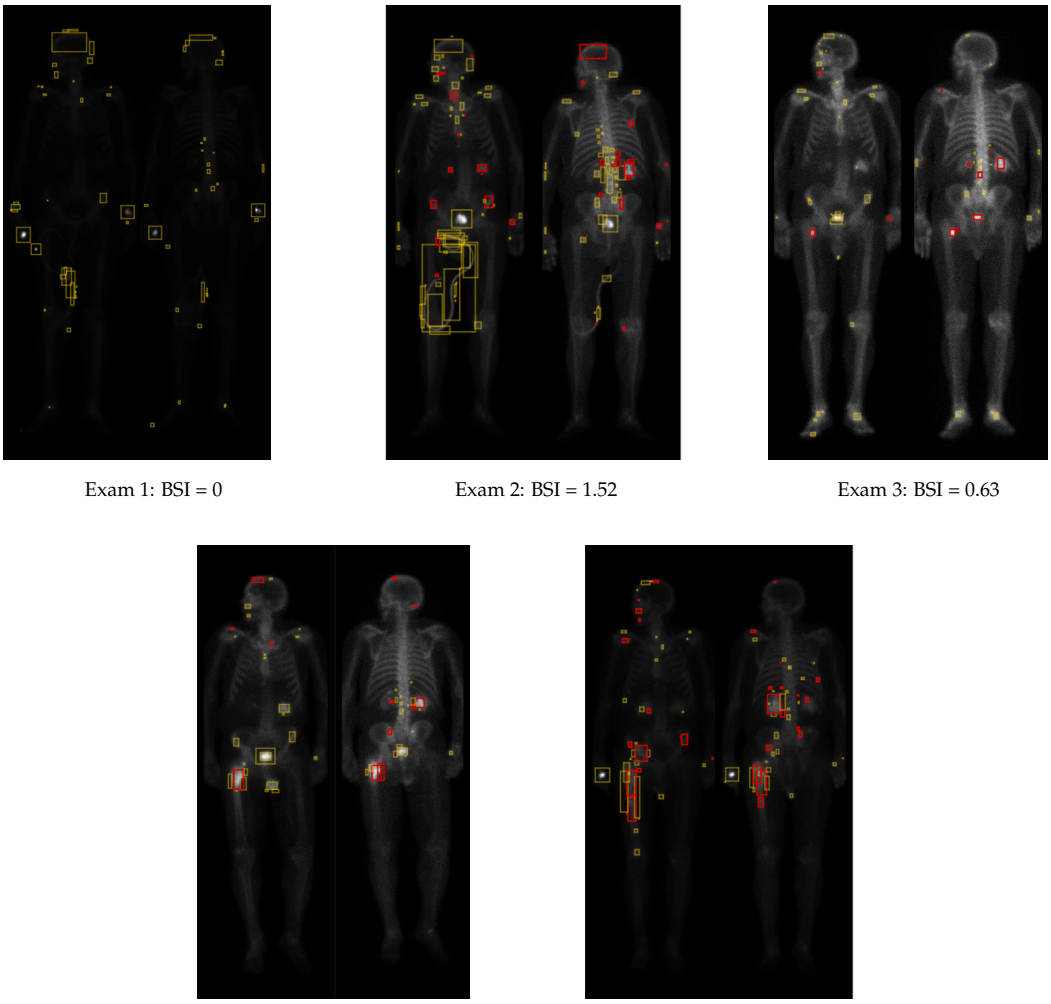
- **Exam 2:** The medical reports indicate a possible metastasis in the femur trochanter, which was detected with the here proposed algorithm and classified as malignant in the AP view and as benign in the PA view. The value obtained for the BSI, however, seems to be considerably higher than what it should be. While the medical report only suggests a possible metastasis in the femur trochanter, the here proposed algorithm found a lot more metastases in the head, pelvic bones, ribs, etc., explaining the high BSI value.

- **Exam 3:** The medical report refers again the a possible metastasis in the femur trochanter, which was correctly identified with the here proposed algorithm. Although a few other hotspots are falsely classified as malign, this did not happen as often as it did in the previous exam. The obtained BSI score is, therefore, smaller than the one obtained in the previous exam, and closer to what the real BSI must be.

- **Exam 4:** The medical report indicates an aggravation of the metastasis in femur trochanter, with no other possible malignant lesions being found. The here proposed algorithm also detects a malignant hotspot in the femur trochanter with an increased side relatively to the previous exams, and the BSI score also increases. It should be noted, however, that the real BSI value is likely smaller than the one obtained, as some hotspots are incorrectly classified as malignant.

- **Exam 5:** The medical reports indicate an increase of scintigraphic findings (disease aggravation). In addition to the previously detected metastasis in the femur, which

now appears with greater extent, suspicious radiotracer fixation is also found in the costal arch (lower edge of the chest formed by the bottom edge of the rib cage), left iliac wing and right acetabulum. As it can be observed in Figure 5.16, the metastases in the femur, iliac and acetbulum are correctly classified as such; some hotspots detected in the costal arch are classified as malignant and others as non-malignant. The BSI score increases with respect to the previous exam, which is in agreement with the clinical report.



Exam 1: BSI = 0 Exam 2: BSI = 1.52 Exam 3: BSI = 0.63

Exam 4: BSI = 1.45 Exam 5: BSI = 2.93

FIGURE 5.16: Exams for Patient B, with the detection and classification of hotspots according to the proposed algorithm. In each image, the AP (left) and PA (right) views are shown, as well as the BSI score in percentage.

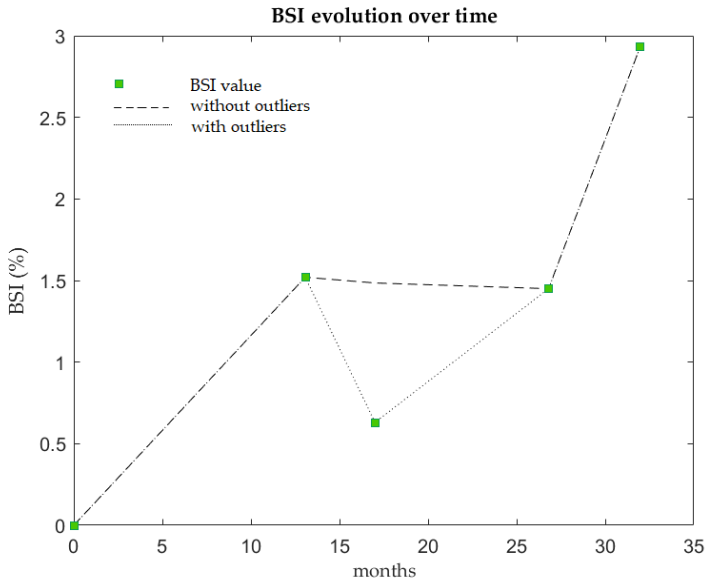


FIGURE 5.17: Evolution of the BSI obtained from Patient B bone scintigraphy exams throughout the months

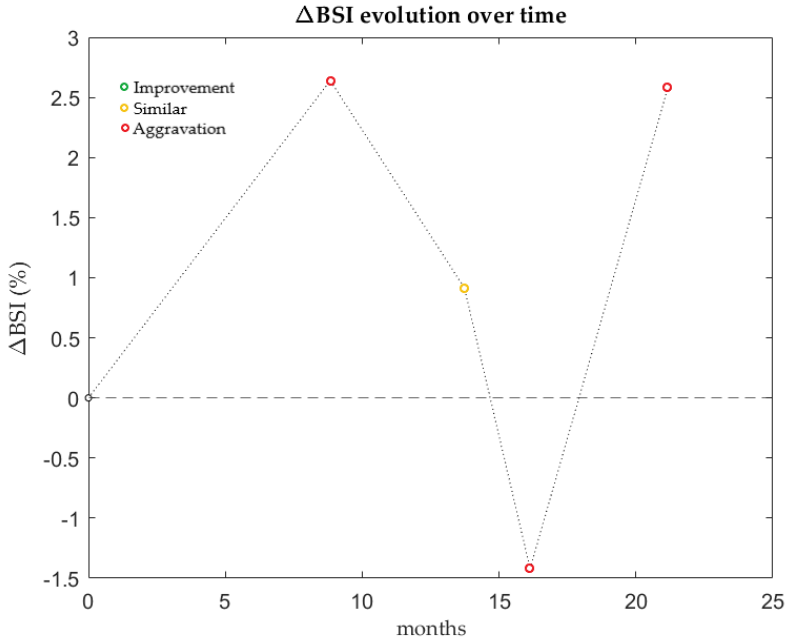


FIGURE 5.18: Variation of the BSI obtained from Patient B bone scintigraphy exams throughout the months. The ordinate of each point corresponds to the difference in BSI score between the current and the previous exams. The colour of the points translate the information given by the respective medical report.

