FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Learning models for bone marrow edema characterization in radiological images

Gonçalo José Marques Ribeiro

Mestrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Hélder Oliveira

Co-Supervisor: Tânia Pereira

Co-Supervisor: Sílvia Costa Dias

August 21, 2022

Abstract

Bone marrow edema (BME) is the generic term applied to bone marrow signal intensity changes in magnetic resonance imaging (MRI). More recently, the term edema-like marrow signal intensity (ELMSI) replaces the term BME. This condition is often detected in the knee, hip, foot, ankle and shoulder, and is characterized by areas of intermediate to low signal intensities on T1-weighted images, and areas of high signal intensities on fluid-sensitive sequences. The detection of ELMSI by clinicians on X-rays is currently considered unreliable. The identification of areas of ELMSI can be time consuming and subject to reader variability. Nowadays, the use of deep learning techniques has helped finding valid solutions to complex problems, notably in medical imaging. Diagnosing ELMSI can be time consuming, and can depend on the physician conducting the MRI, the device, and the reader. Furthermore, there are no guidelines for quantitative measurement. In children and young adults, natural changes that occur in the bone marrow can cause visible alterations in MRIs. Therefore, applying deep learning techniques to predict ELMSI on young patients can have vast advantages. In this work, a Convolutional Neural Network (CNN) will be trained to classify ELMSI slices on a dataset containing children and young adults. Data augmentation and transfer learning will be used to reduce over-fitting and improve generalization, as well as pre-processing techniques to improve the performance of the model.

Resumo

Edema da medula óssea (BME) é o termo genérico aplicado às alterações da intensidade do sinal da medula óssea em ressonância magnética (MRI). Mais recentemente, o termo intensidade de sinal da medula óssea (ELMSI) substitui o termo BME. Esta condição é frequentemente detectada no joelho, anca, pé, tornozelo e ombro, e é caracterizada por áreas de intensidades de sinal intermédias a baixas nas imagens ponderadas em T1, e áreas de intensidades de sinal altas nas sequências sensíveis a fluidos. A detecção de edema de medula óssea por clínicos em X-rays é actualmente considerada pouco fiável. A identificação de áreas de ELMSI pode ser demorada e sujeita à variabilidade do leitor. Actualmente, a utilização de técnicas de deep learning tem ajudado a encontrar soluções válidas para problemas complexos, nomeadamente na imagiologia médica. O diagnóstico do ELMSI pode ser demorado, e pode depender do médico que realiza a ressonância magnética, do dispositivo, e do leitor. Além disso, não existem directrizes para a medição quantitativa. Em crianças e adultos jovens, as alterações naturais que ocorrem na medula óssea podem causar alterações visíveis em MRIs. Portanto, a aplicação de técnicas de deep learning para prever o ELMSI em crianças e jovens adultos pode ter grandes vantagens. Neste trabalho, uma Convolutional Neural Network (CNN) será treinada para classificar ELMSI nas imagens. Data augmentation e transfer learning serão utilizados para reduzir o over-fitting, bem como técnicas de pré-processamento para melhorar o desempenho do modelo.

Acknowledgements

This thesis could not be possible without the contribution and support of many people.

I'd like to express my gratitude to my supervisors, Hélder Oliveira and Tânia Pereira, who consistently followed the progress of my work, always guiding and supporting me; and also Sílvia Costa Dias, for always showing readiness to help and answer my questions. I would also like to thank Francisco Silva for always being available to answer my doubts, even if some were basic, and to offer his advice.

I'm very grateful for my family, for always being supportive throughout these years; for my sister, for always encouraging me and making me laugh; and for my girlfriend and my friends, who shared this ride of enthusiasm, stress, distress and accomplishment with me.

“You know what a learning experience is? A learning experience is one of those things that says, ’You know that thing you just did? Don’t do that.”’

Douglas Adams

Contents

List of Figures

List of Tables

Abreviaturas e Símbolos

BM	Bone Marrow
BME	Bone Marrow Edema
CNN	Convolutional Neural Network
ELMSI	Edema-Like Marrow Signal Intensity
ML	Machine Learning
MRI	Magnetic Resonance Imaging
OA	Osteoarthritis
ROI	Region Of Interest
STD	Standard Deviation

Chapter 1

Introduction

1.1 Context

Bone marrow edema or, bone marrow lesion, is the generic term applied to bone marrow signal intensity changes in magnetic resonance imaging. More recently, the term edema-like marrow signal intensity replaces the term BME [1], but for historical coherence, the BME term will be utilized in this thesis whenever that was the term used in the referenced publication. It is characterized by areas of high signal intensity in fluid-sensitive sequences (T2 or proton-density (PD) weighted images with fat-suppression, or short tau inversion-recovery (STIR)), and areas of intermediate to low signal intensity in T1-weighted images. For BME detection, fluid-sensitive sequences are more sensitive than T1-weighted images [2]. Histological studies on BME are scarce, however, the lesion is not a typical edema by histologic criteria [3]. Normal bone marrow is heavily composed of fat and adipocytes, while in BME areas there are accumulations of immune cells and microvessels [4]. It is this change in tissue composition that leads to atypical readings in MRI scans, as can be seen in Figure 1.1. Medical causes for BME are diverse, although it is often associated with conditions such as arthritis, trauma, osteoporosis, infection, among others. In some cases, pain and limb motion difficulty can be present. In most patients, the lesions are reversible and can be accelerated with treatment. The clinical relevance of BME is still being assessed by clinicians, but it is currently a topic of interest namely in professional sports [5].

Despite MRI being the favoured method for detection of BME, it has certain challenges caused by biochemical level changes that occur in the bone marrow associated with aging, which make diagnosis more difficult. Visually identifying BME in X-rays is considered unreliable by clinicians due to the imaging's lack of sensitivity in detecting changes in bone marrow structures. Due to the lack of guidelines in identifying and quantifying BME, a computer-automated method would greatly decrease human error and the diagnosis period.

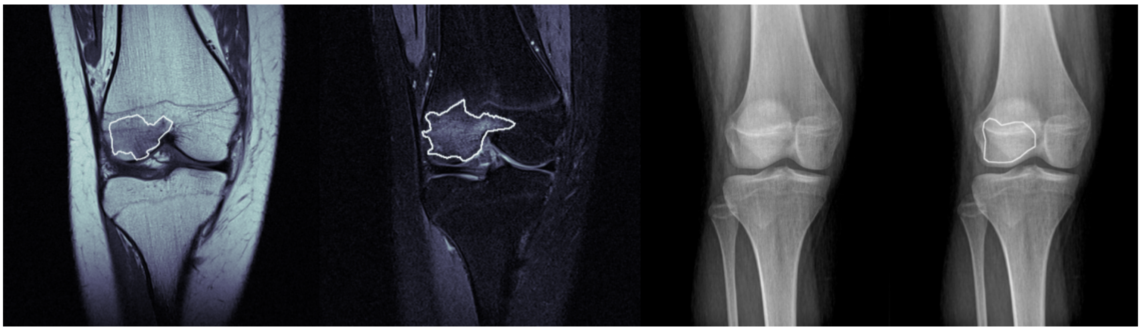


Figure 1.1: Different scans of a patient with BME. From left to right, the T1- weighted MRI image, the fluid-sensitive sequence image, the original X-ray and the X-ray with annotated region where the edema was found in MRI [6].

1.2 Motivation and objectives

Developing a machine learning (ML) model for classification of ELMSI in MRI scans would help clinicians by reducing their workload, decreasing human error and accelerating the time to correctly diagnose. Developing a model for detection of ELMSI in X-rays would be extremely beneficial to patients, as this imaging technique is very often the first to be performed in clinical trials, and would benefit clinicians as it is frequently not possible to visually detect it. An automated method for classification of ELMSI is extremely relevant, as there are no base guidelines for quantitative measurements. There is a margin of error in this process, since it also depends on the skills of the physician taking the MRI, the equipment and the experience of readers. Changes in the bone marrow can occur naturally in children and young adults, as the skeleton is developing, which make ELMSI detection more difficult. A machine learning model that can detect ELMSI on this demographic group would assist clinicians in the diagnosis of ELMSI.

The purpose of this thesis is to explore the challenges surrounding the visual classification of ELMSI in MRI sequences of children and young adults. In an attempt to classify ELMSI areas in these medical scans through an automated computer environment, CNNs will be trained and tested. A dataset of under-18 patients was provided by the University Hospital Center of São João, comprising two versions: one containing 36 edema patients and 36 non-edema patients, and another containing 28 of the same edema patients, but with different annotations. Since the datasets provided for this work are relatively small, and deep learning methods require large amounts of data to be trained without over-fitting, this thesis will focus mainly on image pre-processing techniques, transfer learning and data augmentation techniques, which have proven successful in recent works [7, 8]. These techniques offer an alternative to the patch extraction technique employed by Franco [6], where patches of arbitrary size are taken from original images and fed to the ML model, thus increasing the size of the dataset.

1.3 Contributions

The following contributions are expected from this thesis:

- an updated literature review on topics concerning BME detection using Machine Learning algorithms;
- implementation and evaluation of a deep learning model for detection of ELMSI on under-18 patients, using transfer learning;
- implementation and evaluation of a deep learning model for detection of ELMSI on under-18 patients, using data augmentation;
- implementation and evaluation of a deep learning model for detection of ELMSI on under-18 patients, using pre-processing techniques;
- a scientific publication with relevant results of this thesis (in development).

1.4 Document structure

The document is organized in the following way:

- chapter 1, Introduction, contains the context of the work, problem statement, objectives, motivations, possible contributions and techniques that will be used in the implementation;
- chapter 2, Background, describes theoretical aspects regarding bone marrow edema and succinctly explains deep learning and CNNs;
- chapter 3, Literature review, details past automated machine learning methods for detection of BME;
- chapter 4, Materials and methods, describes the dataset used to train the CNN, explains the techniques employed in each experiment, and the experimental and solution design;
- chapter 5, Results and discussion, displays and discusses the results of the trained models;
- chapter 6, Conclusions and future work, contains the conclusions and future work.

Chapter 2

Background

In this chapter, ELMSI will be described, as well as its causes, symptoms, treatments and detection techniques. Furthermore, CNNs will be briefly detailed with knowledge acquired mainly from Goodfellow et al's [9] "Deep Learning" book.

2.1 Bone marrow edema

2.1.1 Fundamentals of bone marrow edema

Bone marrow (BM) is a highly cellular connective tissue contained within the bones. It replaces the liver as the organ responsible for haemopoiesis after birth [10] and it is one of the largest organs by weight in the body behind bone, muscle and fat [11]. Healthy bone marrow is composed of fat-rich tissue, especially in peripheral sites, where haemopoies is not notable [4]. The cellular components include stem cells, erythrocytes, myeloid cells and megakaryocytes, which are important for maintaining oxygenation, immunity and coagulation [12]. Upon examination, bone marrow appears red (hematopoietic marrow) or yellow (fatty marrow) depending on its main components. However, fat represents the major component in both types, although to a lesser degree in red marrow. More specifically, when measured in anterior iliac crest, red marrow is composed of 40% fat cells and 60% hematopoietic cells, with a chemical composition of around 40% fat,40% water and 20% protein. On the other hand, yellow is composed of 95% fat cells and 5% nonfat cells, which equate to 80% fat, 15% water and 5% protein [13].

Bone marrow edema is a term used to describe high-signal-intensity alterations detected on MRI fluid-sensitive sequences of BM. More recently, the term edema-like marrow signal intensity replaces the term BME. It occurs when the fat present in bone marrow is partially replaced by water-rich components, as is illustrated in Figure 2.1, which leads to different readings on MRI scans when compared to healthy marrow scans. In a study relating osteoarthritis with BME, it is referenced that "Bone marrow fat is replaced by immune cells, particularly T and B lymphocytes, which form aggregates in the bone marrow and are associated with the accumulation of blood vessels" [4]. This process is an explanation for the changes in BM composition, although it may not apply to all BME cases.

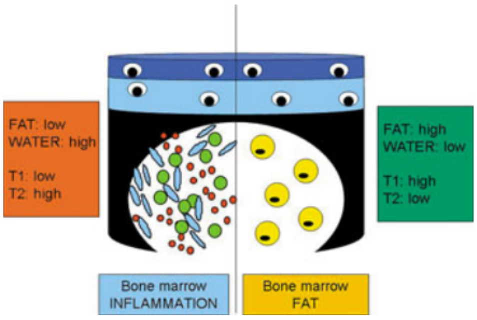


Figure 2.1: Illustration of changes in composition of bone marrow due to BME. Normal BM (right) composition is fat-rich, and can be detected by high signal intensities in T1 images, and low intensities in T2 images. On the contrary, in BME areas (left) immune cells and microvessels replace the adipocytes, increasing the water content. Therefore, BME can be detected by low signal intensities in T1 images and high intensities in T2 images. Figure from [4].

The term BME was first employed in 1988 in a study of 10 patients diagnosed with transient osteoporosis of the hip or knee, all of whom showed altered signal intensities of bone marrow in T1-weighted images and T2-weighted images. Because ischemic necrosis and metastasis were excluded by biopsy, and symptoms resolved spontaneously in all cases, the authors suggested naming these findings “transient bone marrow edema syndrome” [14]. In recent years, bone marrow edema syndrome has been more established as a change in intensity of normal marrow MRI scans, associated with pain in joints and lack of signs of avascular necrosis, antecedent trauma or infection [10]. Bone marrow edema is a common but nonspecific signal pattern of several diseases, and can be found on bone parts of joints in areas like the hip, knee, ankle, foot or shoulder. In a study by Hofmann et al [15], three main groups were categorized as causes for BME in the knee:

- Ischemic BME
    - Osteonecrosis
    - Bone marrow edema syndrome (BMES)
    - Osteochondritis dissecans (OCD)
    - Complex regional pain syndrome (CRPS)
- Mechanical BME
    - Bone contusion (bone bruise)
    - Microfracture
    - Stress-related BME
    - Stress fracture
- Reactive BME
    - Gonarthritis

- Osteoarthritis
- Postoperative BME
- Tumor-related BME

More recently, Maraghelli et al [1] proposes the categorization of ELMSI into: "ELMSI with known etiology", for cases when ELMSI is a non-specific but important finding and indicates the presence of a disease; "ELMSI with unknown etiology" for cases when ELMSI is an isolate finding with no apparent cause. Causes for ELMSI with unknown etiology are present in Table 2.1.

Table 2.1: Classification of ELMSI with known etiology. Adapted from [1].

Type	**Etiology**
Trauma	Direct or indirect damage, fracture, complex regional pain syndrome
Trauma/degenerative	Subchondral insufficiency fractures
Degenerative	Osteoarthritis
Degenerative/inflammatory	Modic changes
Inflammatory	Inflammatory arthritis, enthesitis
Vascular	Avascular necrosis
Infectious	Bone and articular infections
Neoplastic	Benign lesions
Neoplastic	Malignant lesions
Iatrogenic	After surgery or RT, steroids or calcineurin inhibitors
Metabolic	Hydroxyapatite deposition disease, calcium pyrophosphate deposition, gout
Neurological	Charcot's joints

In a study by Eriksen [2], an alternative yet similar BME aetiology is presented. Treatment for BME ranges from surgical interventions to physical modalities and medication. An example of the first case is drilling holes in the BME affected area for core decompression along with possible injections of hydroxyapatite cement or autologous bone marrow stem cells. The subchondroplasty technique is also a promising, yet underdeveloped technique [14]. For the second case, extracorporeal shock wave therapy can be applied, which consists of mechanical shocks of a certain magnitude in the BME affected area. As for pharmaceutical options, bisphosphonates have shown to reduce pain and extension of BME by inhibiting osteoclast activity, and so reduce bone resorption. Vasodilators like prostacyclin also show promising results. In most asymptomatic cases, BME fades away by itself, although this process can be accelerated by treatment. Clinical significance of BME is still under discussion, since it is not easy to identify symptoms directly associated with bone marrow lesions, as most of the times it is associated with soft tissue alterations.

