FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Computer Vision approach for bone marrow edema detection in MRI and X-ray images

João Pedro Viveiros Franco



Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Hélder P. Oliveira

Co-supervisor: Tânia Pereira

Co-supervisor: Silvia Costa Dias

August 2, 2021

Computer Vision approach for bone marrow edema detection in MRI and X-ray images

João Pedro Viveiros Franco

Mestrado Integrado em Engenharia Informática e Computação

August 2, 2021

Abstract

Bone marrow edema (BME) or bone marrow lesion (BML) is the name attributed to an observed signal change within the bone marrow in magnetic resonance imaging (MRI), more commonly detected in the hip, knee, ankle or foot. The condition is generally diagnosed when areas of high signal in fluid-sensitive sequences - T2 or proton-density (PD) weighted images with fat-suppression, or short tau inversion-recovery (STIR) - and medium or low signal intensity areas in T1-weighted images are detected in MRI. The recent developments in machine learning techniques, such as deep learning with the use of convolutional neural networks(CNN), allow increasingly more complex problems, such as the ones found on medical imaging, to be solved with good results. The purpose is to explore bone marrow edema, focusing on their detection in MRI and in X-ray imaging with the help of appropriate machine learning algorithms. The possibility of detecting the condition in X-ray imaging could prove to be a very useful tool as it is a more widespread technique and thus would help on an early diagnosis and intervention; however, this type of imaging appears to have very low correlation with the detection of BME since X-rays usually don't reveal signs to suggest the pathology. This low correlation is coherent with the fact that the best classifier used in the tests performed in this experiment achieving merely 56.3% of accuracy. With gathered images from 72 patients, through a patch-based strategy, both a random forest and a CNN were trained and tested, with the latter achieving 72.1% balanced accuracy.

Keywords: Machine Learning, Deep learning, Computer Vision, MRI, X-ray, Bone Marrow Edema, Bone Marrow Lesions

Resumo

O edema medular ósseo é o nome dado a uma alteração de sinal observada na medula óssea em imagens de ressonância magnética (RM), mais frequentemente encontrada na área da anca, joelho, tornozelo ou pé. Habitualmente este problema é diagnosticado quando são detetadas em RM zonas de hipersinal em sequências sensíveis ao líquido - imagens ponderadas em T2 ou densidade protónica com supressão de gordura, ou short tau inversion-recovery (STIR) - e áreas de sinal intermédio ou hipossinal em T1. Os últimos avanços em aprendizagem automática aplicada ao contexto médico, como é o exemplo da aprendizagem profunda com o uso de redes neuronais convolucionais, permitem que problemas cada vez mais complexos, tais como os de imagiologia médica, serem resolvidos com bons resultados. O objetivo é explorar o problema do edema medular ósseo com foco na sua deteção em imagens de RM e radiografia (Rx) com a ajuda das técnicas de aprendizagem automática apropriadas. A possibilidade de deteção deste achado em Rx poderá ser útil visto ser uma técnica mais acessível, facilitando assim o seu diagnóstico mais precoce e intervenção. Contudo, este tipo de imagem aparenta ter um grau de correlação muito baixo com a deteção de edema medular ósseo, dado que Rx habitualmente não revela sinais que sugiram edema medular ósseo. Esta baixa correlação é coerente com o facto do melhor classificador utilizado nos testes desta experiência apenas atingir uma precisão de 56%. Com imagens de 72 pacientes, utilizando um estratégia à base de fragmentos de imagens, tanto um *random forest* tal como uma rede neuronal convolucional foram trainadas e testados, sendo que a última atingiu 72.95% de precisão equilibrada.

Keywords: Aprendizagem Automática, Aprendizagem Profunda, Visão por Computador, Ressonância Magnética, Raio X, Edema Medular Ósseo

Acknowledgements

This thesis could not have been possible without the support of many people. Firstly, I would like to thank my close family for the undivided support and love they have given me throughout the years, which has made me the person I am today. For that and much more, no words could fully express my gratitude. I would also like to thank Hélder Oliveira, Tânia Pereira, Silvia Dias and Diogo Carvalho for going above and beyond in helping me during this entire process. Thank you to Miguel and Gabriela Viveiros for all the support these last years. I would like to also thank my friends.

João Pedro Viveiros Franco

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Motivation . . . 2
- 1.3 Objectives . . . 2
- 1.4 Contributions . . . 2
- 1.5 Document Structure . . . 3

2 State of the Art **5**
- 2.1 Background . . . 5
  - 2.1.1 Identifying bone marrow edema on MRI scans . . . 5
  - 2.1.2 Identifying bone marrow edema on X-Ray imaging . . . 6
  - 2.1.3 Treatment . . . 7
- 2.2 Machine Learning . . . 7
  - 2.2.1 Traditional machine learning . . . 7
  - 2.2.2 Deep learning . . . 12
  - 2.2.3 Metrics and the confusion matrix . . . 16

3 Methodology and Experimental Setup **19**
- 3.1 Data Preparation . . . 19
  - 3.1.1 Class imbalance . . . 24
  - 3.1.2 Age in BME detection . . . 24
- 3.2 Algorithm Validation . . . 26
- 3.3 Models . . . 27