Chapter 6

Conclusions and Future Work

In this work, an algorithm for the automatic quantification of bone scintigraphy images from patients with prostate cancer was developed. Such an algorithm will be extremely useful in the medical community as it will provide the physicians with an aiding tool to quantitatively assess whole-body bone scans from patients with bone metastases, giving them information about disease staging, prognosis and therapy response. Not only will this make it easier for physicians to analyse a bone scan, it will also bring homogeneity within the medical community, as it will reduce the dependency and subjectivity inherent to a bone scan evaluation which is made 100% by visual assessment at the present time. The development of such an algorithm involved three stages: hotspot detection, false-positive attenuation, and BSI calculation.

6.1 Detection Algorithm

Main Conclusions

The developed algorithm for hotspots detection had as main goal the detection of all malignant lesions. This goal was attained: the algorithm achieved a sensitivity of 100% when it comes to detecting metastases. However, despite the high sensitivity for malignant hotspots, it also detected a lot of false-positives, having a false-positive rate of 73%, which corresponded to 32 false-positive detections per image. This was expected, as the algorithm was programmed to find brighter regions in the scans, which can represent anything from healthy physiological processes and benign lesions to metastases. Achieving

a higher specificity with a detection algorithm whose working principle is based solely on the brightness of the hotspots is therefore a very difficult or even impossible task. For this reason, and because the patient condition is determined through the assessment of metastases, two methods for false-positives attenuation were proposed.

6.2 False-positives Attenuation

Two methods for false-positive attenuation were developed.

6.2.1 Method 1

Main Conclusions

The first method used image analysis techniques to remove hotspots that were known *a priori* to be non-malignant. This lead to the removal of several false-positive detections: hotspots found in the bladder, hands and feet, as well as symmetrical hotspots. The algorithm had a specificity of 30%, meaning that 30% of the non-malignant hotspots from the test set were correctly identified as such. The number of false-positive detections per image also suffered a decrease of 30%, dropping from 32 to 22. The achieved sensitivity of 87% shows that, even though almost all malignant hotspots were correctly classified as such, 13% of the metastases of the test set were lost. This mostly happened because some metastases fell under the symmetry conditions and were therefore wrongly classified as false-positives.

Future Work

Improvements on the algorithm include finding symmetry conditions that would keep the specificity at a maximum value (more ability to identify false-positive detections), under the condition that no metastases were being lost (sensitivity = 100%). This could be achieved, for example, by building an optimisation algorithm.

6.2.2 Method 2

Main Conclusions

The first method for false-positives attenuation still left us with a lot of detections that were non-malignant. The second method aimed to reduce even more these false-positives,

by building an algorithm capable of distinguishing between malignant from non-malignant hotspots. Although a few machine learning algorithms for the classification of hotspots from bone scans had already been proposed (Section 3), none of them would tackle the biggest challenge one deals with when building such an algorithm: the lack of a fully labelled data set. Here, we tried to overcome that problem by using algorithms that worked in a completely unsupervised way (k-means algorithm) or that only required knowledge about the type of bone scan from which the hotspots were extracted from (OCC and iterative algorithm). The best model was the iterative algorithm, hotBSI, trained with Support Vector Machine and ResNet18 features, which achieved a sensitivity and false negative rate of 63% and 37%, respectively, compared to 17%/20% and 83%/74% obtained with the best k-means and one-class classification algorithms. With this hotBSI algorithm, the initial false positive rate of 73% obtained with the detection algorithm decrease to 42%. The number of false positive detections per image suffered a decrease of 57%, going from 32 to 14 FPPI. The hotBSI was originally proposed in this work, and has shown to outperform the state-of-the-art algorithms k-means and OCC.

Future Work

Despite showing to be superior to state-of-the-art algorithms, analysis of the performance metrics obtained for the hotBSI shows that this algorithm is still not ready to be used in the clinical practice: the not so high scores for sensitivity (63%), specificity (58%) and AUC (0.66) are still a concern; the false negative rate (37%), despite clearly inferior to the state-of-the-art algorithms, is also still high. Improvements on the algorithm are therefore needed. These include:

- Finding features that are more discriminative, for instance, by using a different pretrained network, by extracting features from different layers or by extracting features from autoencoders;
- Using other classifiers to train the hotBSI;
- Apply variations in the hotBSI, for example, by choosing a stopping criterion in the iteration that is not the number of iterations and fine tune the value of the threshold;
- Retrain the algorithm with a more balanced data set.

An improved classifier will result in a more accurate classification of the hotspots and consequently in a more accurate elimination of false-positives.

6.3 BSI Calculation

Main Conclusions

This work ended by merging the detection and false-positive attenuation algorithms together to calculate the Bone Scan Index for patients in the database who had had bone scintigraphy exams done regularly. Even though some clear outliers could be found during this analysis, as a consequence of flaws in the algorithm, overall it was possible to see a tendency in the evolution of the BSI scores that was in accordance with the medical reports.

Future Work

In the future, it is hoped that instead of a qualitative evaluation of the BSI, a quantitative evaluation can be done. This will require access to a labelled database, so that the real BSI value can be calculated and compared to the one obtained experimentally.

6.4 Overall Conclusions

In this dissertation, an automatic algorithm to assist physicians during bone scans assessment was proposed. The main contributions of this work include:

- Proposal of an algorithm for detection of hotspots in bone scintigraphy images;
- Development of two methods for false-positive attenuation, including a new, iterative semi-supervised algorithm for hotspots classification;
- Extensive experiments on a real data set of scintigraphy images from 102 patients with prostate cancer;
- Quantitative evaluation of the detection and false-positive attenuation algorithms;
- Calculation of an imaging biomarker, the Bone Scan Index, capable of quantifying bone scan images;

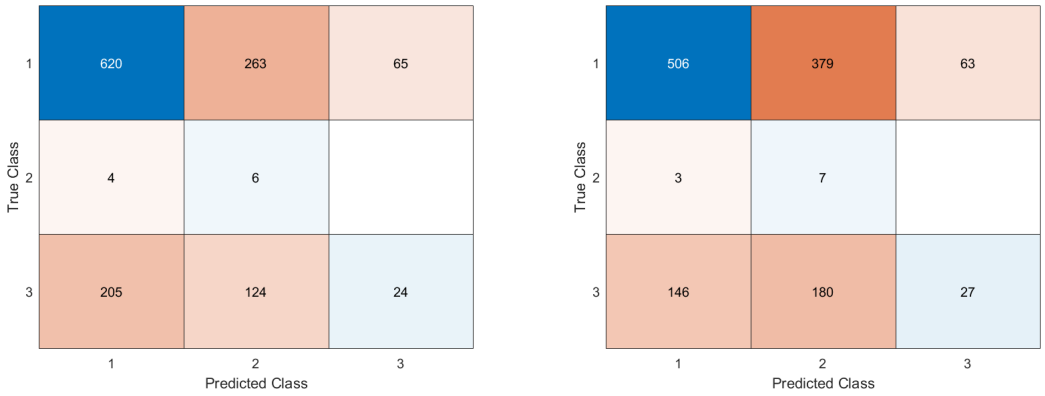
- Study of the evolution of the BSI with two patients from the database.

Once the overall algorithm is improved and a performance that is considered good enough is obtained, it can be used as an aiding-tool by physicians in the medical practice. The final goal is to build a software that can be used in the clinical context, capable of not only quantifying a given bone scintigraphy but also of giving information about disease progression, response to treatment and disease prognosis. Such a software will make the process of assessing a bone scan more objective, simpler and faster, and will for sure be an asset in the medical community.

Appendix A

Appendix: Results of the classification algorithms

In this appendix the results obtained with classification algorithms (Section 5.2.3), along with the respective confusion matrices, are presented.