2.1.2 Identifying bone marrow edema

MRI has been, for the last decades, the most common imaging modality for evaluating patients affected by pain in bones or joints, whose X-rays are first assessed by clinicians as inconclusive [15]. Since BME is characterized by the replacement of fat-rich components by water based components, MRI stands as the most adequate method for detection of BME. Imaging modalities such as X-rays are unable to detect changes in marrow structures with sufficient sensitivity, and

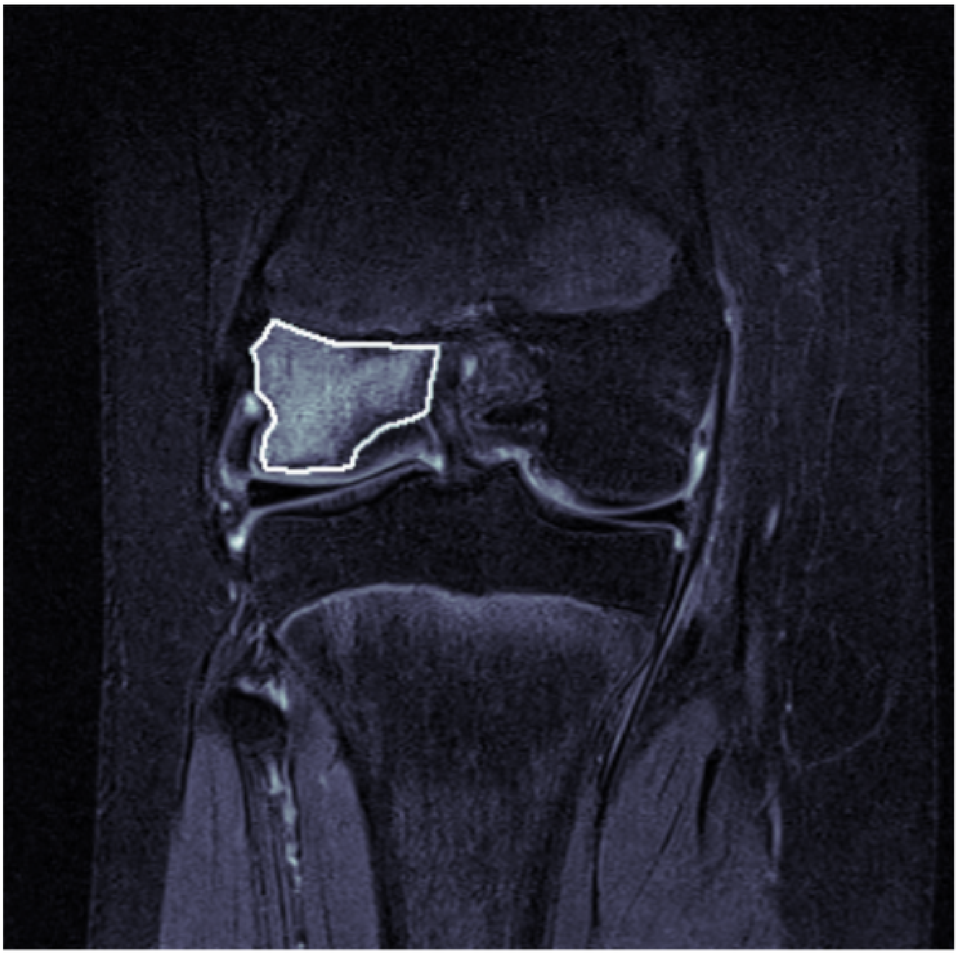


Figure 2.2: Example of a coronal fluid-sensitive sequence of the knee, with findings of BME. Figure from [6].

are therefore unreliable for clinicians to visually diagnose. The differences of BME readings between MRI scans and X-ray images can be seen in Figure 1.1, where it is noticeable that X-ray detection is very poor. On T1-weighted images it is characterized by intermediate to low signal intensity compared with unaffected bone marrow, while on T2-weighted images, especially when fat- suppression techniques are used, it is characterized by high signal intensity areas. An example of BME detection in an MRI scan can be seen in Figure 2.2.

Several classification systems have been elaborated by different authors, but there is no overall agreement. Most authors differentiate between reticular and geographic/demarcated patterns, while others highlight the location (subchondral versus at distance of joint space). Costa-Paz et al [16] classified BME in the knee in 3 types: type I was defined as alterations of medullary component, often reticular and distant from the subjacent articular surface; type II was defined as localized/geographic signal with contiguity to the subjacent articular surface; type III was defined as disruption or depression of the normal contour of the cortical surface/subchondral lamella.

Despite MRI being the favorable imaging technique for the detection of BME, it is not always

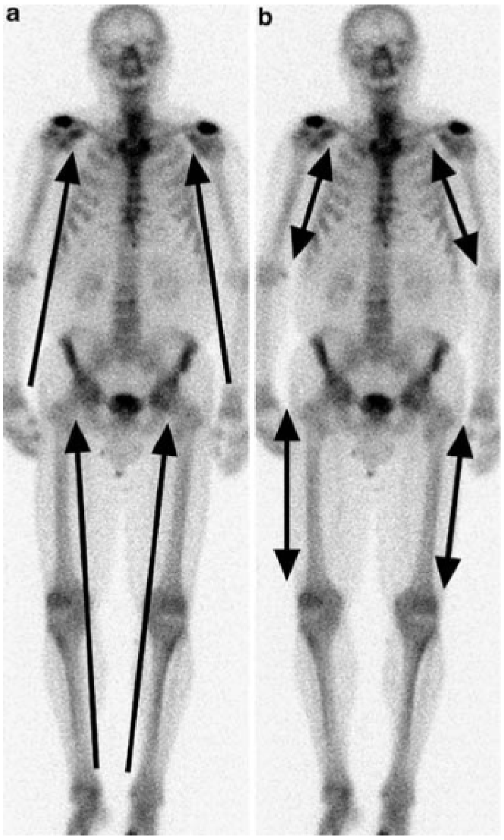


Figure 2.3: Convertion of bone marrow due to age, shown by two superimposed sequences. **a** shows convertion of red marrow to yellow marrow from the peripheral to the central skeleton. **b** is superimposed on that sequence, showing red marrow converting to yellow in long bones [12].

reliable. Changes at a biochemical level in the bone marrow can occur physiologically along with the development of the skeleton, which represents additional difficulty in diagnosing this clinical condition, specially in children and young adults. For example, at birth red marrow is present throughout the whole skeleton. Over the next two decades, various regions start converting from red to yellow marrow, starting in the periphery of the skeleton and extending into the central skeleton [12]. In Figure 2.3 this process is illustrated, where 2.3a shows the conversion from peripheral to the central skeleton, and 2.3b shows the conversion from diaphyses to metaphyses in long bones. In adults, the proportions of red and yellow marrow in the medullary cavity are approximately evenly split. Because healthy marrow is a fatty tissue, it can be identified in MRI scans through bright signals on T1-weighted images and dark signals on T2-weighted images.

2.2 Machine learning

Machine learning is a type of Artificial Intelligence that focuses on algorithms that map historical data as input to a certain output. Traditional machine learning models are defined as models that were used before the "Big Data Era", where deep learning models are gaining popularity. Traditional machine learning models are faster during the training process than deep learning algorithms, but generally slower when making predictions. They also require more time and domain knowledge in the preparation of data than deep learning methods. Nonetheless, deep learning models offer great benefits and are becoming more popular as they continue to solve different challenges, namely in computer vision.

2.2.1 Deep learning - CNNs

The Popularity of deep learning models has sky-rocketed in past years due to the increased size of data. These methods use multiple layers which transform the input in a way that increases abstraction and allow modeling of complex functions. Using techniques from representation learning, models are able to find efficient representations of data for automatic feature selection using prior knowledge, which cuts the need for manual feature engineering processes. Abstraction is a key concept because it can be compared to how a human would look for particular features when trying to visually recognize an object. For example, to identify the species of an animal a human would look for features like shapes, patterns or high contrast areas.

Convolutional neural networks (CNNs) were initially inspired by the relationship between the animal visual cortex and specific neurons. Over the years, its architecture was improved until it consisted of multiple layers with many types, as can be seen in Figure 2.4. The 3 main types consist of:

- convolutional layers;
- pooling layers;
- fully-connected layers.

Convolutional layers are the first and principal type of layers. In these layers, different filters are applied to image pixels through the convolution operation, using kernels of arbitrary shape. The goal is to generate features in order to produce a new representation of the image - the feature map. The purpose is to find the ideal filter so that the feature set is extracted from the images efficiently.

Pooling layers increase abstraction and discard information by downsampling the feature map. This layer is important because it reduces the dimension of the feature map, thus reducing the number of parameters that need to be learned and computed, and summarises features which make the model more robust. In the downsampling process, each independent matrix of pixels of specific size generates a numeric value, which is usually selected as the maximum or average value of the matrix.

Finally, after the feature set is processed, the model connects the feature map to the fully-connected layers of neurons, into the output layer with the desired output shape.

In conclusion, the advantages and disadvantages of CNNs are summarised in Table 2.2

Table 2.2: Advantages and disadvantages of CNNs.

Advantages	Disadvantages
• ability to automatically extract features relevant to the problem from videos and images • ability to outperform conventional machine learning methods in a lot of imaging problems, e.g. in the medical field [17] • capacity to identify objects anywhere in the image	• consumes large amounts of computing power, even if it takes advantage of GPU power • requires large datasets to outperform other machine learning algorithms • requires domain knowledge to be used effectively while preparing the dataset • are considered black boxes, since it is hard to understand their internal evolution

2.2.2 Metrics and confusion matrix

To accurately evaluate the behaviour of a model, certain metrics need to be calculated. An intuitive way of analyzing the output of classification problems is the confusion matrix (Figure 2.5). This matrix is composed of 4 numeric values: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). With these values, metrics like accuracy (Equation 2.1), precision (Equation 2.2), recall (Equation 2.3), specificity (Equation 2.4), F1 (Equation 2.5), among others, can be calculated. For regression problems, mean squared error and mean absolute error can be used to assess results.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.3}$$

$$Specificity = \frac{TN}{FP+TN} \tag{2.4}$$

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} \tag{2.5}$$

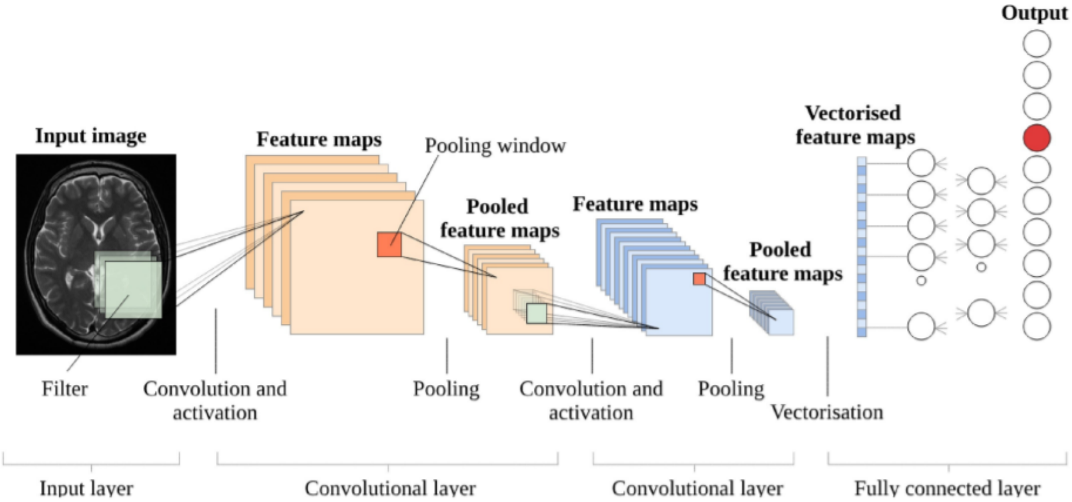
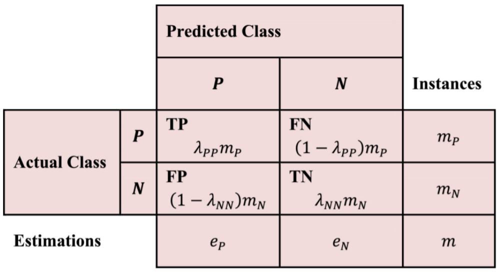


Figure 2.4: Visual representation of inner workings of a CNN. Figure from [18].



		Predicted Class		
		P	*N*	Instances
Actual Class	*P*	TP $\lambda_{PP}m_P$	FN $(1-\lambda_{PP})m_P$	m_P
	N	FP $(1-\lambda_{NN})m_N$	TN $\lambda_{NN}m_N$	m_N
Estimations		e_P	e_N	m

Figure 2.5: Layout of a confusion matrix. Figure from [19].

To evaluate performance on unbalanced datasets, metrics such as balanced accuracy (Equation 2.6), Receiver Operating Characteristic (ROC), and its respective Area Under the Curve (AUC) are more reliable. The latter metrics are illustrated in Figure 2.6.

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2} \tag{2.6}$$

2.3 Summary

Bone marrow edema is a term used to describe intensity modifications when compared to normal bone marrow in MRI sequences. It is a common, although nonspecific signal pattern of several diseases, and can be found on bone parts of joints such as the knee. It is associated with pain in joints

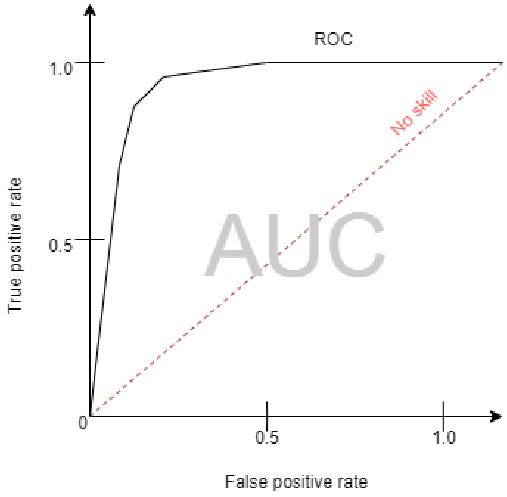


Figure 2.6: Illustration of ROC and AUC.

and limb motion difficulty. Treatments range from surgical interventions to physical modalities and medication. Clinical significance of BME is still under discussion. To identify BME, MRI has been the preferred modality for the last decades, since X-rays are not sensitive enough to changes in bone marrow. However, detecting BME in MRI exams is challenging, specially in children and young adults, due to natural changes that occur in the bone marrow structure. Additionally, there are still no clear guidelines for the detection of BME.

Machine learning is a branch of Artificial Intelligence which focuses on algorithms that utilize a sample of data to make predictions. A Convolutional Neural Network is a Deep Learning algorithm which is often applied to solve problems which involve images and videos. It is inspired by the connectivity between the animal visual cortex and specific neurons. Convolutional Neural Networks offer great benefits and have proven to outperform other algorithms in multiple problems, including medical imaging. However, it has some drawbacks, like the fact that it is a black box method.