(A) (A) (B) (B)

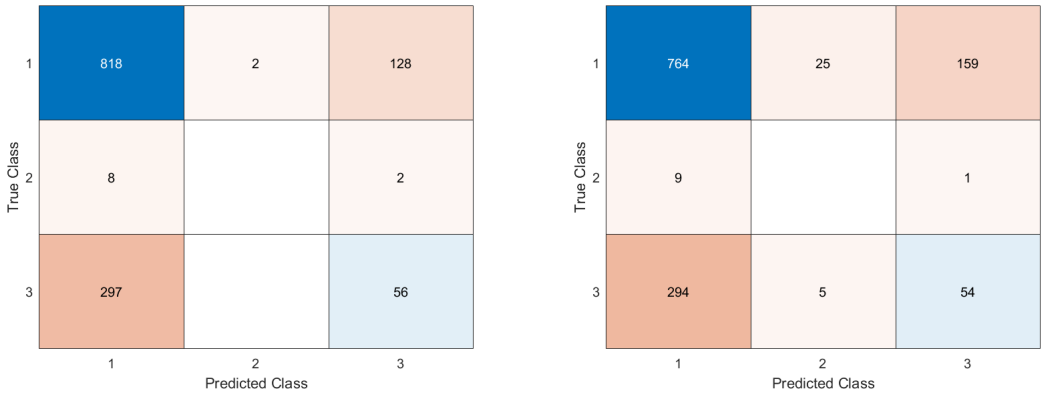
FIGURE A.1: Confusion matrices obtained with the 3-class k-means algorithm trained with handcrafted (A) and ResNet18 (B) features

TABLE A.1: Results obtained with the 3-class k-means algorithm (macro metrics)

Metric	Handcrafted	ResNet18
Sensitivity	0.44	0.44
Specificity	0.68	0.70
Accuracy	0.66	0.61
FPR	0.31	0.30
FNR	0.56	0.56
Precision	0.34	0.36
F1	0.38	0.40

TABLE A.2: Results obtained with the 3-class k-means algorithm

Metric	Handcrafted	ResNet18
Sensitivity	0.07	0.08
Specificity	0.93	0.93
Accuracy	0.70	0.70
FPR	0.07	0.06
FNR	0.93	0.92
Precision	0.27	0.30
F1	0.11	0.12



(A) (B)

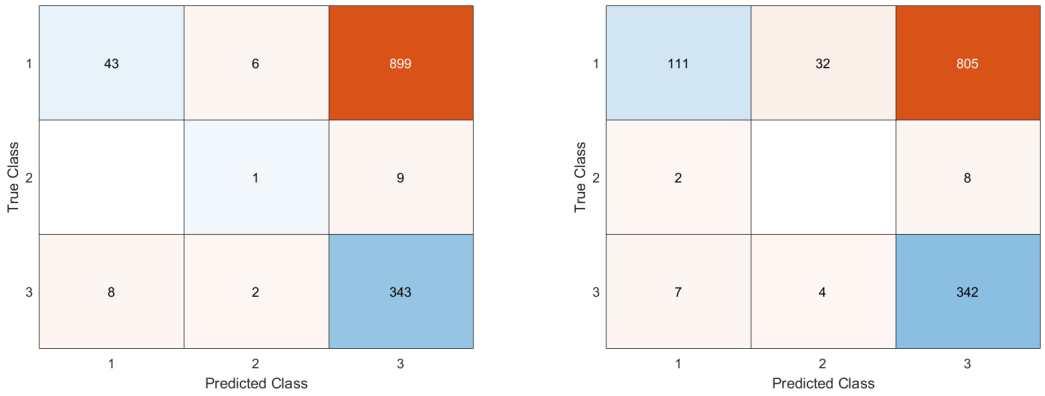
FIGURE A.2: Confusion matrices obtained with 3 class hotBSI-SVM trained with handcrafted (A) and ResNet18 (B) features

TABLE A.3: Macro-averaged metrics obtained with 3 class hotBSI-SVM

Metric	Handcrafted	ResNet18
Sensitivity	0.34	0.32
Specificity	0.67	0.66
Accuracy	0.78	0.74
FPR	0.32	0.34
FNR	0.66	0.68
Precision	0.34	0.32
F1	0.11	0.32

TABLE A.4: "Malign" metrics obtained with 3 class hotBSI-SVM

Metric	Handcrafted	ResNet18
Sensitivity	0.16	0.15
Specificity	0.86	0.83
Accuracy	0.67	0.65
FPR	0.14	0.17
FNR	0.84	0.85
Precision	0.30	0.25
F1	0.21	0.19



(A) (B)

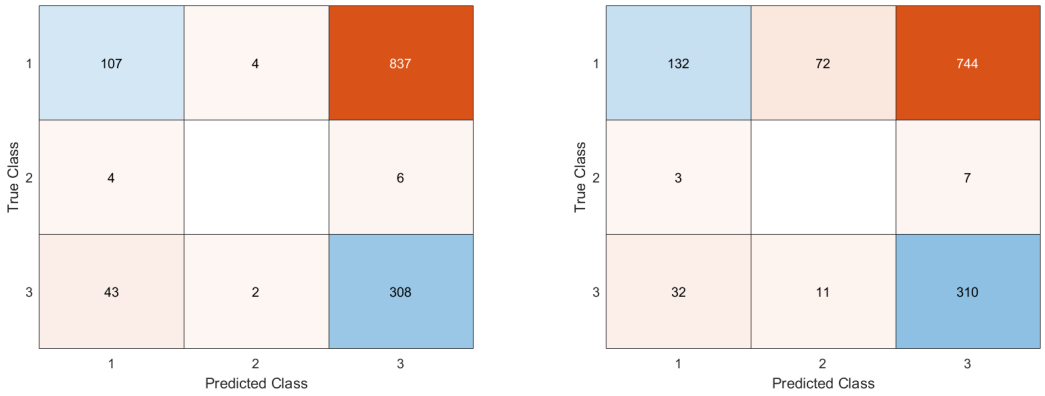
FIGURE A.3: Confusion matrices obtained with 3 class hotBSI-KNN trained with handcrafted (A) and ResNet18 (B) features

TABLE A.5: Macro-averaged metrics obtained with 3 class hotBSI-KNN

Metric	Handcrafted	ResNet18
Sensitivity	0.37	0.36
Specificity	0.67	0.70
Accuracy	0.53	0.56
FPR	0.63	0.30
FNR	0.63	0.64
Precision	0.41	0.41
F1	0.13	0.38

TABLE A.6: "Malign" metrics obtained with 3 class hotBSI-KNN

Metric	Handcrafted	ResNet18
Sensitivity	0.97	0.97
Specificity	0.05	0.15
Accuracy	0.30	0.37
FPR	0.95	0.84
FNR	0.03	0.03
Precision	0.27	0.30
F1	0.43	0.45



(A) (B)

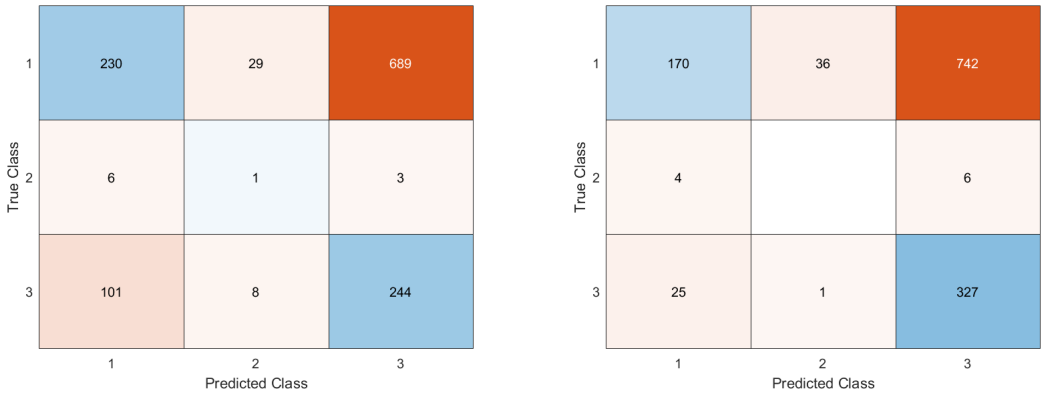
FIGURE A.4: Confusion matrices obtained with 3 class hotBSI-DTs trained with handcrafted (A) and ResNet18 (B) features

TABLE A.7: Macro-averaged metrics obtained with 3 class hotBSI-DTs

Metric	Handcrafted	ResNet18
Sensitivity	0.33	0.34
Specificity	0.66	0.68
Accuracy	0.54	0.56
FPR	0.34	0.31
FNR	0.67	0.66
Precision	0.32	0.36
F1	0.11	0.35

TABLE A.8: "Malign" metrics obtained with 3 class hotBSI-DTs

Metric	Handcrafted	ResNet18
Sensitivity	0.87	0.88
Specificity	0.12	0.22
Accuracy	0.32	0.40
FPR	0.88	0.78
FNR	0.13	0.12
Precision	0.27	0.29
F1	0.41	0.44



(A) (B)