Chapter 3

Literature review

In this chapter, relevant work previously published by multiple authors is discussed. It starts off by discussing some examples of traditional machine learning architectures developed in order to detect and quantify BME, and then explores more recent studies using deep learning techniques for the same purpose. Given that BME and deep learning in medical imaging are very recent and evolving concepts by themselves, there is very little available literature that combines both. Nevertheless, some conclusions from related works were extracted. Each section is organized by relevance to the thesis in descending order.

3.1 Detection using traditional Machine Learning

Aizenberg et al [20] studied the feasibility of quantifying BME using traditional ML techniques, in patients with early Arthritis. First, to fuse coronal and axial scans into a 3D image, super-resolution reconstruction was applied. The carpal bones were then segmented through atlas-based segmentation in order to detect BME. Then, BME was quantified within each bone by fuzzy clustering. To implement this process, the authors evaluated the fraction of voxels containing an intensity higher than a threshold compared to normal bone voxels in each segmentation. This means that the gravity of BME is being measured by its spread in each bone, instead of the intensity of BME affected areas. Four different levels of BME were assessed by trained readers: 0, which equates to no BME; 1, for patients with 1-33% of BME volume; 2, for 34-66% volume; and 3, for 67-100% volume. The algorithm described in this study showed good results, however, a few outliers were very clear. It was considerably limited by segmentation errors and noise related to the scanning techniques. Despite this, the idea of studying carpal bones is interesting since it is a part of the body that provides complex multi-object images, which can be used to study the resilience of the algorithm. Also, the notion of quantifying BME through the fraction of affected bone should be considered.

In previous work developed by Chuah et al [21], histogram-based textures such as mean intensity or standard deviation were calculated in 2D and 3D MRI scans. These values were compared in healthy and BME affected patients, and two-sample t-tests were carried out in order to assess

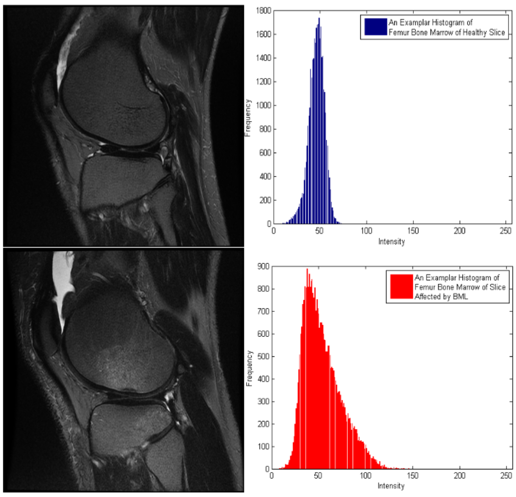


Figure 3.1: Example of histogram of a healthy femur slice (top), and a slice affected by BME (bottom). The bright patch at distal femur represents BME and its affect in the histogram is very clear. Figure from [21].

the difference between both sets of individuals. In Figure 3.1, the difference between readings is visible. This approach is obviously limited, however, some key aspects can be extracted. First, it was concluded that healthy slices of MRI scans show no apparent difference in healthy and BME affected patients. Furthermore, these statistics are able to detect differences in healthy and unhealthy slices, and 2D scans showed more robust results than 3D. Big limitations to this study include the spatial relationship lost between pixels with different intensities, and the lack of robustness to noise and artifacts which can heavily skew a histogram.

3.2 Detection using deep learning

In recent years, deep learning techniques have revolutionised the imaging paradigm. In previous methods, the general idea was to segment an image into a region of interest (ROI) and quantify/detect using handcrafted features relevant to the pathology. Through deep learning algorithms, the image can now be classified in one step, since the neural network is the one deciding which features are relevant for the application. However, this automatic feature selection reduces the transparency of classification and quantification methods, since the process becomes a sort of black box. Furthermore, most algorithms require supervised learning, which means human observers must carefully build a large dataset with detailed, conventionally approved annotations. Stoel [17] predicts the traditional ML methods for quantifying BME will be replaced by more recent AI techniques, which supports the focus of this literature review in deep learning.

This thesis will present itself as a sort of alternative to Franco's [6] work, as it will share the same dataset, but use different techniques to train the CNN. The dataset was collected from 72 patients aged under 18 and was provided by the University Hospital Center of São João. It is a balanced set, as 36 individuals had BME detected by health professionals and 36 were healthy. In

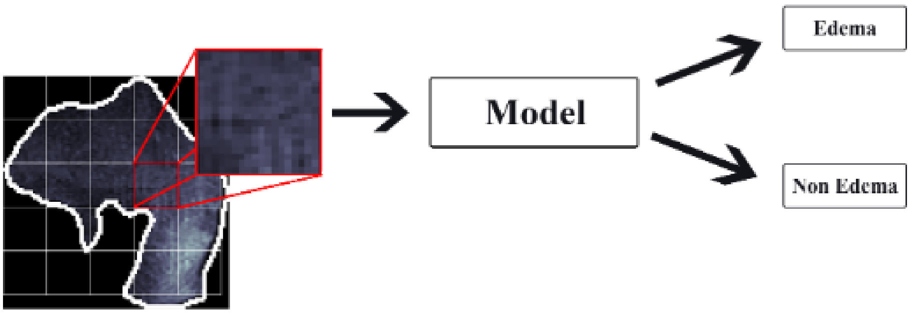
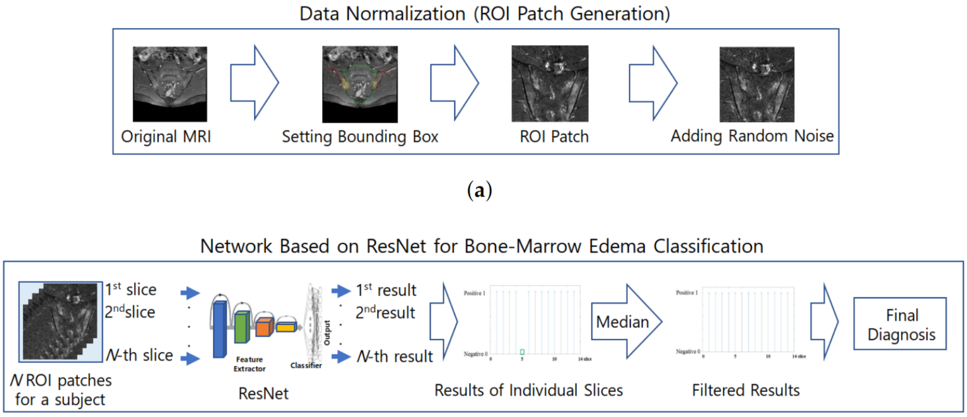


Figure 3.2: Illustration of the patch removal strategy. The patches are extracted from the ROI and fed to the model. Figure from [6].

healthy individuals, the healthy bone was annotated so that the algorithm can learn healthy composition, while the BME areas were annotated in affected patients. Franco started off by applying a mask in order to isolate the ROIs containing healthy bone or BME in each slice. Afterwards, due to the relatively small size of the dataset, tiny patches of an arbitrary size were extracted from the resulting masked image. This process can be visualized in Figure 3.2, and it is a common strategy when dealing with a reduced number of samples. It should be noted that this strategy induces class imbalance, as BME regions are typically smaller than healthy bone regions. To deal with this, undersampling and oversampling were performed, with the latter showing better, although not relevant enough results. The patches were then fed to Random Forests and a CNN model in order to classify healthy and unhealthy individuals. Random Forests are a type of traditional ML that work by having multiple decision trees, where each classifies the input, and the output is the majority of votes. The best results in MRI scans for the CNN were 72.10% balanced accuracy and 69.11% for the Random Forests. These represent decent results, although a margin for improvement is noticeable. It was also concluded that X-rays offer poor detection of BME, with the CNN achieving the best result of 56.32% balanced accuracy. Furthermore, the dataset is built by scans of teenagers and children, which has the added difficulty of containing natural chemical changes in the marrow. Fluid-sensitive sequences offered better results than T1-weighted sequences, possibly due to the increased amount of textures in the former images which may be interpreted by the algorithms as noise. The fact that the CNN was able to slightly outperform Random Forests with such a small number of samples proves the potential of deep learning techniques.

A more relevant study to the thesis is the work by Lee et al [7], where a CNN was trained using data augmentation and transfer learning to detect BME of sacroiliac joint with axial Spondyloarthritis, in a dataset containing mostly adults. MRI is useful for the detection of this medical condition, as consecutive readings of BME in this part of the body are a relatively strong indicator. In the first part of the work, the ROIs containing the sacral and iliac bones were extracted from the MRI slices. In this process, random noise is added so that the model is robust to noise formed during MRI acquisitions due to inhomogeneities of the receiving sensors, or from image reconstruction. Afterwards, these patches containing the ROI were fed to the ResNet, which will



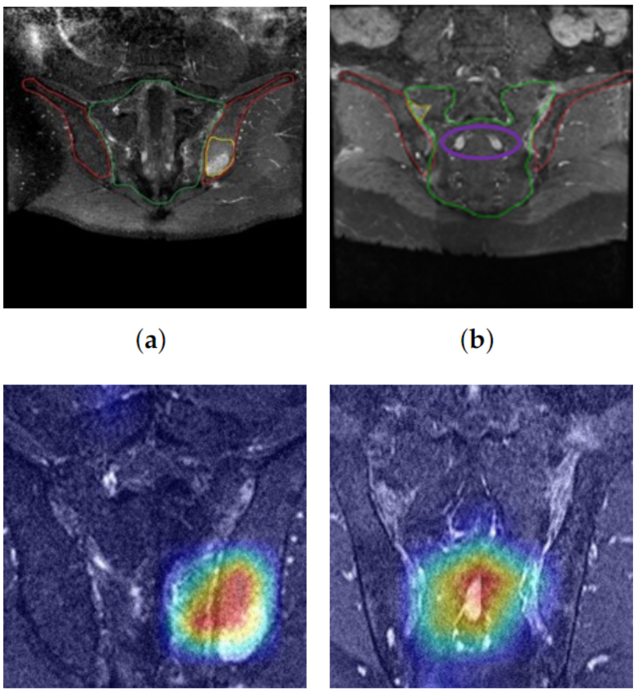
(a)

(b)

Figure 3.3: System architecture for detecting active sacroiliitis. **a** shows the process of obtaining the ROI. **b** ResNet for detecting BME in each slice, and median filter to diagnose active sacroiliitis in each patient. Figure from [7].

classify the presence or absence of BME. If BME was detected in at least two consecutive slices, or if 2 BM lesions in the same image are found, then active sacroiliitis is determined. To reduce misclassifications of active sacroiliitis due to a slice not containing BME, a median filter was applied. This system architecture can be visualized in Figure 3.2.

Since deep learning requires large amounts of data, transfer learning and data augmentation techniques were applied to the ResNet classifier. Transfer learning consists of using knowledge acquired from one task to solve another task. In this case, the CNN will learn how to recognize patterns in large databases, which enables the learning of BME patterns in a more efficient way using a smaller dataset. The authors pre-trained the ResNet using the publicly available ImageNet. Data augmentation is the process of increasing the dataset size by slightly modifying existing samples. It is useful to prevent over-fitting and increase the model's robustness to differences in spatial changes. In this case, the authors suggest not using the "random crop" technique since it could lead to false negatives due to the affected area being cut off. Thus, only horizontal flipping and rotation were applied. The results were very promising, with the CNN achieving a BME classification accuracy of 83.45%, recall of 85.13% and precision of 81.90% for images containing the full, non-cropped MRI scan. For the images containing ROI only, the results were 93.55% accuracy, 92.87% recall and 94.69% precision. It can be concluded that limiting the learning of the model to the ROIs increased its performance significantly. Additionally, by using the Grad-Cam technique, it was possible to analyze in which areas of the images the model was focusing on in order to make predictions. An example of this result can be seen in Figure 3.4, where BME is signaled by the yellow line, and bones by green and red. The heat maps are the result of extracting the class activation map for the unhealthy class. The degree of activation decreases in the order of red, orange, yellow, green and blue. In Figures 3.4a and 3.4c, it is possible to conclude that the model correctly focuses on the BME affected area. However, as is shown in Figures 3.4b and 3.4d,



(a) (b)

(c) (d)

Figure 3.4: Examples of class activation mapping compared to original images. **a** and **b** show examples of two different slices, where the bones are signaled with red and green lines. The purple line represents the structural features of the sacral bone that may cause a false positive. **c** shows the Gradient-based class activation map of (a). **d** shows the Gradient-based class activation map of (b). Figure from [7].

high intensity noise caused by the structural features of the sacral bone (shown in purple) deviate the model's attention away from the BM lesion. This type of noise can cause false positives, and in quantification applications, could cause the wrong prediction of BME gravity. In conclusion, this study supports the belief that employing transfer learning and data augmentation techniques in order to train the CNN is worth pursuing.

In order to automatically detect morphological and degenerative changes in patients with osteoarthritis (OA) of the hip joint, Tibrewala et al [8] developed an architecture using deep learning. OA can be detected in MR sequences by identifying abnormalities in the bone and cartilage, such as BME. The authors first segmented the slices into ROIs using a pre-trained ResNet-50, and then classified individuals into healthy or unhealthy using a DenseNet-100 architecture. The classification model was pre-trained with the CIFAR-32 dataset. Results were solid, with a sensitivity of 73% and specificity of 92%. A saliency map obtained using Grad-Cam can be seen in Figure 3.5, where the model correctly focuses on the affected area. This study is yet another example of a robust pipeline for the detection of abnormalities in MRI scans.

Von Brandis et al [22] developed a feasibility study on the automated segmentation of BM signal in MRI sequences. Although it is a segmentation work rather than a classification work,

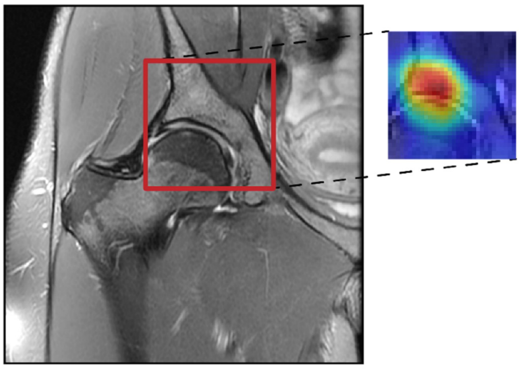


Figure 3.5: Saliency map generated by Grad-Cam. Figure from [7].

it is an interesting study since, among other key conclusions, it stresses the difficulties of BM signal detection in MR images. The authors also used a dataset containing mainly children, and mentioned the difficulties in detecting abnormalities in such samples. They converted all scans to NIfTI image format, instead of keeping the DICOM format, due to the former being simpler and more standardised. To quantify BM signal intensity, three levels were created by two experienced radiologists: 1 is a slightly increased with diffuse intensity distribution area; 2 is a focal and mildly increased area; 3 is a focal and moderate to highly increased area. Examples of the different levels are present in Figure 3.6. The authors theorized that the highest signal-intensity level is likely pathological, and in children with inflammatory changes, level 3 was almost always associated with level 2 and level 1 signals. These are interesting hypotheses because they might help separate pathological causes from natural composition causes. More specifically, an extensive level 2 pattern might be a better indicator of pathology than a level 3. The authors also detailed the challenges attributed to the fact that there is no general consensus yet on BM intensities, which can affect the training model due to the ground-truth being developed with no globally accepted guidelines.

A similar quantification strategy was studied by Han et al [23], in order to detect hip BME in ankylosing spondylitis using deep learning. The quantification is expressed as a percentage of marrow inflammation, using 4 different grades: 0 for healthy patients; 1 for mild inflammation (BME<15%); 2 for moderate (BME 15-30%); 3 for severe (BME>30%). In this work, the 3D MRI sequences were segmented using a variant of a U-Net and the resulting inflamed regions are sliced in 2D. Then, these resulting slices are concatenated with the original MRI slice and fed to a ResNet classifier. The results were very close to the manual annotations, although the classifier performance depended on the quality of segmentation. The authors also mention the fact that the quality of the MRI images depends on the radiographer's skills and equipment, and thus a larger and more diverse dataset is needed in order to obtain a more robust model.