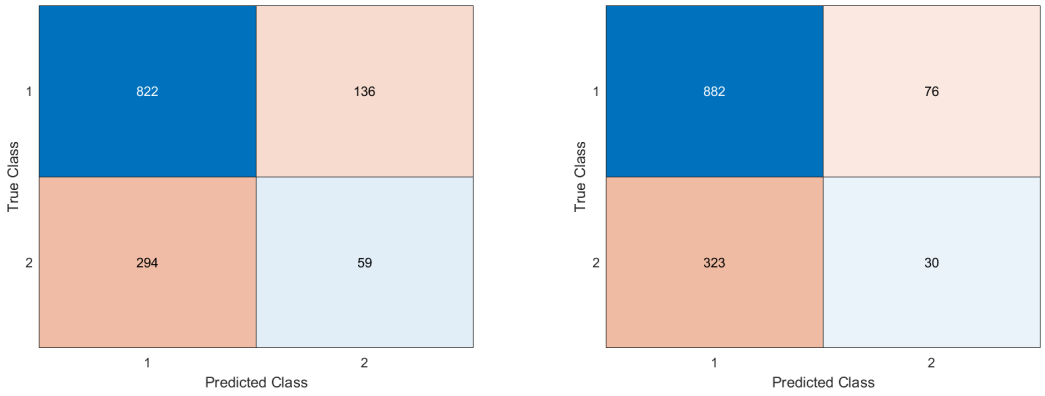
FIGURE A.5: Confusion matrices obtained with 3 class hotBSI-LDA trained with handcrafted (A) and ResNet18 (B) features

TABLE A.9: Macro-averaged metrics obtained with 3 class hotBSI-LDA

Metric	Handcrafted	ResNet18
Sensitivity	0.34	0.37
Specificity	0.65	0.70
Accuracy	0.57	0.59
FPR	0.35	0.30
FNR	0.66	0.63
Precision	0.32	0.39
F1	0.11	0.38

TABLE A.10: "Malign" metrics obtained with 3 class hotBSI-LDA

Metric	Handcrafted	ResNet18
Sensitivity	0.69	0.93
Specificity	0.28	0.22
Accuracy	0.40	0.41
FPR	0.72	0.78
FNR	0.31	0.07
Precision	0.26	0.30
F1	0.38	0.46

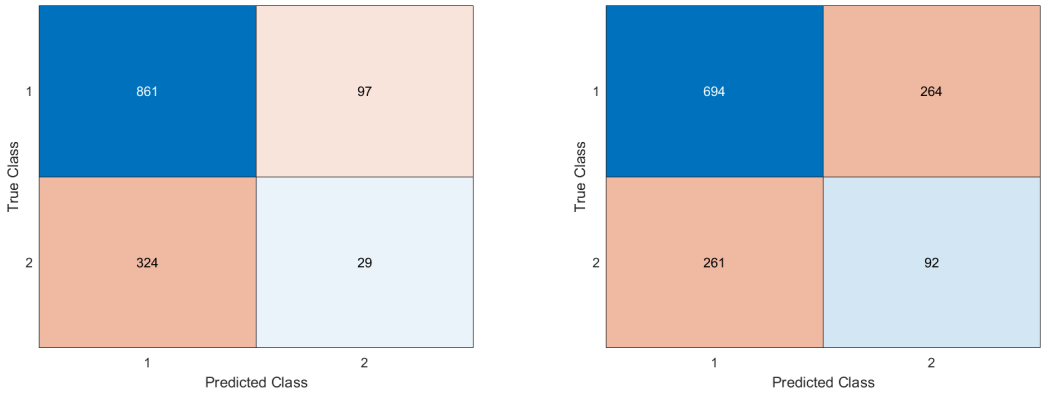


(A) (A) (B) (B)

FIGURE A.6: Confusion matrices obtained with the 2-class k-means algorithm trained with handcrafted (A) and ResNet18 (B) features

TABLE A.11: Results obtained with the k-means algorithm

Metric	Handcrafted	ResNet18
Sensitivity	0.17	0.08
Specificity	0.86	0.92
Accuracy	0.67	0.70
FPR	0.14	0.10
FNR	0.83	0.92
Precision	0.30	0.28
F1	0.21	0.13
FPPI	4	4

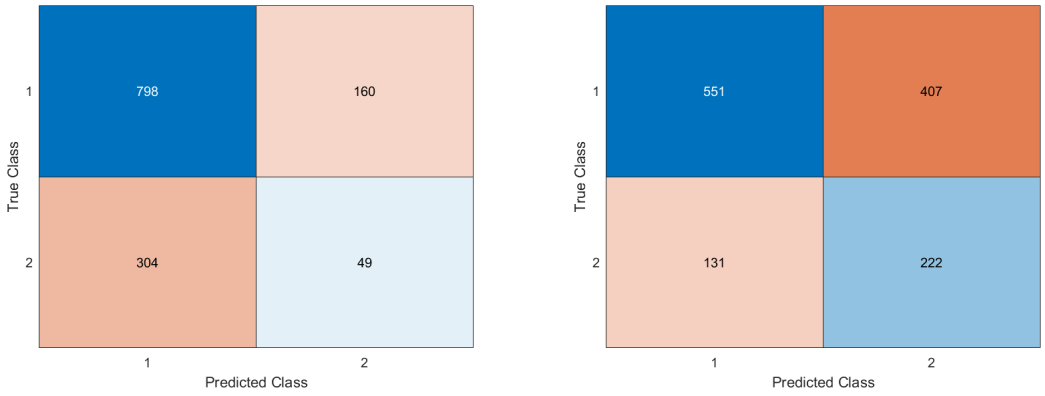


(A) (B)

FIGURE A.7: Confusion matrices obtained with the OCC algorithm trained with handcrafted (A) and ResNet18 (B) features

TABLE A.12: Results obtained with the OCC algorithm

Metric	Handcrafted	ResNet18
Sensitivity	0.08	0.26
Specificity	0.90	0.72
Accuracy	0.68	0.60
FPR	0.10	0.28
FNR	0.92	0.74
Precision	0.23	0.26
F1	0.12	0.12
AUC	0.51	0.50
FPPI	3	9

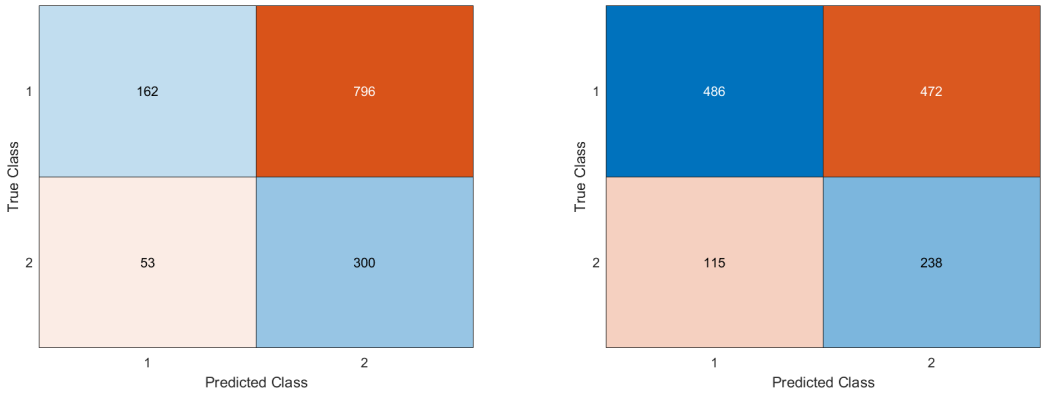


(A) (B)

FIGURE A.8: Confusion matrices obtained with binary hotBSI-SVM trained with handcrafted (A) and ResNet18 (B) features

TABLE A.13: Results obtained with the binary hotBSI-SVM

Metric	Handcrafted	ResNet18
Sensitivity	0.13	0.63
Specificity	0.83	0.58
Accuracy	0.65	0.59
FNR	0.86	0.37
FPR	0.18	0.42
Precision	0.23	0.35
F1	0.17	0.46
AUC	0.50	0.66
FPPI	5	14

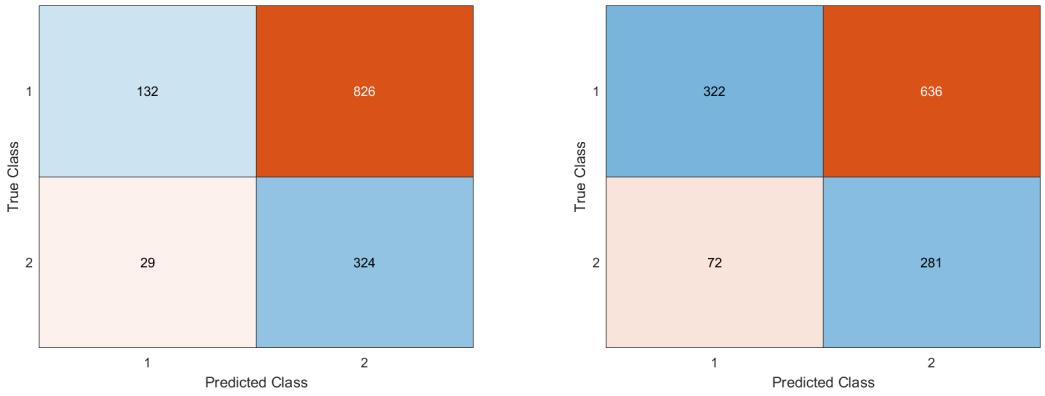


(A) (B)

FIGURE A.9: Confusion matrices obtained with binary hotBSI-KNN trained with handcrafted (A) and ResNet18 (B) features

TABLE A.14: Results obtained with the binary hotBSI-KNN

Metric	Handcrafted	ResNet18
Sensitivity	0.85	0.67
Specificity	0.17	0.51
Accuracy	0.35	0.55
FNR	0.15	0.32
FPR	0.83	0.49
Precision	0.27	0.34
F1	0.41	0.45
AUC	0.52	0.62
FPPI	27	18

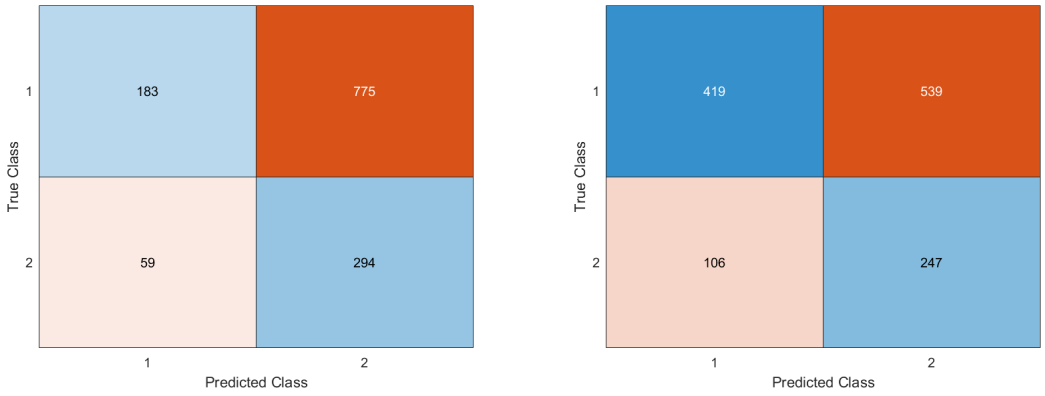


(A) (B)

FIGURE A.10: Confusion matrices obtained with binary hotBSI-DTs trained with handcrafted (A) and ResNet18 (B) features

TABLE A.15: Results obtained with the binary hotBSI-DTs

Metric	Handcrafted	ResNet18
Sensitivity	0.92	0.80
Specificity	0.14	0.33
Accuracy	0.35	0.46
FNR	0.08	0.20
FPR	0.86	0.66
Precision	0.28	0.31
F1	0.43	0.44
AUC	0.46	0.57
FPPI	28	21



(A) (B)

FIGURE A.11: Confusion matrices obtained with binary hotBSI-LDA trained with handcrafted (A) and ResNet18 (B) features

TABLE A.16: Results obtained with the binary hotBSI-LDA

Metric	Handcrafted	ResNet18
Sensitivity	0.83	0.70
Specificity	0.19	0.43
Accuracy	0.36	0.51
FNR	0.17	0.30
FPR	0.81	0.56
Precision	0.28	0.31
F1	0.41	0.43
AUC	0.44	0.59
FPPI	26	18

Bibliography

World Health Organization, "Latest global cancer data:cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018," accessed on 19-12-2020. [Online]. Available: https://www.who.int/cancer/PRGlobocanFinal.pdf [Cited on page 1.]

L. Bubendorf, A. Schöpfer, U. Wagner, G. Sauter, H. Moch, N. Willi, T. C. Gasser, and M. J. Mihatsch, "Metastatic patterns of prostate cancer: An autopsy study of 1,589 patients," *Human Pathology*, vol. 31, no. 5, pp. 578 – 583, 2000. [Cited on pages 1 and 6.]

G. Gandaglia, F. Abdollah, J. Schiffmann, V. Trudeau, S. Shariat, S. Kim, P. Perrotte, F. Montorsi, A. Briganti, Q.-D. Trinh, P. Karakiewicz, and M. Sun, "Distribution of metastatic sites in patients with prostate cancer: A population-based analysis," *The Prostate*, vol. 74, 02 2014. [Cited on pages 1 and 6.]

M. S. Soloway, S. W. Hardeman, D. Hickey, B. Todd, S. Soloway, J. Raymond, and M. Moinuddin, "Stratification of patients with metastatic prostate cancer based on extent of disease on initial bone scan," *Cancer*, vol. 61, no. 1, pp. 195–202, 1988. [Cited on pages 1, 7, and 11.]

M. Norgaard, A. O. Jensen, J. B. Jacobsen, K. Cetin, J. P. Fryzek, and H. T. Sørensen, "Skeletal related events, bone metastasis and survival of prostate cancer: A population based cohort study in denmark (1999 to 2007)," *Journal of Urology*, vol. 184, no. 1, p. 162–167, 2010. [Cited on pages 1 and 7.]

IJUP, "Quantification of whole-body bone scans with imaging processing and machine learning algorithms," 2021, accessed on 24.06.2021. [Online]. Available: https://ijup.up.pt/2021/ [Cited on page 3.]

ENJIO, "Algoritmo para a quantificação automática de cintigrafias ósseas de pacientes com cancro da próstata," 2021, accessed on 21.09.2021. [Online]. Available: https://www.ligacontracancro.pt/enjio/ [Cited on page 3.]

L. Providência, I. Domingues, and J. Santos, "An iterative algorithm for semisupervised classification of hotspots on bone scintigraphies of patients with prostate cancer," *Journal of Imaging*, vol. 7, no. 8, 2021. [Cited on page 4.]

——, "False-positives attenuation of automatically detected hotspots on bone scintigraphy images using image analysis techniques," 2021, submitted. [Cited on page 4.]

World Health Organization, "Data visualization tools for exploring the global cancer burden in 2020," accessed on 09-03-2021. [Online]. Available: https://gco.iarc.fr/today/home [Cited on page 5.]

E. A. Klein and J. S. Jones, *Management of prostate cancer*. Springer, 2013. [Cited on page 5.]

Cancer Net, "Prostate cancer - types of treatment," Jan 2020, accessed on 16-03-2021. [Online]. Available: https://www.cancer.net/cancer-types/prostate-cancer/types-treatment [Cited on page 6.]

J. F. Worthington, "Focal therapy for prostate cancer: If it sounds too good to be true..." accessed on 16-03-2021. [Online]. Available: https://www.pcf.org/c/focal-therapy-for-prostate-cancer-if-it-sounds-too-good-to-be-true/ [Cited on page 6.]

P. Msaouel, N. Pissimissis, A. Halapas, and M. Koutsilieris, "Mechanisms of bone metastasis in prostate cancer: clinical implications," *Best Practice & Research Clinical Endocrinology & Metabolism*, vol. 22, no. 2, pp. 341 – 355, 2008, endocrinology and the Prostate. [Cited on page 6.]

A. I. Brenner, J. Koshy, J. Morey, C. Lin, and J. DiPoce, "The bone scan," *Seminars in Nuclear Medicine*, vol. 42, no. 1, pp. 11 – 26, 2012, planar Imaging in the Age of SPECT. [Cited on page 7.]

M. Ohta, Y. Tokuda, Y. Suzuki, M. Kubota, H. Makuuchi, T. Tajima, S. Nasu, S. Yasuda, and A. Shohtsu, "Whole body PET for the evaluation of bony metastases in patients with breast cancer: comparison with 99Tcm-MDP bone scintigraphy," *Nuclear medicine communications*, vol. 22, pp. 875–9, 09 2001. [Cited on page 7.]

E. Even-Sapir, U. Metser, E. Mishani, G. Lievshitz, H. Lerman, and I. Leibovitch, "The detection of bone metastases in patients with high-risk prostate cancer: 99mTc-MDP planar bone scintigraphy, single- and multi-field-of-view SPECT, 18f-fluoride PET, and 18F-fluoride PET/CT," *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, vol. 47, no. 2, p. 287—297, February 2006. [Cited on page 7.]