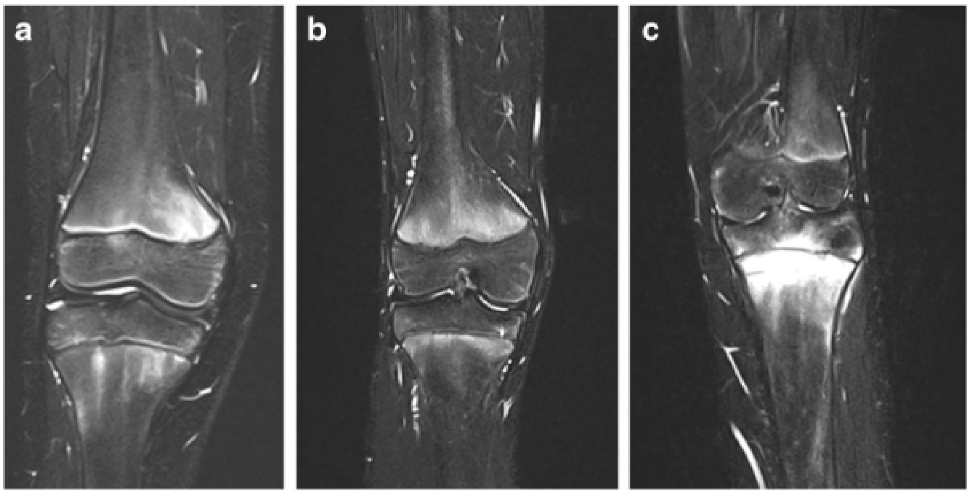


Figure 3.6: Levels of high intensity BM signals on coronal T2-W Dixon water-only images of the knee. **a** is an MR scan taken from a 12 year-old boy with chronic non-bacterial osteomyelitis and knee symptoms. **b** is from a healthy 14-year-old boy who showed no symptoms. **c** is from a 15-year-old girl with chronic non-bacterial osteomyelitis and symptoms in the knee. Adapted from [22].

3.3 Summary

In this chapter, a literature review of studies relevant to the thesis was performed. An emphasis was given to authors who used transfer learning and data augmentation in their works, as these will be the techniques employed in this thesis. Since ELMSI and deep learning techniques in medical imaging are hot topics of investigation, combining both concepts into a solution that helps clinicians is of utmost interest. It would also solve many issues such as reader variability. Due to the high prices of acquiring MRI images, and non-existent large datasets of these sequences, developing a model that uses data augmentation and transfer learning is extremely relevant.

Chapter 4

Materials and methods

This chapter focuses on the dataset used for this thesis, as well as the techniques developed regarding the pre-processing of images and strategies to avoid over-fitting. Finally, the experimental and solution design are described.

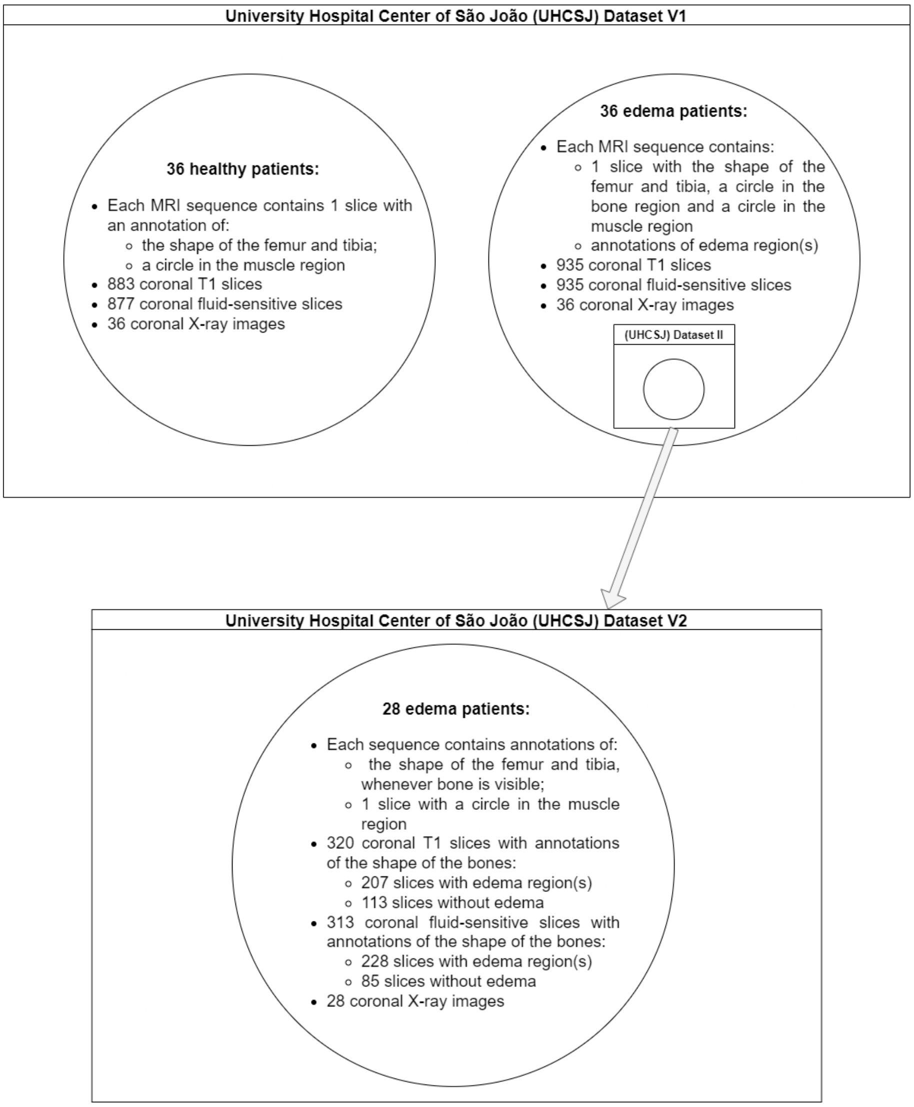
4.1 Datasets

For the training of the CNN, a dataset provided by the University Hospital Center of São João was used. This dataset contains T1, fluid-sensitive and X-ray sequences of 36 non-edema and 36 edema patients, all aged under-18. These exams were collected from different devices. Initially, in the first version of the dataset, non-edema patients contained a single slice with segmentations of the tibia and femur, while edema patients contained annotations of the edema regions. Afterward, in the second version of the dataset, most edema patients were annotated with the tibia and femur segmentations in every slice. In other words, the second version contains a portion of the edema patients included in the first version, but with different annotations. The contents of both versions are briefly detailed in Figure 4.1. This distinction between versions was created to facilitate the interpretation of the experiments, since both versions of this dataset were used as separate datasets. MRI annotations were performed by one consultant radiologist and one radiology resident, and all the annotations double checked with a concordance review.

4.1.1 University Hospital Center of São João (UHCSJ) Dataset V1

The UHCSJ Dataset V1, provided by the University Hospital Center of São João, contains T1, fluid-sensitive and X-ray sequences of 36 healthy and 36 edema patients. This dataset contains 2D MRI examinations of coronal, axial, and sagittal planes of the knee, although annotations are only present in coronal views. X-ray images include sagittal and coronal planes of the knee. The dataset distribution is represented in Table 4.1.

In both T1 and fluid-sensitive sequences, each healthy patient contains a single slice with annotations of the shape of the bones, and a circle in the muscle region. Alternatively, each edema patient, in both T1 and fluid-sensitive sequences, contains annotations of edema regions, as well as


University Hospital Center of São João (UHCSJ) Dataset V1
(UHCSJ) Dataset II
University Hospital Center of São João (UHCSJ) Dataset V2

36 healthy patients:

- Each MRI sequence contains 1 slice with an annotation of:
  - the shape of the femur and tibia;
  - a circle in the muscle region
- 883 coronal T1 slices
- 877 coronal fluid-sensitive slices
- 36 coronal X-ray images

36 edema patients:

- Each MRI sequence contains:
  - 1 slice with the shape of the femur and tibia, a circle in the bone region and a circle in the muscle region
  - annotations of edema region(s)
- 935 coronal T1 slices
- 935 coronal fluid-sensitive slices
- 36 coronal X-ray images

28 edema patients:

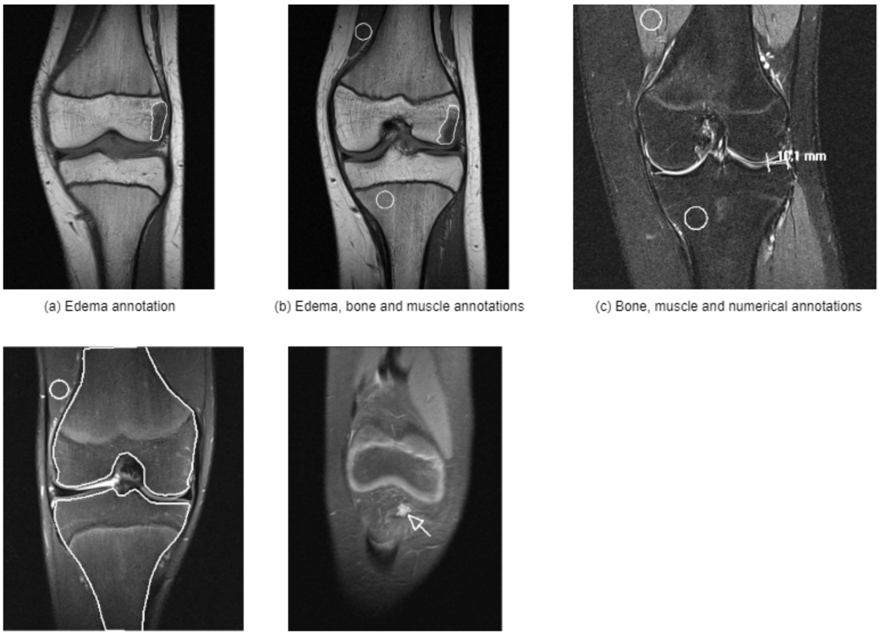
- Each sequence contains annotations of:
  - the shape of the femur and tibia, whenever bone is visible;
  - 1 slice with a circle in the muscle region
- 320 coronal T1 slices with annotations of the shape of the bones:
  - 207 slices with edema region(s)
  - 113 slices without edema
- 313 coronal fluid-sensitive slices with annotations of the shape of the bones:
  - 228 slices with edema region(s)
  - 85 slices without edema
- 28 coronal X-ray images

Figure 4.1: Diagram summarising the contents of both datasets used in this thesis.

a single slice with a circle in the bone region and a circle in the muscle region. Examples of these annotations are present in Figure 4.2. Furthermore, both sets of patients contain annotations that needed to be removed, such as numbers and arrows, as is shown in Figure 4.2c and Figure 4.2e. These unwanted annotations were registered in the original exams, and could not be removed by the clinicians. In X-ray images, edema patients contain annotations of edema regions, a circle in the bone, and a circle in muscle. In contrast, healthy patients include a circle in the muscle region and the segmentation of the bones, as shown in Figure 4.3.

Table 4.1: UHCSJ Dataset V1 distribution

	# Slices			
	Non-edema patients	With edema regions in edema patients	Without edema regions in edema patients	**Total**
T1	883	334	601	1818
Fluid-sensitive	877	381	554	1812
X-ray	36	36	-	72



(a) Edema annotation

(b) Edema, bone and muscle annotations

(c) Bone, muscle and numerical annotations

(d) Muscle and bone segmentation annotations

(e) Arrow annotation

Figure 4.2: Examples of MRI slices with annotations present in the UHCSJ Dataset V1. Slices **(a)**, **(b)** and **(c)** belong to edema patients, while slices **(d)** and **(e)** are extracted from healthy patients. Slices **(a)**, **(b)** and **(d)** are T1 sequences, while **(c)** and **(e)** are fluid-sensitive sequences.

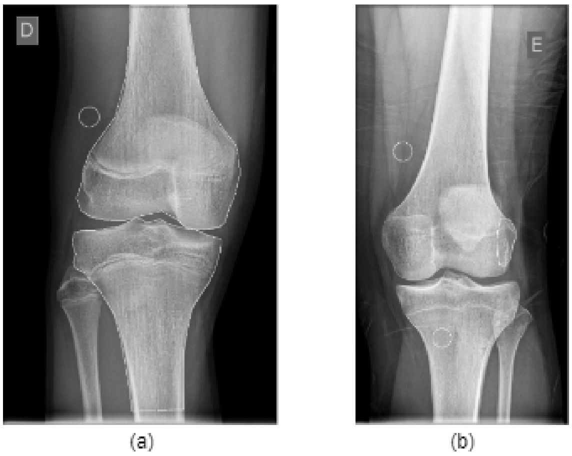


Figure 4.3: Examples of X-ray slices with annotations present in the UHCSJ Dataset V1. Slice **(a)** belongs to a healthy patient; Slice **(b)** belongs to an edema patient

4.1.2 University Hospital Center of São João (UHCSJ) Dataset V2

The UHCSJ Dataset V2, provided once more by the University Hospital Center of São João, contains the same MRI and X-ray sequences of 28 of the edema patients present in the UHCSJ Dataset V1. However, it now contains annotations of every bone shape whenever bone is visible, i.e., annotations with the femur and tibia segmented, and a circle in the muscle area in one of the slices. Examples of these annotations are present in Figure 4.4. Table 4.2 depicts the number of slices with bones segmented, as well as how many of these contain edema regions.

Table 4.2: UHCSJ Dataset V2 distribution

	# Slices with bones segmented		
	With edema regions	Without edema regions	**Total**
T1	207	113	320
Fluid-sensitive	228	85	313
X-ray	36	-	36

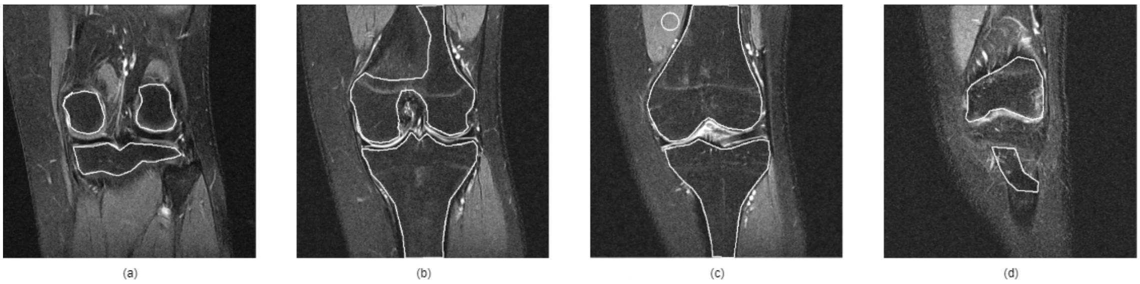


Figure 4.4: Examples of fluid-sensitive sequences, taken from a single patient, with annotations present in the UHCSJ Dataset V2.

4.2 Pre-processing

Each patient folder contains the exams in the DICOM format, as well as an XML file with details about the exams, which made the isolate extraction of annotation and slice files possible. Each slice was matched with its respective annotation file using their identifiers. The images and masks were then converted to *Numpy* arrays for easier manipulation.

4.2.1 Cleaning annotations

As was previously mentioned, some of the annotations present in the datasets were of no use to this work. In the case of healthy patients, only bone masks and the circle in the muscle region were of interest; for edema patients, only edema regions and circles in bone and muscle areas were useful. Bone masks are useful to obtain segmentations of the bones; the circles are useful to gather information regarding pixel intensities in their respective regions; and edema annotations are used to determine if a slice contains ELMSI. To remove unwanted annotations in healthy patients, the following algorithm was developed:

1. Small dilation (e.g. using a 3x3 kernel) to guarantee there are no gaps in annotation lines;
2. Detect closed contours using *OpenCV*'s *FindContours()* [1] function, and maintain the inner contours;
3. Ignore annotations that do not contain 2 closed contours with a large area, e.g. annotations that only contain numbers or arrows;
4. Remove every contour beyond bone limits.

The kernel chosen for the dilation might be a limitation if new cases are added to the dataset, as it may need adjustments for cases where it doesn't close the gaps properly. Another limitation is the threshold that defines the cutoff for the area of bones. A result of this algorithm is present in Figure 4.5.

For edema patient annotations, simply removing small closed contours (caused, for example, by numbers) was enough to remove annotations that only contained unwanted masks.

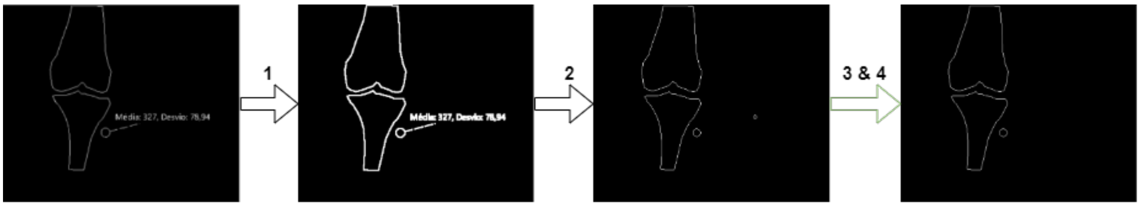


Figure 4.5: Example of unwanted annotations being removed.

4.2.2 Normalization

Slices are 2^{12} bits images, which implies pixel intensities vary between 0 and 4095. Muscle and bone intensities were acquired by averaging the pixel intensities in the circles annotated by the clinicians. It was observed that muscle intensities are relatively similar in all slices, despite the use of different image acquisition devices, while bone intensities were quite different. To convert the image intensities to the [0,1] range, three different methods were applied:

- **Max intensity** - every pixel intensity is divided by 4095. This method could limit the range of bone intensities in sequences with lower than average bone pixel values;
- **Max image intensity** - compute the maximum intensity of each image, and divide each intensity by its corresponding image maximum;

[1] https://docs.opencv.org/4.x/d3/dc0/group__imgproc__shape.html#gadf1ad6a0b82947fa1fe3c3d497f260e0

- **Muscle ratio** - intensities of muscle tissue are relatively similar in all slices. Therefore, muscle intensities can be used to standardize bone intensities in every slice. To do so, for each patient a constant value of 150 was divided by their respective muscle annotation, and the result was multiplied in all pixels. After that, the result was divided by 4095.

4.2.3 High intensity masking

Due to results further described in Chapter 5, a method to extract the ELMSI features from MRI images was developed. The method is described in Figure 4.6.

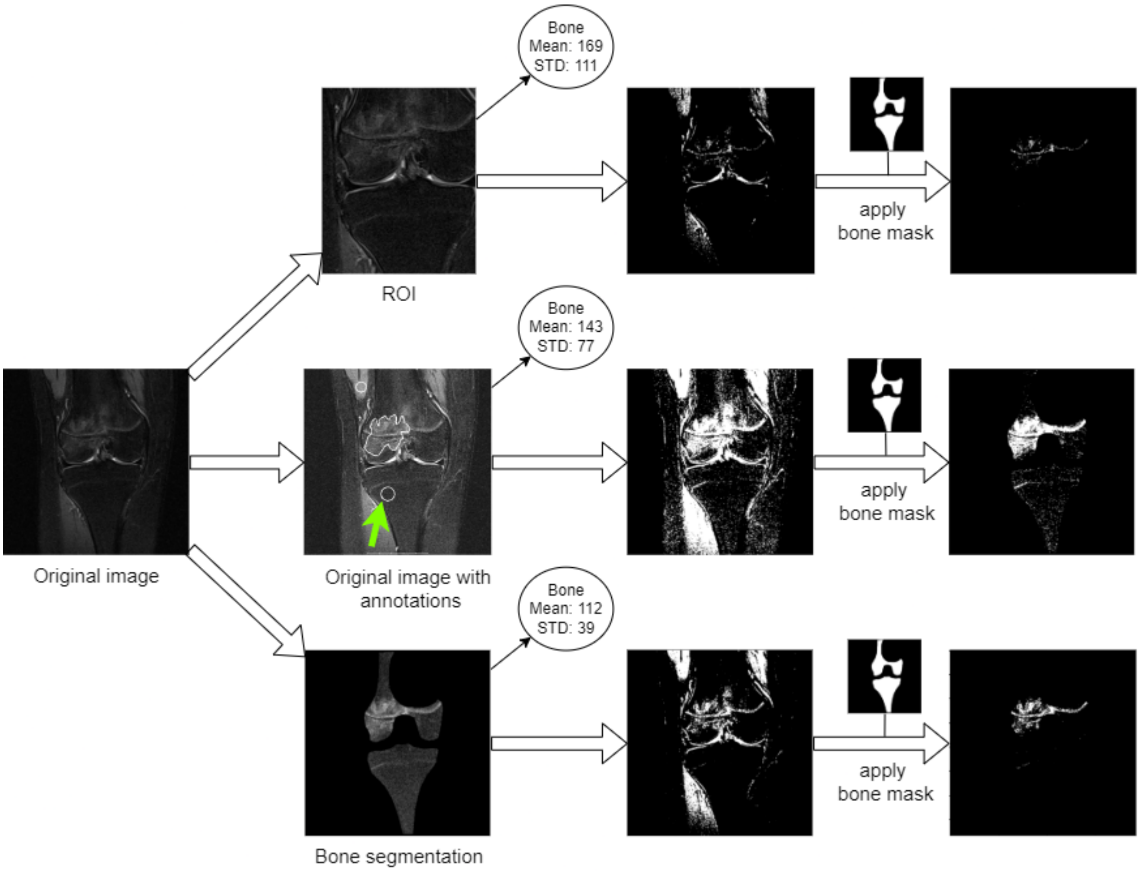


Figure 4.6: High intensity masking technique, developed to extract ELMSI features on fluid-sensitive sequences. First, the mean and STD of the bone signal intensities are calculated; then, using these values, the image intensities are clamped, and a histogram equalization using *MinMax* algorithm with values 0 and 255 is applied; afterward, the bone mask extracted from UHCSJ Dataset V2 is applied. Three ways of calculating the mean and STD of bone intensities were tested: using a ROI centered in the joints (upper row); using the circle annotation located in the bone (middle row); using the bone segmentations extracted from the annotations (lower row).

Edema regions are defined by lower than normal bone intensities in T1 sequences, and higher than normal in fluid-sensitive sequences. This method attempts to extract the range of intensities of edema areas on fluid-sensitive sequences. Since the range and value of bone intensities vary between patients, so does the range of intensities of edema regions. This means that a fixed range can not be used to isolate the edema regions in all patients. Therefore, the mean bone intensities must be computed in each patient to extract the ELMSI regions. This calculation was attempted using three different methods:

- **Region of Interest (ROI)** - by setting a ROI in the central part of the scan, it is possible to get a satisfactory approximation of the bone mean and standard deviation (STD). However, other body tissues are present in the ROI, which can affect the calculations and cause the detection of noise;
- **Bone annotation** - compute the mean and STD of bone intensities using the circle annotated by the clinicians;
- **Bone segmentation** - compute the mean and STD of bone intensities using the segmentations of the bones provided by clinicians. This method appeared to show the best results, as it was more robust to the detection of noise.

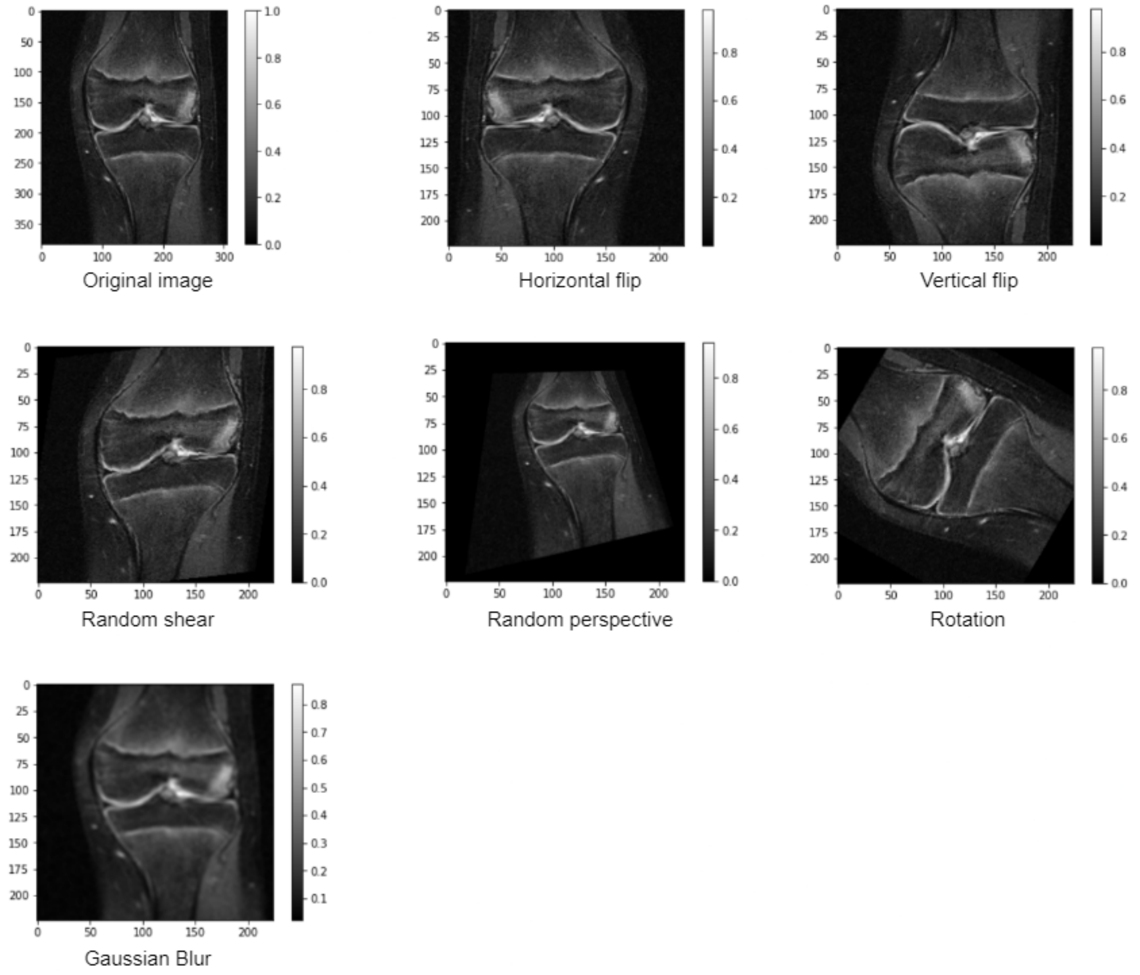
Afterwards, the values of all pixels are limited to a range given by *[mean + 1.5*STD, mean + 2*STD]*. In other words, every value lower or higher than these thresholds is assigned the respective limit. These thresholds were determined by visually assessing the final image, with the goal of detecting the edema regions while avoiding the detection of noise and structures such as the growth plate. After that, the images are normalized using the *MinMax* algorithm [2] with 0 and 255 as lower and upper limits. This last procedure is merely a shift in the range of intensities, which makes, for example, every pixel that was assigned the lower threshold have intensity 0 (black). It also makes the normalization process easier, since it is only necessary to divide every intensity by 255 to obtain a [0, 1] range. After obtaining the features, the bone mask is applied, leaving only the features detected on the bones. The ideal result for non-edema slices would be a black image, which means no noise and structures in BM were captured; for edema slices, the ideal result would contain only the edema intensities.

4.3 Transfer Learning and Data Augmentation

In order to decrease over-fitting, data augmentation techniques were applied to the images before being fed to the CNN. These techniques were applied with a concern for not overly distorting the edema regions, which would negatively affect the learning process. Examples of the applied data augmentation techniques can be seen in Figure 4.7. Each technique had a 50% probability of being applied.

Transfer learning is another helpful technique in avoiding over-fitting. The Resnet-18 [24] architecture has been employed in previous ELMSI classification works [7], and has proven itself efficient in multiple applications, namely when making use of its pre-training in the ImageNet

[2] https://docs.opencv.org/3.4/d2/de8/group__core__array.html#ga87eef7ee3970f86906d69a92cbf064bd



Original image

Horizontal flip

Vertical flip

Random shear

Random perspective

Rotation

Gaussian Blur

Figure 4.7: Data augmentation techniques.

dataset. To make use of transfer learning efficiently, usually layers are frozen, which means their weights can not be updated, and are therefore not trainable. To fine-tune the model, two approaches were used:

- progressively freezing layers before initializing the training, only training the classifier and the unfrozen layers throughout the learning process. I.e. in the first experiment, first the weights are loaded from the pre-training on ImageNet, then 9 of the 10 Resnet-18 layers are frozen before initializing training (meaning only the classifier is trainable), and finally the model is trained and tested; in the second experiment, the weights are again loaded from the pre-training on ImageNet, but now 8 layers are frozen before initializing training (meaning the classifier and the layer before it are trainable), and then the model is trained and tested; and so on until no layer is frozen;

- freeze all layers except the classifier before initializing the training, and unfreeze layers one by one throughout the training.

4.4 Experimental and solution design

In order to predict if slices contain ELMSI, the Resnet-18 was fed with 4 different types of images. Initially, the original images from UHCSJ Dataset V1 were used with normalization techniques as the only pre-processing. Afterward, ROIs centered on the joints of the bones were fed into the CNN. After that, using the UHCSJ Dataset V2, bone segmentations were used as input to the model. Finally, the result of the high intensity masking technique on fluid-sensitive sequences was fed to the CNN. A broad view of the developed work is present in Figure 4.8.