G. J. O'Sullivan, "Imaging of bone metastasis: An update," *World Journal of Radiology*, vol. 7, no. 8, p. 202, 2015. [Cited on pages 7 and 9.]

D. Hadjidakis and I. Androulakis, "Bone remodeling," *Annals of the New York Academy of Sciences*, vol. 1092, pp. 385–96, 01 2007. [Cited on page 7.]

N. Chun-Yi, "What is the difference between 'bone turnover' and 'bone remodeling'?" Sep 2020. [Online]. Available: https://www.researchgate.net/post/What_is_the_difference_between_bone_turnover_and_bone_remodeling [Cited on page 7.]

X. Feng, "Chemical and biochemical basis of cell-bone matrix interaction in health and disease," *Current chemical biology*, vol. 3, pp. 189–196, 05 2009. [Cited on page 8.]

J. Jeong, K. Jung Hun, J. Shim, N. Hwang, and C.-Y. Heo, "Bioactive calcium phosphate materials and applications in bone regeneration," *Biomaterials Research*, vol. 23, 12 2019. [Cited on page 8.]

F. A. Mettler and M. J. Guiberteau, *Essentials of nuclear medicine and molecular imaging*. Elsevier, 2019. [Cited on page 8.]

J. Purden, "Nuclear medicine 2: principles and technique of bone scintigraphy," *Nursing Times*, vol. 115, no. 4, p. 48–49, 2019. [Cited on page 8.]

E. Even-Sapir, U. Metser, E. Mishani, G. Lievshitz, H. Lerman, and I. Leihovitch, "1553: The detection of bone metastases by 99Tcm-MDP planar bone scintigraphy, single and multi-field-of-views SPECT, 18F-Fluoride PET and 18f-fluoride PET-CT, prospective study in 44 patients with high-risk prostate cancer," *Journal of Urology*, vol. 175, no. 4S, p. 501, 2006. [Cited on page 9.]

World Health Organization and International Programme on Chemical Safety, "Biomarkers in risk assessment : validity and validation," pp. 235–238, 2001. [Cited on page 10.]

National Cancer Institute, "Prostate-specific antigen (PSA) test," accessed on 23-02-2021. [Online]. Available: https://www.cancer.gov/types/prostate/psa-fact-sheet [Cited on page 10.]

——, "Prostate-specific antigen (PSA) test," accessed 23-02-2021. [Online]. Available: https://www.cancer.gov/types/prostate/psa-fact-sheet [Cited on page 10.]

N. Mustansar, "Utility of bone scan quantitative parameters for the evaluation of prostate cancer patients," *Journal of Nuclear Medicine & Radiation Therapy*, vol. 09, 01 2018. [Cited on pages 11, 12, and 54.]

M. Noguchi, H. Kikuchi, M. Ishibashi, and S. Noda, "Percentage of the positive area of bone metastasis is an independent predictor of disease death in advanced prostate cancer," *British journal of cancer*, vol. 88, pp. 195–201, 02 2003. [Cited on page 12.]

E. Dennis, X. Jia, I. Mezheritskiy, R. Stephenson, H. Schoder, J. Fox, G. Heller, H. Scher, S. Larson, and M. Morris, "Bone scan index: A quantitative treatment response biomarker for castration-resistant metastatic prostate cancer," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 30, pp. 519–24, 02 2012. [Cited on page 12.]

D. Li, H. Lv, X. Hao, Y. Dong, H. Dai, and Y. Song, "Prognostic value of bone scan index as an imaging biomarker in metastatic prostate cancer: a meta-analysis," *Oncotarget*, vol. 8, no. 48, pp. 84 449–84 458, 2017. [Cited on pages 12 and 56.]

R. Kaboteh, J.-E. Damber, P. Gjertsson, P. Höglund, M. Lomsky, M. Ohlsson, and L. Edenbrandt, "Bone scan index: A prognostic imaging biomarker for high-risk prostate cancer patients receiving primary hormonal therapy," *EJNMMI research*, vol. 3, p. 9, 02 2013. [Cited on pages 12 and 55.]

W. S. Snyder, *Report of the Task Group on Reference Man: a report prepared by a task group of Committee 2 of the International Commission on Radiological Protection*. Pergamon Press, 1981. [Cited on page 12.]

I. Ito, K. Ito, S. Takahashi, M. Horibe, R. Karita, C. Nishizaka, T. Nagai, K. Hamada, H. Sato, and N. Shindo, "Association between bone scan index and activities of daily living in patients with advanced non-small cell lung cancer," *Supportive Care in Cancer*, vol. 25, pp. 1779–1785, 2016. [Cited on page 13.]

A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959. [Cited on page 13.]

A. Burkov, *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019. [Online]. Available: https://books.google.pt/books?id=0jbxwQEACAAJ [Cited on pages 13 and 15.]

T. M. Mitchell, *Machine Learning*, 1st ed. USA: McGraw-Hill, Inc., 1997. [Cited on page 14.]

R. Bhatia, "Top 6 regression algorithms used in analytics & data mining," Sep 2017, accessed on 23-03-2021. [Online]. Available: https://analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/ [Cited on page 15.]

U. R. Hodeghatta and U. Nayak, *Business analytics using R - A practical approach*. Apress, 2017. [Cited on page 16.]

D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *SODA '07: Proceedings of the eighteenth annual ACM-S*, vol. 8. Society for Industrial and Applied Mathematics, 01 2007, pp. 1027–1035. [Cited on page 17.]

L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965. [Cited on page 18.]

M. Choudhry and R. Kapoor, "Performance analysis of fuzzy c-means clustering methods for MRI image segmentation," *Procedia Computer Science*, vol. 89, pp. 749–758, 12 2016. [Cited on page 18.]

M. Huang, Z. Xia, H. Wang, Q. Zeng, and Q. Wang, "The range of the value for the fuzzifier of the fuzzy c-means algorithm," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2280–2284, 2012. [Cited on page 18.]

F. Klawonn and F. Höppner, "What is fuzzy about fuzzy clustering? understanding and improving the concept of the fuzzifier," *Advances in Intelligent Data Analysis V Lecture Notes in Computer Science*, p. 254–264, 2003. [Cited on page 18.]

N. S. Chauhan, "What is hierarchical clustering?" 2019, accessed on 26-03-2021. [Online]. Available: https://www.kdnuggets.com/2019/09/hierarchical-clustering.html [Cited on page 19.]

X. Zhu and A. Goldberg, *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers, 01 2009, vol. 3. [Cited on page 20.]

I. Domingues and J. S. Cardoso, "Max-ordinal learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 7, pp. 1384–1389, 2014. [Cited on page 20.]

R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*. Prentice Hall, 2007. [Cited on pages 21 and 52.]

A. Carreño, I. Inza, and J. Lozano, "Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework," *Artificial Intelligence Review*, vol. 53, 06 2020. [Cited on page 21.]

Q. Wei, Y. Ren, R. Hou, B. Shi, J. Y. Lo, and L. Carin, "Anomaly detection for medical images based on a one-class classification," in *Medical Imaging 2018: Computer-Aided Diagnosis*, N. Petrick and K. Mori, Eds., vol. 10575, International Society for Optics and Photonics. SPIE, 2018, pp. 375 – 380. [Cited on page 21.]

O. Mazhelis, "One-class classifiers: A review and analysis of suitability in the context of mobile-masquerader detection," *South African Computer Journal*, vol. 36, pp. 29–48, 01 2006. [Cited on page 21.]

S. S. Khan and M. G. Madden, "One-class classification: taxonomy of study and review of techniques," *The Knowledge Engineering Review*, vol. 29, no. 3, p. 345–374, 2014. [Cited on pages 21, 22, 23, and 25.]

B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," in *Advances in Neural Information Processing Systems 12*, Max-Planck-Gesellschaft. MIT Press, 06 2000, pp. 582–588. [Cited on pages 22, 23, and 77.]

G. Zhang, "What is the kernel trick? Why is it important?" Nov 2018, accessed on 31-03-2021. [Online]. Available: https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-important-98a98db0961d [Cited on page 23.]

L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and

A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 4393–4402. [Cited on pages 23 and 25.]

N. Japkowicz, C. Myers, and M. Gluck, "A novelty detection approach to classification," *Proceedings of the Fourteenth Joint Conference on Artificial Intelligence*, 10 1999. [Cited on page 24.]

M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *MLSDA'14: Machine Learning for Sensory Data Analysis*, 2014. [Cited on page 24.]

S. Chaurasia, S. Goyal, and M. Rajput, "Outlier detection using autoencoder ensembles: A robust unsupervised approach," in *2020 International Conference on Contemporary Computing and Applications (IC3A)*, 2020, pp. 76–80. [Cited on page 24.]