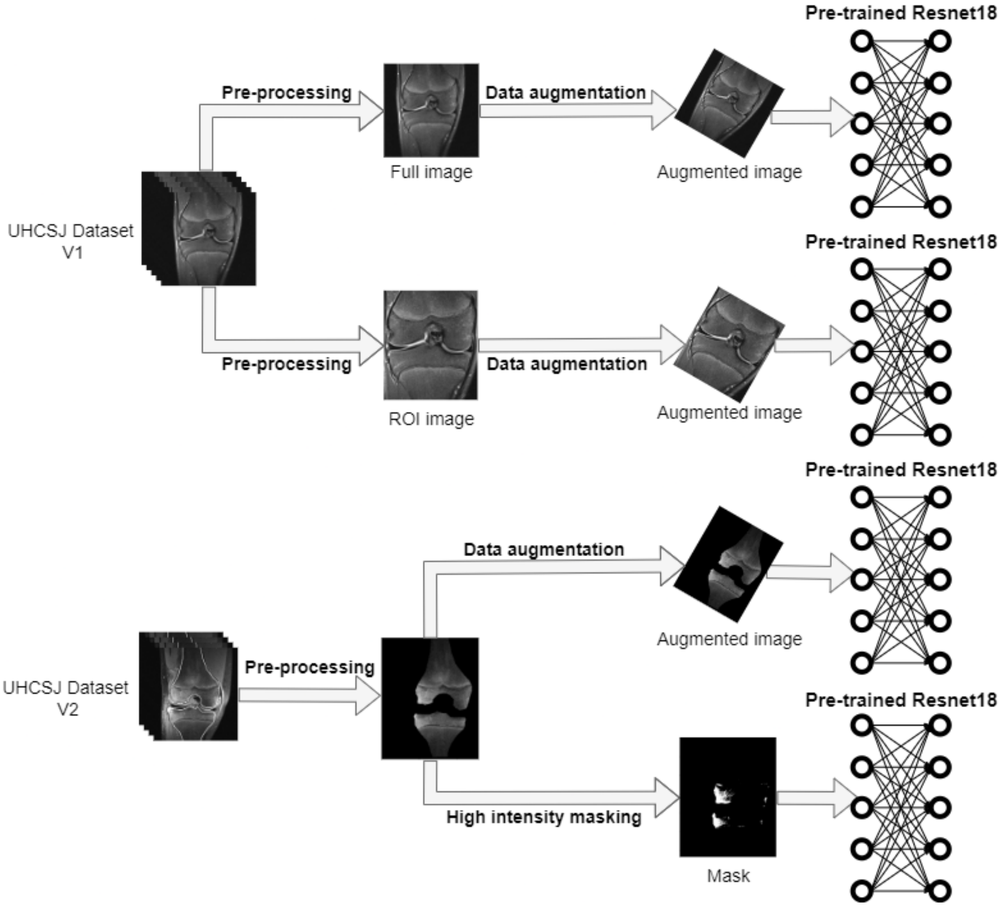


Figure 4.8: General overview of the developed pipelines.

In every experiment, the dataset was split into train/validation/test sets and was shuffled by patient following a 60/20/20 ratio. This means that, of the 72 patients contained in the UHCSJ Dataset V1, 44 were in the training set, 14 on the validation set and 14 on the test set. Similarly, of the 28 patients present in the UHCSJ Dataset V2, 17 were in the training set, 6 were in the validation set and 5 were in the test set. Initially, experiments with hyper-parameters were conducted to reach a stable and efficient training. The results of these experiments were a learning rate 0.001, batch size 113 and the best optimizer was Stochastic Gradient Descent with momentum 0. These hyper-parameters were kept throughout the experiments, and cross entropy was used as the loss function. On the validation set, the model with the highest accuracy (or balanced accuracy when the dataset was unbalanced) was saved, and later evaluated on the test set. In every experiment, different combinations of data augmentation were tested to assess their effect. To make use of transfer learning on ImageNet, only the classifier layer was trained, except when performing fine-tuning. In experiments where transfer learning was not utilized, every layer was unfrozen and the weights were initialized randomly. All experiments were performed and developed on the *Google Colab* environment using the *Pytorch* [3] framework.

4.5 Summary

In this chapter, both versions of the dataset provided by University Hospital Center of São João were described. The second version of the dataset contains only a portion of the edema patients present in the dataset, but with different annotations that were useful for segmenting the femur and tibia. After that, the process of removing unwanted annotations was explained, as well as the 3 different normalization techniques that were applied. The goal of these techniques is to shift the range of pixel intensities from [0, 4095] to [0,1]. Additionally, in an attempt to extract the ELMSI features from the slices, the method of high intensity masking was described. The goal is to obtain a black image that only contains edema intensities. Afterward, transfer learning and data augmentation were mentioned as the techniques used to improve the model and reduce over-fitting. The chosen architecture was the Resnet-18 with pre-training on the ImageNet dataset. Finally, the experimental and solution design was detailed.

[3] https://pytorch.org/

Chapter 5

Results and discussion

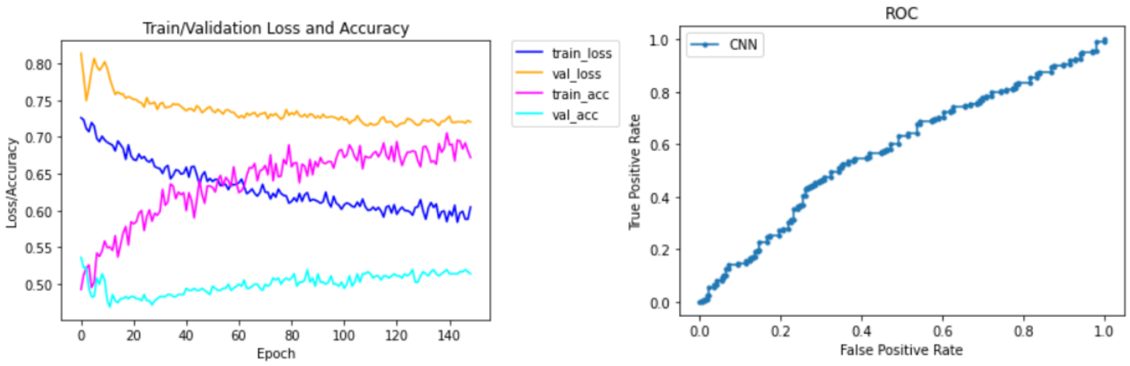
In this chapter, the results of binary classification of ELMSI using a Resnet-18 are presented and discussed. Initially, using the UHCSJ Dataset V1, the whole images were fed to the CNN. Afterward, due to lackluster results, a ROI was cropped from the images and fed to the model. To try a different approach, the UHCSJ Dataset V2 was used to obtain segmentations of the bones, which were then used as input to network. Despite the use of data augmentation techniques and transfer learning, to try to reduce over-fitting and improve the generalization of the model, the results did not considerably improve. After that, the images containing features extracted by the high intensity masking technique were fed to the CNN, which produced much better results.

5.1 Prediction using whole images

Initially, all scans were fed to the CNN, with no pre-processing other than normalization and data augmentation. All images that belonged to edema patients were labeled as edema, and every scan that belonged to non-edema patients was labeled as non-edema. The results are present in Table 5.1. The training curves and ROC of the model with the highest test accuracy are present in Figure 5.1.

Table 5.1: Results on the test set using whole images.

MRI	Transfer Learning	Data augmentation	Normalization	Best epoch	Accuracy	AUC
T1	-	-	Max image intensity	317	0.474	0.473
T1	ImageNet	-	Max image intensity	56	0.471	0.425
T1	ImageNet	Random perspective Horizontal flip Vertical flip	Max intensity	1	0.521	0.521
T1	ImageNet	Random perspective Horizontal flip	Max image intensity	1	0.566	0.578
T1	ImageNet	Horizontal flip Gaussian blur	Muscle ratio	1	0.554	0.551
Fluid-sensitive	ImageNet	Random perspective Horizontal flip Vertical flip	Muscle ratio	1	0.407	0.407



(a) Accuracy and loss of the train and validation sets throughout the training.

(b) ROC in the test set.

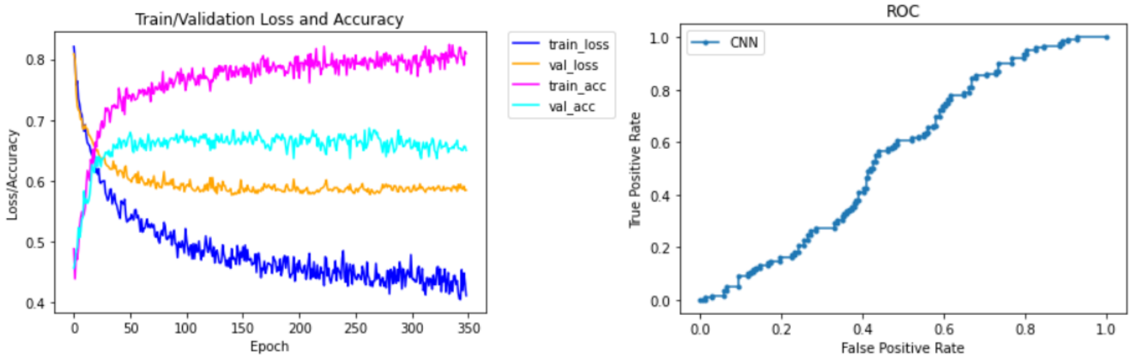
Figure 5.1: Training curves and ROC using: whole images; T1 sequences; with pre-training; random perspective and horizontal flip as data augmentation; max image intensity as normalization.

It is possible to see by the lackluster results and training curves, that the network was not properly learning the necessary features to predict ELMSI. Applying different normalization techniques, transfer learning and data augmentation techniques did not improve the results. The validation loss was quite high and the training process was overall not very stable.

Since ELMSI is a condition found in bones, it may not make sense to label images as containing edema when they do not contain any visible bone, or edema areas annotated. Therefore, in the following experiment only images containing annotations of edema regions were labeled as edema. Oversampling on the edema label was performed in order to balance the dataset. The best results on the test set are displayed in Table 5.2. The training curves, as well as the ROC, of the best performing model in the test set are present in Figure 5.2. It is possible to see that the evaluation on the test set did not improve when compared to the previous experiment, although data augmentation was an improvement over no data augmentation. Regardless, the training curves look more ordinary, although the validation loss is still high. It is also possible to conclude by the convergence of the curves, that the model has enough capacity for the complexity of the dataset. The model still shows signs of heavy over-fitting quite early, as the validation loss stops decreasing, or even starts increasing, while the training loss keeps decreasing. The big gap between the validation loss/accuracy and training loss/accuracy, combined with over-fitting, might indicate that the model is not generalizing well enough. A possible reason for this may be the high complexity and resolution of the images, as well as the small size of the dataset.

Table 5.2: Results on the test set using whole images when the dataset only contains, for the edema label, slices with edema regions annotated.

MRI	Transfer Learning	Data augmentation	Normalization	Best epoch	Accuracy	AUC
T1	ImageNet	-	Muscle ratio	57	0.457	0.483
T1	ImageNet	Random perspective Horizontal flip Gaussian blur	Max image intensity	98	0.553	0.552



(a) Accuracy and loss of the train and validation sets throughout the training.

(b) ROC in the test set.

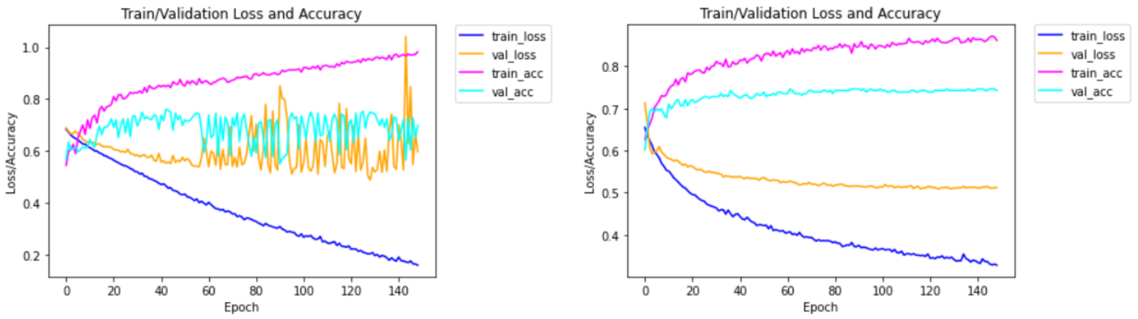
Figure 5.2: Training curves and ROC using: whole images; edema label only containing slices with edema annotations; T1 sequences; with pre-training; random perspective, horizontal flip and gaussian blur as data augmentation; max image intensity as normalization.

5.2 Prediction using ROIs

In order to improve the results obtained thus far, and inspired by the work developed by Lee et al [7], a ROI was extracted from the central region of the scans. The goal of this approach is to reduce the noise present in the images, and therefore reduce their complexity, so that the CNN can learn the necessary features to predict ELMSI. Since the position and orientation of the knee slightly varies depending on the patients, a ROI defined by constant coordinates would be either overly aggressive (by cutting out edges of the bones, and therefore edema regions), or ineffective. To extract an efficient ROI, the width was shortened until the dark background was removed, and then the images were cropped from 20% to 80% of their height. The relevant results on the test set are presented in Table 5.3, where only slices containing edema annotations were labelled as edema. The training curves of the models with and without transfer learning, both without data augmentation and max image intensity as normalization, are present in Figure 5.3. The training curves and ROC of the model with highest accuracy on the test set are present in Figure 5.4.

Table 5.3: Results on the test set using ROI when the dataset only contains, for the edema label, slices with edema annotations.

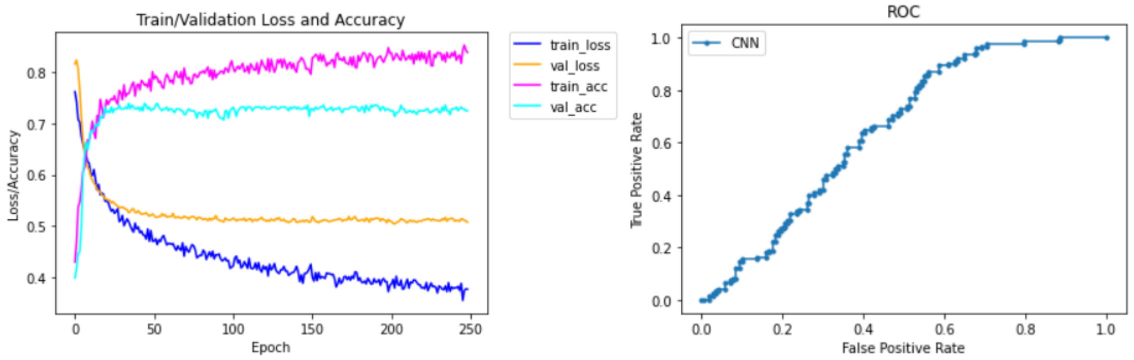
MRI	Transfer Learning	Data augmentation	Normalization	Best epoch	Accuracy	AUC
T1	-	-	Max image intensity	129	0.567	0.664
T1	ImageNet	-	Max image intensity	62	0.546	0.587
T1	ImageNet	Random perspective Horizontal flip	Max intensity	39	0.553	0.568
T1	ImageNet	Random perspective Horizontal flip	Max image intensity	51	0.608	0.651
T1	ImageNet	Random perspective Horizontal flip	Muscle ratio	44	0.583	0.595
T2	ImageNet	Random perspective Horizontal flip	Max image intensity	49	0.587	0.611



(a) Training curves using: ROI; T1 sequences; no pre-training; no data augmentation; max image intensity as normalization.

(b) Training curves using: ROI; T1 sequences; with pre-training; no data augmentation; max image intensity as normalization.

Figure 5.3: Training curves using ROI.



(a) Accuracy and loss of the train and validation sets throughout the training.

(b) ROC in the test set.

Figure 5.4: Training curves and ROC of the best model using: ROI; T1 sequences; with pre-training; random perspective and horizontal flip as data augmentation; max image intensity as normalization.

It is possible to see that transfer learning improved the training curves, specially by stabilizing the validation curves. It is again possible to conclude by the convergence of the curves, that the model has enough capacity for the complexity of the dataset. Although the test set accuracy did not improve considerably, compared to the previous experiment of using whole images, the behaviour of the training curves of the best model looks better. The validation accuracy during training is higher, and the validation loss is lower. The model is still over-fitting in relatively short period, and the relatively big gap between the validation loss and training loss shows that the model is still not able to generalize well.