T. Schlegl, P. Seeböck, S. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International Conference on Information Processing in Medical Imaging*, 03 2017, pp. 146–157. [Cited on page 24.]

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Cited on page 24.]

N. Sebe, *Machine learning in computer vision*. Springer, 2005. [Cited on page 26.]

A. F. A. Fernandes, J. R. R. Dórea, and G. J. d. M. Rosa, "Image analysis and computer vision applications in animal sciences: An overview," *Frontiers in Veterinary Science*, vol. 7, 2020. [Cited on page 26.]

D. H. Ballard and C. M. Brown, *Computer Vision*, 1st ed. Prentice Hall Professional Technical Reference, 1982. [Cited on page 26.]

J. Gao, Y. Yang, P. Lin, and D. Park, "Computer vision in healthcare applications," *Journal of Healthcare Engineering*, vol. 2018, pp. 1–4, 03 2018. [Cited on page 26.]

R. Bala and R. P. Loce, *Introduction*. John Wiley & Sons, Ltd, 2017. [Cited on page 26.]

D. I. Patrício and R. Rieder, "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review," *Computers and Electronics in Agriculture*, vol. 153, pp. 69–81, 2018. [Cited on page 26.]

A. Bohr and K. Memarzadeh, "Chapter 2 - the rise of artificial intelligence in healthcare applications," in *Artificial Intelligence in Healthcare*, A. Bohr and K. Memarzadeh, Eds. Academic Press, 2020, pp. 25–60. [Cited on page 26.]

J. Le, "The 5 computer vision techniques that will change how you see the world," Jan 2020, access on 05-04-2021. [Online]. Available: https://heartbeat.fritz.ai/the-5-computer-vision-techniques-that-will-change-how-you-see-the-world-1ee19334354b [Cited on page 27.]

B. P. K. Reddy and A. Chatterjee, "Encrypted classification using secure k-nearest neighbour computation," in *Security, Privacy, and Applied Cryptography Engineering*, S. Bhasin, A. Mendelson, and M. Nandi, Eds. Cham: Springer International Publishing, 2019, pp. 176–194. [Cited on page 27.]

L. Kabbai, M. Abdellaoui, and A. Douik, "Image classification by combining local and global features," *The Visual Computer*, vol. 35, 05 2019. [Cited on page 28.]

ImageNet, "Imagenet large scale visual recognition challenge (ilsvrc)." [Online]. Available: http://image-net.org/challenges/LSVRC/ [Cited on page 28.]

A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Cited on page 28.]

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2014. [Cited on page 28.]

K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, 09 2015. [Cited on page 28.]

K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Cited on page 28.]

S. Siddiqui, I. Malik, F. Shafait, A. Mian, M. Shortis, and E. Harvey, "Automatic fish species classification in underwater videos: Exploiting pretrained deep neural network models to compensate for limited labelled data," *ICES Journal of Marine Science*, vol. 75, 05 2017. [Cited on page 29.]

I. Tabian, H. Fu, and Z. Sharif Khodaei, "A convolutional neural network for impact detection and characterization of complex composite structures," *Sensors*, vol. 19, no. 22, 2019. [Cited on page 30.]

N. Chawla, N. Japkowicz, and A. Kołcz, "Editorial: Special issue on learning from imbalanced data sets," *SIGKDD Explorations*, vol. 6, pp. 1–6, 06 2004. [Cited on page 31.]

F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Mach. Learn.*, vol. 42, no. 3, Mar. 2001. [Cited on pages 31 and 34.]

W. S. Lee and B. Liu, "Learning with positive and unlabeled examples using weighted logistic regression," in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*. AAAI Press, 2003, p. 448–455. [Cited on page 36.]

M. Brown, G. Chu, G. H. Kim, M. Allen-Auerbach, C. Poon, J. Bridges, A. Vidovic, B. Ramakrishna, J. Ho, M. Morris, S. Larson, H. Scher, and J. Goldin, "Computer-aided quantitative bone scan assessment of prostate cancer treatment response," *Nuclear medicine communications*, vol. 33, pp. 384–94, 04 2012. [Cited on pages 39, 40, 43, 45, 47, and 53.]

M. Koenigkam Santos, J. Ferreira Junior, D. Wada, A. Tenório, M. Nogueira-Barbosa, and P. Azevedo-Marques, "Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine," *Radiologia Brasileira*, vol. 52, 09 2019. [Cited on pages 39 and 42.]

M. Brown, G. Kim, G. Chu, B. Ramakrishna, M. Allen-Auerbach, C. P. Fischer, B. Levine, P. Gupta, C. Schiepers, and J. Goldin, "Quantitative bone scan lesion area as an early surrogate outcome measure indicative of overall survival in metastatic prostate cancer," *Journal of Medical Imaging*, vol. 5, 2018. [Cited on pages 40 and 67.]

A. Shimizu, H. Wakabayashi, T. Kanamori, A. Saito, K. Nishikawa, H. Daisaki, S. Higashiyama, and J. Kawabe, "Automated measurement of bone scan index from a whole-body bone scintigram," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, pp. 1–12, 12 2019. [Cited on pages 40, 44, and 45.]

J.-Y. Huang, P.-F. Kao, and Y.-S. Chen, "A set of image processing algorithms for computer-aided diagnosis in nuclear medicine whole body bone scan images," *Nuclear Science, IEEE Transactions on*, vol. 54, pp. 514 – 522, 07 2007. [Cited on pages 40, 41, 43, 45, and 67.]

N. Papandrianos, E. Papageorgiou, A. Anagnostis, and A. Feleki, "A deep-learning approach for diagnosis of metastatic breast cancer in bones from whole-body scans," *Applied Sciences*, vol. 10, no. 3, p. 997, Feb 2020. [Cited on pages 41, 49, 50, and 53.]

N. Papandrianos, E. Papageorgiou, A. Anagnostis, and K. Papageorgiou, "Bone metastasis classification using whole body images from prostate cancer patients based on convolutional neural networks application," *PLOS ONE*, vol. 15, no. 8, pp. 1–28, 08 2020. [Cited on pages 41, 49, 50, and 53.]

——, "Efficient bone metastasis diagnosis in bone scintigraphy using a fast convolutional neural network architecture," *Diagnostics*, vol. 10, p. 532, 07 2020. [Cited on pages 41, 49, 50, and 53.]

J. Dang, "Classification in bone scintigraphy images using convolutional neural networks," Master's thesis, Lund University, Sweden, 2016. [Cited on pages 41, 51, and 53.]

L. Belcher, "Convolutional neural networks for classification of prostate cancer metastases using bone scan images," 2017, student Paper. [Cited on pages 41, 51, and 53.]

Y. Guo and A. S. Ashour, "11 - neutrosophic sets in dermoscopic medical image segmentation," in *Neutrosophic Set in Medical Image Analysis*, Y. Guo and A. S. Ashour, Eds. Academic Press, 2019, pp. 229 – 243. [Cited on page 42.]

A. Aslantas, M. Çakıroğlu, and E. Dandıl, "Comparison of segmentation algorithms for detection of hotspots in bone scintigraphy images and effects on CAD systems," *Biomedical Research*, vol. 28, pp. 676–683, 01 2017. [Cited on page 42.]

M. Sadik, D. Jakobsson, F. Olofsson, M. Ohlsson, M. Suurkula, and L. Edenbrandt, "A new computer-based decision-support system for the interpretation of bone scans," *Nuclear medicine communications*, vol. 27, pp. 417–23, 06 2006. [Cited on pages 42, 45, 47, 48, 49, 50, and 53.]

M. Sadik, I. Hamadeh, P. Nordblom, M. Suurkula, P. Höglund, M. Ohlsson, and L. Edenbrandt, "Computer-assisted interpretation of planar whole-body bone scans," *Journal of*

nuclear medicine : official publication, Society of Nuclear Medicine, vol. 49, pp. 1958–65, 12 2008. [Cited on pages 42, 45, 47, 48, and 53.]

O. Tobias and R. Seara, "Image segmentation by histogram thresholding using fuzzy sets," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 11, pp. 1457–65, 02 2002. [Cited on page 43.]

M. Ohlsson, R. Kaboteh, M. Sadik, M. Suurkula, M. Lomsky, P. Gjertsson, K. Sjostrand, J. Richter, and L. Edenbrandt, "Automated decision support for bone scintigraphy," in *2009 22nd IEEE International Symposium on Computer-Based Medical Systems*, 2009, pp. 1–6. [Cited on pages 43, 45, and 47.]