5.3 Prediction using bone segmentations

The annotations present in UHCSJ Dataset V2 made it possible to obtain bone segmentations. Since there are no non-edema patients in this version of the dataset, segmentations with edema regions were labelled as edema, while segmentations without edema annotations were labelled as non-edema. This process was influenced by Chuah et al [21], where slices of edema patients that did not contain ELMSI in them, appeared to show no differences compared to slices of non-edema patients. Since the dataset was unbalanced, oversampling on the non-edema label was performed for the training set. The goal of using bone segmentations was to reduce the noise and complexity of the images, by removing every tissue and structure other than bones. Since the dataset is still slightly unbalanced, with the predominant label being edema, accuracy was replaced by balanced accuracy as the metric to choose the best model. The relevant results are detailed in Table 5.4, while the training curves of the model using no transfer learning and no data augmentation are present in Figure 5.5. The training curves and ROC of the model with highest balanced accuracy on the test set, using transfer learning and data augmentation, are present in Figure 5.6.

Table 5.4: Results on the test set using bone segmentations.

MRI	Transfer Learning	Data augmentation	Normalization	Best epoch	Balanced Accuracy	AUC
T1	-	-	Max image intensity	31	0.59	0.641
T1	ImageNet	-	Max intensity	12	0.477	0.551
T1	ImageNet	-	Max image intensity	11	0.556	0.586
T1	ImageNet	-	Muscle ratio	8	0.511	0.464
T1	ImageNet	Random perspective Horizontal flip	Muscle ratio	6	0.579	0.559
T2	ImageNet	Random perspective Horizontal flip	Max image intensity	60	0.583	0.618

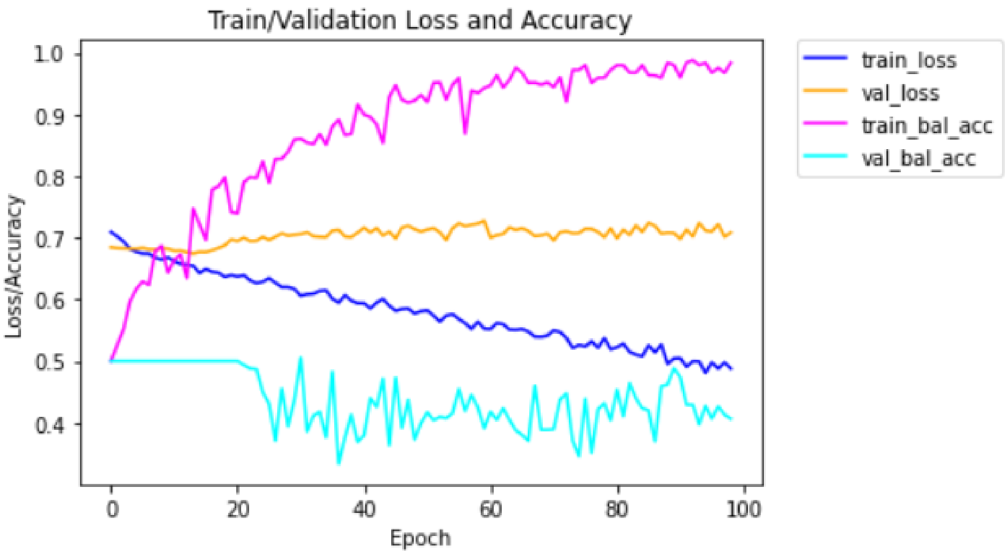
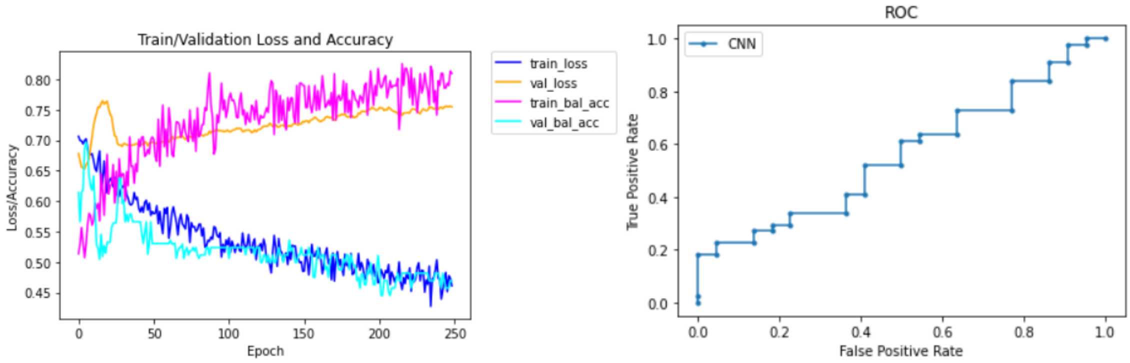


Figure 5.5: Training curves using: bone segmentations; T1 sequences; no pre-training; no data augmentation; max image intensity as normalization.

Using transfer learning and data augmentation improved the behaviour of the training curves, but even with these techniques it is apparent that the dataset was too small. The training curves were not very stable, the validation loss maintained high and the rapid decrease of the validation



(a) Accuracy and loss of the train and validation sets throughout the training.

(b) ROC in the test set.

Figure 5.6: Training curves and ROC using: bone segmentations; T1 sequences; with pre-training; random perspective and horizontal flip as data augmentation; muscle ratio as normalization.

accuracy is not what is desired. It is possible that changing hyper-parameters such as batch size could have made the curves more stable. The results on the test set also did not improve compared to using a ROI. It is also still not possible to assess which normalization technique provides better results, since all results were rather poor.

Since adding more samples to the dataset was not possible, and in an attempt to reduce the resolution of the images even further, experiments were ran where the images were resized to 28x28, 56x56 and 112x112 before being resized to 224x224. This also did not improve the classification of ELMSI, which means a new technique needed to be explored for this dataset.

5.4 Prediction using high intensity masking

In order to find a different alternative to the experiments conducted earlier, the high intensity masking technique was applied to the UHCSJ Dataset V2. The experiments were ran on fluid-sensitive sequences due to their higher sensitivity to bone marrow changes. The results are present in Table 5.5, the training curves of the model without transfer learning and data augmentation are present in Figure 5.7. The training curves of the model with highest balanced accuracy on the test set are present in Figure 5.8.

Table 5.5: Results on the test set using high intensity masking.

MRI	Transfer Learning	Data augmentation	Best epoch	Balanced Accuracy	AUC
Fluid-sensitive	-	-	37	0.55	0.971
Fluid-sensitive	ImageNet	-	32	0.588	0.683

Training the model without transfer learning yielded a high AUC in the test set due to the model predicting all 42 edema slices correctly; however, it misclassified 9 of the 10 non-edema images, which led to the low balanced accuracy. Perhaps using a different threshold for the predictions would improve the results. In comparison, the pre-trained model with unfrozen classifier correctly

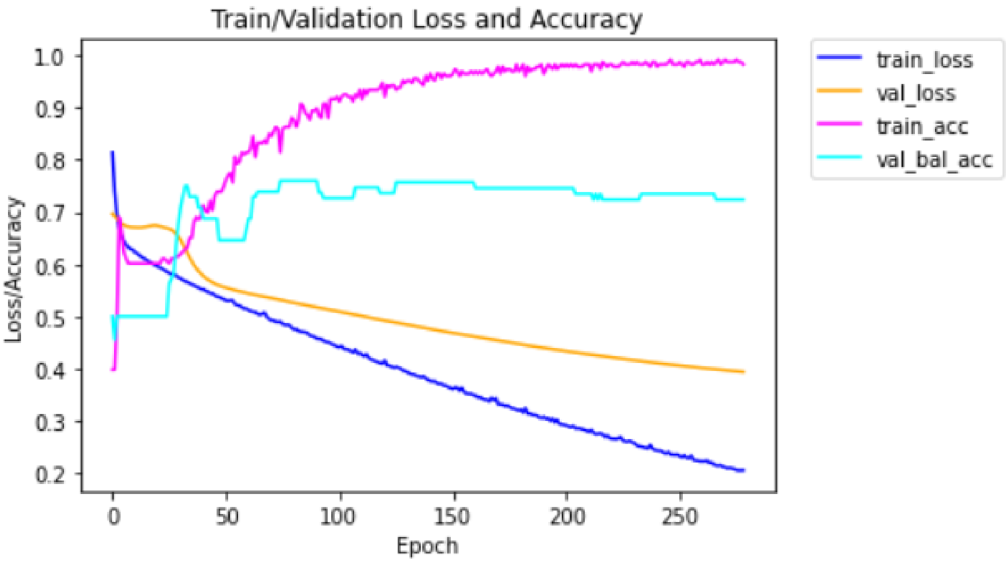
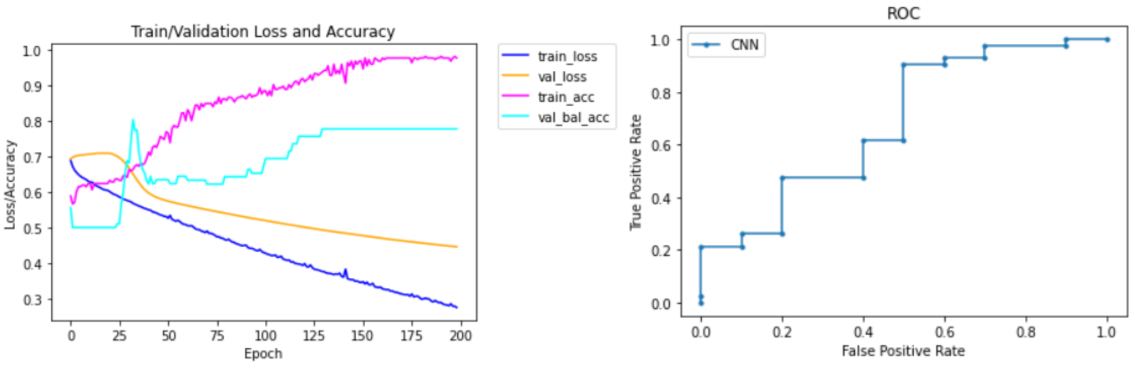


Figure 5.7: Training curves using: high intensity masking; fluid-sensitive sequences; no transfer learning; no data augmentation.



(a) Accuracy and loss of the train and validation sets throughout the training.

(b) ROC in the test set.

Figure 5.8: Training curves and ROC using: high intensity masking; fluid-sensitive sequences; with pre-training; no data augmentation.

predicted 2 of the 10 non-edema slices, while classifying 41 of the 42 edema samples correctly. It seems that the model is biased towards the edema label. Data augmentation did not improve the results, and using transfer learning did not seem to have a big impact on the training curves. The validation balanced accuracy has high values during training, although the loss also continues to be high, probably due to the very small size of the dataset. The model also appears to over-fit quickly. Despite these drawbacks, the method showed promising results if the false positives could be improved.

5.5 Fine-tuning the models

To fine-tune the models, two different approaches were used: progressively unfreezing layers before beginning the training; progressively unfreezing layers during training. For the former case, for the high intensity masking technique, the best result is present in Table 5.6, which corresponds to freezing 6 of the 10 layers of the Resnet-18; for the latter, for the same method, the result of progressively unfreezing a layer every 20 epochs is shown in Figure 5.9. The confusion matrix for the best result of the high intensity masking method is present in Table 5.7. The results on the test set of fine-tuning the best models of previous experiments is present in Table 5.8.

Table 5.6: Best result on the test set, after fine-tuning the model trained with the high intensity masking technique.

MRI	Transfer Learning	Data augmentation	Best epoch	Balanced Accuracy	AUC
Fluid-sensitive	ImageNet w/ 6 frozen layers	-	61	0.852	0.964

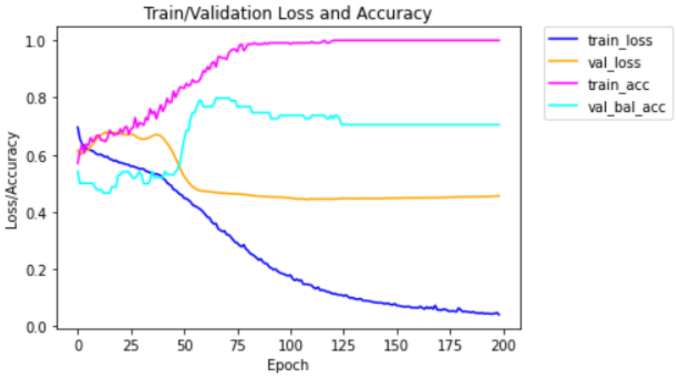


Figure 5.9: Training curves when unfreezing a layer every 20 epochs, using: high intensity masking; fluid-sensitive sequences; with pre-training; no data augmentation.

Table 5.7: Confusion matrix of the best model.

		Predicted class	
		Non-edema	Edema
True class	Non-edema	8	2
	Edema	4	38

Table 5.8: Best result on the test set, after fine-tuning the best performing models of previous experiments.

Image type	MRI	Transfer Learning	Data augmentation	Best epoch	Accuracy	AUC
Whole image	T1	ImageNet w/ 6 frozen layers	Random perspective Horizontal flip Gaussian blur	72	0.551	0.562
ROI	T1	ImageNet w/ 7 frozen layers	Random perspective Horizontal flip	88	0.591	0.673
Bone segmentation	T1	ImageNet w/ 6 frozenl layers	Random perspective Horizontal flip	12	0.568	0.550

It appeared that fine-tuning the best models of previous experiments did not lead to better results. It is possible that these approaches are way too hindered by the small size of the datasets. Looking at the training curves when progressively unfreezing layers during training, using the high intensity masking technique, the highest validation balanced accuracy occurred around epoch 70, which corresponds to the period where 6 layers were frozen. This may support the fact that the best performing model on the test set was the one with 6 frozen layers from the beginning of training. After that mark, the validation balanced accuracy decreases, but that may be due to the model already being over-fit. Perhaps changing the learning rate at every checkpoint would lead to more conclusive results. The balanced accuracy on the test set improved massively due to the model now correctly predicting 8 of the 10 non-edema patients. This result was a big improvement on techniques used previously in this thesis, and had a very promising outcome.

In an attempt to assess the performance of the high intensity masking method, the dataset was shuffled 5 times to produce 5 different train/validation/test sets, and the CNN was trained and tested in each experiment. Additionally, in order to determine the robustness of this method on non-edema slices, the 36 segmentations of bones of non-edema patients extracted from UHCSJ Dataset V1 were added to the test set. The results are present in Table 5.9.

Table 5.9: Results on the test set, using 5 different distributions of the dataset.

Run	MRI	Transfer Learning	Data augmentation	Balanded accuracy without new non-edema slices	Balanded accuracy with new non-edema slices
#1	Fluid-sensitive	ImageNet w/ 6 frozen layers	-	0.852	0.739
#2	Fluid-sensitive	ImageNet w/ 6 frozen layers	-	0.791	0.687
#3	Fluid-sensitive	ImageNet w/ 6 frozen layers	-	0.801	0.630
#4	Fluid-sensitive	ImageNet w/ 6 frozen layers	-	0.759	0.613
#5	Fluid-sensitive	ImageNet w/ 6 frozen layers	-	0.760	0.693
Average				0.792	0.672

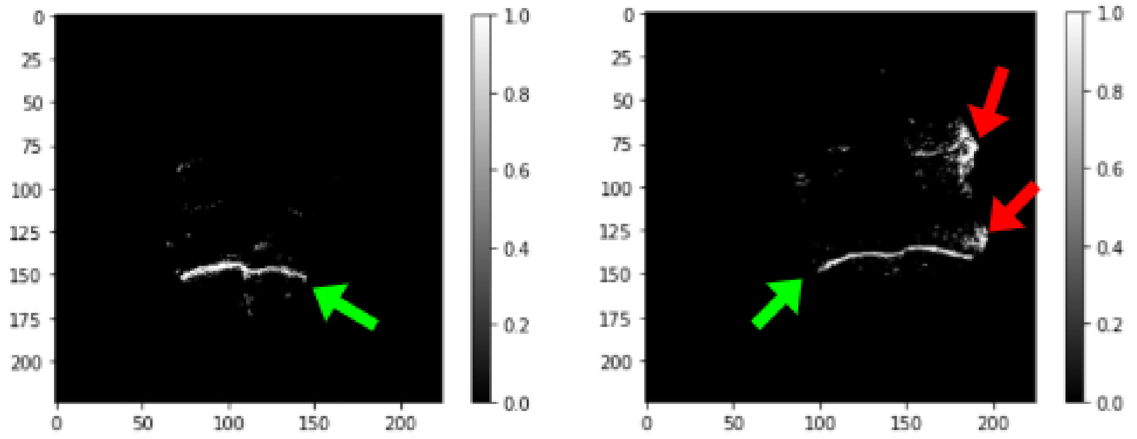
The model achieved satisfactory predictions, reaching an average balanced accuracy of 0.792 on the test sets without non-edema patients segmentations, and 0.672 on the test sets with non-edema patients segmentations. When looking at the slices where the model misclassified, it was clear that it was under-performing more on the non-edema label, specially in samples which contained noise that looked like edema regions. This technique should be further explored in a larger dataset to truly determine its viability.

5.6 Discussion and limitations

Regarding the previously presented results, when feeding whole images from the UHCSJ Dataset V1 to the model, the training curves seemed fairly standard, although the validation loss was quite high and the results on the test set were lackluster. For the edema label, it appears that using images that contain edema regions lead to better performances than using every image from edema patients. Using ROIs centered in the joints of the femur and tibia did not improve the test set accuracy, but the training curves were much better, as the validation loss decreased and the accuracy increased. It is possible that using a bigger dataset could considerably improve the results of this last approach, although such a task is hard to accomplish in the medical field. Utilizing the bone segmentations extracted from UHCSJ Dataset V2 lead to disappointing results, as the training curves were somewhat unstable and the validation loss was quite high. The test set results were also not better than the results using ROIs. The behaviour of the training curves could be explained by the even smaller dataset and the complexity of the images. Resizing the image to shapes smaller than 224x224 (e.g. to 112x112), and then reshaping them to 224x224 to be fed to the model, did not improve results despite lowering the resolution of the image. It was expected that this approach could lead to better results than previous approaches, since it removes some complexity of the images by solely feeding bone tissue to the model, but that was not the case, possibly due to the size of the dataset. Using data augmentation and transfer learning was overall beneficial. It was not possible to gather convincing conclusions on which normalization technique was better, since all results were rather poor and similar between each other; however, the max intensity approach was never the best result.

The method that presented the best results was the prediction of ELMSI on a pre-trained and fine-tuned Resnet-18, using images that resulted from the high intensity masking technique. By freezing the first 6 layers of the Resnet-18, the model achieved a balanced accuracy in the test set of 0.852. After shuffling the dataset 5 times and training the CNN in each experiment, the model achieved a mean balanced accuracy of 0.792 in the test set. When adding 36 slices of non-edema patients, extracted from the UHCSJ Dataset V1, the mean test set balanced accuracy was 0.672. These results were promising, although it appears that the model under-performed on the non-edema label. To further assess the viability of this method, a bigger dataset would improve the learning process of the CNN.

It should be noted that applying the high intensity masking technique on fluid-sensitive sequences does not always produce the desired results, which could affect the performance of the model. Ideally, applying this method on non-edema slices would produce near-total black images, as this means there are no outlier regions when comparing with normal bone marrow. However, due to the dataset containing under-18 patients, structures like the growth plate are often present in the result (Figure 5.10a). Additionally, due to the higher sensitivity of fluid-sensitive sequences, the final image can include noise that wrongfully appears to be an edema region (Figure 5.10b). The high intensity masking technique highlighting structures such as growth plate should ideally not be an issue for the CNN, as the model itself needs to be able to distinguish edema regions from such structures, if it wants to succeed in predicting ELMSI in children and young adults. However, wrongfully highlighting noise that in the final image appears like an edema area may hinder the prediction of ELMSI. These noise areas were present on some non-edema slices, which could explain the under-performance of the model on this label, since it could be mistaking them for ELMSI regions. Therefore, the highlighting of noise poses as a limitation to this method. Perhaps a procedure different than using bone mean and STD for the extraction of the range of edema intensities, could lead to more robust results. Also, since T1 sequences are less sensitive, perhaps applying a similar method to those images could lead to the detection of less noise, specially in non-edema patients. An ideal pipeline, using the high intensity masking technique, could combine the prediction of ELMSI on both types of slices of MRI, or perhaps a way to fuse the information of both images before feeding the result to the CNN. Either way, utilizing the T1 sequences could greatly improve the performance of the model on non-edema slices. Another limitation is the fact that this method requires the bone masks in order to remove intensities detected in other tissues. Therefore, this technique is completely reliant on the automatic segmentation of bones, which could be achieved by the use of deep learning.



(a) Example of detection of growth plate on non-edema slice.

(b) Example of detection of growth plate, and noise that appears to be an edema region.

Figure 5.10: Examples of results after applying the high intensity masking technique on non-edema patients. The green arrows highlight the growth plate, while the red arrows indicate noise that looks like an edema region.

5.7 Summary

In this chapter, the relevant results of every experiment were presented and discussed. Starting with the whole images extracted from the UHCSJ Dataset V1, it was concluded that the best approach to define the edema label was to only use slices which contained annotations of ELMSI by clinicians. However, the model was not able to achieve good results on the test set. By feeding a ROI of the bone junctions, the model achieved better training curves, although the results on the test set were still rather poor. This could be due to the high complexity and resolution of the images, as well as the small size of the dataset. By feeding the bone segmentations taken from the UHCSJ Dataset V2, the learning curves were more unstable and the results on the test set were not better. The behaviour of the learning curves could indicate that the model is not generalizing well, possibly due to the even smaller size of the dataset. In an attempt to reduce the complexity and resolution of the images even further, the shape of the images was decreased before being resized to 224x224 and fed to the CNN. This did not improve the results. The ELMSI prediction using the high intensity masking technique showed promising results, although further work could be developed. It was clear that the model struggled more with non-edema images, specially when these contained noise that looked like edema areas. Nonetheless, if an automated method for segmentation of bones is ever created, the high intensity masking method could be a possible solution for the classification of ELMSI.

Chapter 6

Conclusions and future work

ELMSI is characterized by regions of high signal intensity in fluid-sensitive sequences, and areas of intermediate to low signal intensity in T1 sequences. Although its clinical relevance is still being determined, it is a condition often associated with pain and limb motion difficulty. Diagnosing ELMSI has a margin of error, due to the lack of guidelines for quantitative measurement, and dependence on the equipment, experience of the physician conducting the MRI, and skill of the reader. It can also be a time consuming process. In young patients, natural chemical changes that occur in the bone marrow can cause visible alterations in MRIs. Therefore, an automated method for predicting ELMSI on children and young adults would greatly assist clinicians.

In many medical problems, it is hard to obtain large datasets, which hinders the use of deep learning methods. The experiments performed in this thesis attempted to find a viable solution for the prediction of ELMSI using a small dataset, with the additional challenge of containing only under-18 patients. To reduce over-fitting, transfer learning and data augmentation were applied, and the models were fine-tuned in an attempt to increase their performance. Feeding the whole images to the Resnet-18 did not lead to satisfactory results, but it was concluded that, for the edema label, utilizing slices that contain edema areas is the best approach. Using ROIs centered in the joints of the femur and tibia improved the training curves and results, but these were still not satisfying. Feeding the bone segmentations to the model also lead to disapointing results, despite the expectation that, by removing tissues that are irrelevant to the problem, the CNN would perform better. It is likely that, for these last 2 approaches, a larger dataset would produce better results. The high intensity masking technique applied on fluid-sensitive sequences achieved promising results, although it has room for improvement.

For future work, the high intensity masking technique should be refined to improve results on non-edema cases. Perhaps a different method for extracting the ranges of edema intensities could lead to more robust results. Furthermore, developing a variation of this technique for T1 sequences, which are less sensitive than fluid-sensitive sequences, would make the combination of both MRI images possible. A more robust pipeline could combine the prediction of ELMSI on both slices, or contain a way to fuse both results into an image and then feed it to the CNN. Evaluating the performance of the model using Grad-Cam could also be interesting, in order to identify cases where the model is failing. If possible, the approaches conducted in this experiment should be run on larger datasets, which would benefit the training of the CNN.

References

[1] Davide Maraghelli, Maria Luisa Brandi, Marco Matucci Cerinic, Anna Julie Peired, and Stefano Colagrande. Edema-like marrow signal intensity: a narrative review with a pictorial essay. URL: https://doi.org/10.1007/s00256-020-03632-4, doi:10.1007/s00256-020-03632-4/Published.

[2] Erik F Eriksen. Treatment of bone marrow lesions (bone marrow edema). *BoneKEy Reports*, 4, 11 2015. doi:10.1038/bonekey.2015.124.

[3] W A Thiryayi, S A Thiryayi, and A J Freemont. Histopathological perspective on bone marrow oedema, reactive bone change and haemorrhage. *European Journal of Radiology*, 67:62–67, 2008. URL: https://www.sciencedirect.com/science/article/pii/S0720048X08000971, doi:https://doi.org/10.1016/j.ejrad.2008.01.056.

[4] Georg Schett. Bone marrow edema. *Annals of the New York Academy of Sciences*, 1154:35–40, 2009. doi:10.1111/j.1749-6632.2009.04383.x.

[5] F.M. Vanhoenacker and A. Snoeckx. Bone marrow edema in sports: General concepts. *European Journal of Radiology*, 62:6–15, 2007.

[6] João Pedro Viveiros Franco. Computer vision approach for bone marrow edema detection in mri and x-ray images, 2021.

[7] Kang Hee Lee, Sang Tae Choi, Guen Young Lee, You Jung Ha, and Sang Il Choi. Method for diagnosing the bone marrow edema of sacroiliac joint in patients with axial spondyloarthritis using magnetic resonance image analysis based on deep learning. *Diagnostics*, 11, 7 2021. URL: https://www.mdpi.com/2075-4418/11/7/1156, doi:10.3390/diagnostics11071156.

[8] R Tibrewala, E Ozhinsky, R Shah, S C Foreman, V Pedoia, and S Majumdar. Detecting hip osteoarthritic degenerative changes in mri using deep learning. *Osteoarthritis and Cartilage*, 27:S387–S388, 2019. URL: https://www.sciencedirect.com/science/article/pii/S1063458419304297, doi:https://doi.org/10.1016/j.joca.2019.02.387.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[10] W.A. Thiryayi, S.A. Thiryayi, and A.J. Freemont. Histopathological perspective on bone marrow oedema, reactive bone change and haemorrhage. *European Journal of Radiology*, 67:62–67, 2008. URL: https://www.sciencedirect.com/science/article/pii/S0720048X08000971?casa_token=IMoQA6Dddf8AAAAA:

8CrZVJPCzx1PF1LZhz48mb3KWwsH3S1oSmRoPCbZDm_3JYOekjVy_nHuPwLbasS9mxmhREKcp7yH.

[11] J B Vogler and W A Murphy. Bone marrow imaging. *Radiology*, 168:679–693, 9 1988. doi: 10.1148/radiology.168.3.3043546. URL: https://doi.org/10.1148/radiology.168.3.3043546, doi:10.1148/radiology.168.3.3043546.

[12] Sinchun Hwang and David M. Panicek. Magnetic resonance imaging of bone marrow in oncology, part 1. *Skeletal Radiology*, 36:913–920, 10 2007. URL: https://link.springer.com/article/10.1007/s00256-007-0309-3, doi:10.1007/S00256-007-0309-3.

[13] Bruno C Vande Berg Jacques Malghem Frederic E Lecouvet Baudouin Maldague, BC Vande Berg, J Malghem, Fe Lecouvet, and B Maldague. Magnetic resonance imaging of the normal bone marrow. *Skeletal Radiol*, 27:471–483, 1998. URL: https://link.springer.com/article/10.1007/s002560050423.

[14] Marcelo Batista Bonadio, Alipio Gomes Ormond Filho, Camilo Partezani Helito, Xavier MGRG Stump, and Marco Kawamura Demange. Bone marrow lesion: Image, clinical presentation, and treatment. *Magnetic Resonance Insights*, 10:1178623X1770338, 1 2017. doi:10.1177/1178623x17703382.

[15] Siegfried Hofmann, Josef Kramer, Anosheh Vakil-Adli, Nicolas Aigner, and Martin Breitenseher. Painful bone marrow edema of the knee: Differential diagnosis and therapeutic concepts, 7 2004. doi:10.1016/j.ocl.2004.04.005.

[16] Matías Costa-Paz, D.Luis Muscolo, Miguel Ayerza, Arturo Makino, and Luis Aponte-Tinao. Magnetic resonance imaging follow-up study of bone bruises associated with anterior cruciate ligament ruptures. *Arthroscopy: The Journal of Arthroscopic Related Surgery*, 17:445–449, 2001. URL: https://www.sciencedirect.com/science/article/pii/S0749806301821484, doi:https://doi.org/10.1053/jars.2001.23581.

[17] Berend C Stoel. Artificial intelligence in detecting early ra. *Seminars in Arthritis and Rheumatism*, 49:S25–S28, 2019. URL: https://www.sciencedirect.com/science/article/pii/S0049017219306559, doi:https://doi.org/10.1016/j.semarthrit.2019.09.020.

[18] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mr., 2019.

[19] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix., 2019.

[20] Evgeni Aizenberg, Edgar A.H. Roex, Wouter P. Nieuwenhuis, Lukas Mangnus, Annette H.M. van der Helm-van Mil, Monique Reijnierse, Johan L. Bloem, Boudewijn P.F. Lelieveldt, and Berend C. Stoel. Automatic quantification of bone marrow edema on mri of the wrist in patients with early arthritis: A feasibility study. *Magnetic Resonance in Medicine*, 79:1127–1134, 2 2018. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/mrm.26712, doi:10.1002/MRM.26712.

[21] Tong Kuan Chuah, Chueh Loo Poh, and Kenneth Sheah. Quantitative texture analysis of mri images for detection of cartilage-related bone marrow edema. pages 5112–5115, 2011. doi:10.1109/IEMBS.2011.6091266.

[22] Elisabeth von Brandis, Håvard B. Jenssen, Derk F. M. Avenarius, Atle Bjørnerud, Berit Flatø, Anders H. Tomterstad, Vibke Lilleby, Karen Rosendahl, Tomas Sakinis, Pia K. K. Zadig, and Lil-Sofie Ording Müller. Automated segmentation of magnetic resonance bone marrow signal: a feasibility study. *Pediatric Radiology*, 2 2022. URL: https://link.springer.com/10.1007/s00247-021-05270-x, doi:10.1007/s00247-021-05270-x.

[23] Qing Han, Yunfei Lu, Jie Han, AnLin Luo, LuGuang Huang, Jin Ding, Kui Zhang, Zhaohui Zheng, JunFeng Jia, Qiang Liang, Shuiping Gou, and Ping Zhu. Automatic quantification and grading of hip bone marrow oedema in ankylosing spondylitis based on deep learning. *Modern Rheumatology*, page roab073, 10 2021. URL: https://doi.org/10.1093/mr/roab073, doi:10.1093/mr/roab073.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL: https://arxiv.org/abs/1512.03385, doi:10.48550/ARXIV.1512.03385.