A. Aslantaş, E. Dandıl, and M. Çakıroğlu, "Cadboss: A computer-aided diagnosis system for whole-body bone scintigraphy scans," *Journal of Cancer Research and Therapeutics*, vol. 12, 04 2016. [Cited on pages 43, 45, and 47.]

C. Li, C.-Y. Kao, J. Gore, and Z. Ding, "Implicit active contours driven by local binary fitting energy," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1-7, 07 2007, pp. 1–7. [Cited on page 44.]

C. Xu, "Common methods for feature extraction: Pca and lda," Jun 2018. [Online]. Available: https://medium.com/@cxu24/common-methods-for-feature-extraction-pca-and-lda-7b1f5679e3bf [Cited on page 46.]

S. Crommelinck, R. Bennett, M. Gerke, F. Nex, M. Y. Yang, and G. Vosselman, "Review of automatic feature extraction from high-resolution optical sensor data for UAV-based cadastral mapping," *Remote Sensing*, vol. 8, p. 689, 09 2016. [Cited on page 46.]

G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60 – 88, 2017. [Cited on pages 46 and 49.]

D. Kumar, A. Wong, and D. Clausi, "Lung nodule classification using deep features in CTimages," *Proceedings -2015 12th Conference on Computer and Robot Vision, CRV 2015*, pp. 133–138, 07 2015. [Cited on page 46.]

R. Cohn and E. Holm, "Unsupervised machine learning via transfer learning and k-means clustering to classify materials image data," *ArXiv*, vol. abs/2007.08361, 2020. [Cited on page 46.]

M. Alaslani and L. Elrefaei, "Convolutional neural network based feature extraction for iris recognition," *International Journal of Computer Science and Information Technology*, vol. 10, pp. 65–78, 04 2018. [Cited on page 46.]

M. Khan, M. Javed, M. Sharif, T. Saba, and A. Rehman, "Multi-model deep neural network based features extraction and optimal selection approach for skin lesion classification," in *2019 International Conference on Computer and Information Sciences (ICCIS)*, 04 2019, pp. 1–7. [Cited on page 46.]

R. Kumar and M. V. Suhas, "Classification of benign and malignant bone lesions on CTimagesusing support vector machine: A comparison of kernel functions," in *2016 IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, 2016, pp. 821–824. [Cited on page 48.]

B. Ginneken and R. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1153–1159, 05 2016. [Cited on page 49.]

EXINI Diagnostics AB, Nov 2020, accessed on 11.12.2020. [Online]. Available: https://exini.com/ [Cited on page 50.]

aBSI, "510(k) premarket submission to U.S. Food & Drug Administration," https://www.accessdata.fda.gov/cdrh_docs/pdf19/K191262.pdf, 2019, accessed on 12-12-2020. [Cited on page 50.]

D. Ulmert, R. Kaboteh, J. J. Fox, C. Savage, M. J. Evans, H. Lilja, P.-A. Abrahamsson, T. Björk, A. Gerdtsson, A. Bjartell, P. Gjertsson, P. Höglund, M. Lomsky, M. Ohlsson, J. Richter, M. Sadik, M. J. Morris, H. I. Scher, K. Sjöstrand, A. Yu, M. Suurküla, L. Edenbrandt, and S. M. Larson, "A novel automated platform for quantifying the extent of skeletal tumour involvement in prostate cancer patients using the bone scan index," *European Urology*, vol. 62, no. 1, pp. 78 – 84, 2012. [Cited on pages 51 and 55.]

H. Horikoshi, A. Kikuchi, M. Onoguchi, K. Sjöstrand, and L. Edenbrandt, "Computer-aided diagnosis system for bone scintigrams from Japanese patients: Importance of training database," *Annals of nuclear medicine*, vol. 26, pp. 622–6, 06 2012. [Cited on page 51.]

E. Ahn, A. Kumar, D. Feng, M. Fulham, and J. Kim, "Unsupervised feature learning with k-means and an ensemble of deep convolutional neural networks for medical image classification," *ArXiv*, vol. abs/1906.03359, 2019. [Cited on page 52.]

K. Polat, "Classification of Parkinson's disease using feature weighting method on the basis of fuzzy c-means clustering," *International Journal of Systems Science*, vol. 43, no. 4, pp. 597–609, 2012. [Cited on page 52.]

C.-H. Chen, "A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection," *Applied Soft Computing*, vol. 20, pp. 4–14, 2014, hybrid intelligent methods for health technologies. [Cited on page 52.]

H. Alashwal, M. El Halaby, J. J. Crouse, A. Abdalla, and A. A. Moustafa, "The application of unsupervised clustering methods to Alzheimer's disease," *Frontiers in Computational Neuroscience*, vol. 13, p. 31, 2019. [Cited on page 52.]

Z. Alaverdyan, "Unsupervised representation learning for anomaly detection on neuroimaging. application to epilepsy lesion detection on brain MRI," Ph.D. dissertation, Université de Lyon, 01 2019. [Cited on pages 52 and 53.]

M. El Azami, A. Hammers, J. Jung, N. Costes, R. Bouet, and C. Lartizien, "Detection of lesions underlying intractable epilepsy on T1-weighted MRI as an outlier detection problem," *PLOS ONE*, vol. 11, no. 9, pp. 1–21, 09 2016. [Cited on page 53.]

A. Gardner, A. Krieger, G. Vachtsevanos, and B. Litt, "One-class novelty detection for seizure analysis from intracranial EEG." *Journal of Machine Learning Research*, vol. 7, pp. 1025–1044, 06 2006. [Cited on page 53.]

E. Spinosa and A. Carvalho, "Support vector machines for novel class detection in bioinformatics," *Genetics and molecular research : GMR*, vol. 4, pp. 608–15, 02 2005. [Cited on page 53.]

C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain mr images," in *Lecture Notes in Computer Science*. Springer International Publishing, 04 2018. [Cited on page 53.]

X. Chen, N. Pawlowski, B. Glocker, and E. Konukoglu, "Unsupervised lesion detection with locally gaussian approximation," in *Machine Learning in Medical Imaging@MICCAI*, 2019. [Cited on page 53.]

A. Moradi, S. Srinivasan, J. Clements, and J. Batra, "Beyond the biomarker role: prostate-specific antigen (PSA) in the prostate cancer microenvironment," *Cancer and Metastasis Reviews*, vol. 38, 09 2019. [Cited on page 54.]

M. H. Poulsen, J. Rasmussen, L. Edenbrandt, P. F. Høilund-Carlsen, O. Gerke, A. Johansen, and L. Lund, "Bone scan index predicts outcome in patients with metastatic hormone-sensitive prostate cancer," *BJU International*, vol. 117, no. 5, pp. 748–753, 2016. [Cited on page 55.]

American Cancer Society, "Hormone therapy for prostate cancer," accessed on 18-12-2020. [Online]. Available: https://www.cancer.org/cancer/prostate-cancer/treating/hormone-therapy.html [Cited on page 55.]

M. Reza, A. Bjartell, M. Ohlsson, R. Kaboteh, P. Wollmer, L. Edenbrandt, and E. Trägårdh, "Bone scan index as a prognostic imaging biomarker during androgen deprivation therapy," *EJNMMI research*, vol. 4, p. 58, 10 2014. [Cited on page 55.]

A. Inaki, K. Nakajima, H. Wakabayashi, T. Mochizuki, and S. Kinuya, "Fully automated analysis for bone scintigraphy with artificial neural network: usefulness of bone scan index (BSI) in breast cancer," *Annals of Nuclear Medicine*, vol. 33, 07 2019. [Cited on page 55.]

S. Fotso *et al.*, "PySurvival: Open source package for survival analysis modeling," 2019, accessed on 21-12-2020. [Online]. Available: https://square.github.io/pysurvival/metrics/c_index.html [Cited on page 56.]

I. Domingues and J. S. Cardoso, "Using Bayesian surprise to detect calcifications in mammogram images," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 1091–1094. [Cited on page 62.]

M. Cicconet, D. G. C. Hildebrand, and H. Elliott, "Finding mirror symmetry via registration and optimal symmetric pairwise assignment of curves," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 1749–1758. [Cited on page 71.]

GitHub, "SymmetryViaRegistration," accessed on 16.07.2021. [Online]. Available: https://github.com/cicconet/SymmetryViaRegistration [Cited on page 71.]

A. M. Silva, E. Crubézy, and E. Cunha, "Bone weight: new reference values based on a modern portuguese identified skeletal collection," *International Journal of Osteoarchaeology*, vol. 19, no. 5, pp. 628–641, 2009. [Cited on page 84.]