4 Results and Discussion **31**
- 4.1 Age in BME detection . . . 33
- 4.2 Discussion . . . 34

5 Conclusions **37**
- 5.1 Future work . . . 37

A Apendix **39**

References **41**

List of Figures

2.1 Example of a high signal in fluid-sensitive sequence with findings of edema . . . 6
2.2 Comparison of different scans of a patient with edema . . . 6
2.3 The Support Vector Machine model and the separation of classes inside the hyper-space . . . 9
2.4 Classification of hand-written digits using SVMs . . . 9
2.5 Visual representation of bagging and boosting . . . 10
2.6 Example of random forest model performing multi-class classification . . . 12
2.7 The convolution process applied on convolutional layers of CNNs . . . 14
2.8 The abstraction process of a deep learning model . . . 15
2.9 Visual representation of the inner workings of convolutional neural networks . . . 15
2.10 Schema denoting the multiple components of the confusion matrix . . . 17

3.1 A single slice of an MRI scan with multiple annotated edema findings . . . 20
3.2 Image of an MRI scan with two non-edema regions annotated . . . 20
3.3 Example of an X-ray image . . . 21
3.4 The topology of the DICOM format . . . 23
3.5 The process of reading DICOM image files . . . 24
3.6 Schema representing the dataflow of patches to the machine learning models . . . 24
3.7 Image of an MRI scan with annotated circles for normalization and an edema region 25
3.8 Image of a patch erogenously capturing pixels outside of the edema region . . . 25
3.9 The distribution of patches generated between edema and non-edema patients . . . 26
3.10 The box plot of patients with or without edema findings . . . 26
3.11 Visual representation of the chosen CNN topology . . . 28
3.12 The k-fold cross validation method . . . 30

4.1 The ROC curve of an execution of the CNN model . . . 32
4.2 The confusion matrix of an execution of the CNN model . . . 32
4.3 First example of an execution of the CNN model on an edema patch . . . 33
4.4 First example of an execution of the CNN model on an edema patch . . . 33
4.5 The scatter plot of testing accuracy and patient mean age . . . 34

A.1 Second example of the best classifier on an edema image . . . 39
A.2 Third example of an execution of the CNN model on an edema patch . . . 39
A.3 Second example of an execution of the CNN model on a non-edema patch . . . 40
A.4 Third example of an execution of the CNN model on a non-edema patch . . . 40

List of Tables

3.1 The attributes of the DICOM format used 23
3.2 Text-based representation of the chosen CNN topology 27
3.3 The random forest parameters and the multiple values considered 29

4.1 Results obtained in the multiple tests done 32
4.2 Values generated by the calculation of the Spearman's rank correlation coefficient 34

Abbreviations

AUC	Area Under the Curve
BME	Bone Marrow Edema
BML	Bone Marrow Lesions
CNN	Convolutional Neural Network
DICOM	Digital Imaging and Communications in Medicine
ID	Identifier
MCC	Matthews Correlation Coefficient
MRI	Magnetic Resonance Imaging
PD	PD-weighted images
RF	Random Forest
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SVM	Support Vector Machine
SOP	Service-Object Pair
T1	T1-weighted images
T2	T2-weighted images

Chapter 1

Introduction

1.1 Context

Bone marrow edema (BME) is the name given to an observed signal change on magnetic resonance imagining (MRI) scans, more specifically, areas of high signal in fluid-sensitive sequences - T2 or proton-density (PD) weighted images with fat-suppression, or short tau inversion-recovery (STIR) - and intermediated or low signal intensity areas in T1-weighted images in MRI. The term was initially used in 1988 by Wilson et al.[60] and is usually found in areas like the hip, knee, ankle, foot, shoulder or spine. The presence of BME is an unspecific but sensitive sign with a widely differential diagnosis, that may act as a guide that leads to a systematic and correct interpretation of the MR examination. Fluid-sensitive MR sequences are more sensitive than T1-weighted images in the detection of BME. While MRI is the predominant detection method for bone marrow edema it has its own set of problems, one of them being the changes happening at the biochemical level in the bone marrow. More specifically, bone marrow composition changes occur physiologically along with the development of the skeleton and with non physiological events such as traumatic insult, disease onset or with some types of therapy. This makes the detection and diagnosis process more difficult[36]. However, a recent study has shown some correlation between BME and changes in water and fat percentage, which can be measured using techniques such as Magnetic Resonance Spectroscopy[50] or using more robust fat suppresion techniques in fluid-sensitive sequences, e.g. the Dixon method[45]. BME is usually associated with conditions like trauma or osteoarthritis but can be observed in patients with other illnesses. Rheumatoid arthritis, ankylosing spondylitis, psoriasis, osteomyelitis and diabetic foot are just some of the conditions where bone marrow edema has also been identified. This signal change in MRI scans has generated some controversy with regards to its causes and clinical impact; however, it has been increasingly more correlated with pain and making limb motion more difficult[11]. There are various ways to treat it, usually through the means of surgery or medication such as prostacyclin analogs and bisphosphonates. Prostacyclin is a vasodilator that helps with the passage of blood in tissues with bad blood flow and bisphosphonates is a class of medication that help preserve bone density. These medications are very helpful for pain relief[18].

1.2 Motivation

Bone marrow edema is a feature often found in both athletic and non-athletic individuals, with the athletic ones sometimes being asymptomatic[53]. Some studies have suggested that BME is not a good indicator for severe pathological complications in these asymptomatic cases[14][47][39]; however, other studies have associated it with pain[38][33]. Bone marrow edema is also associated with common pathological issues such as arthritis, osteoporosis and ankle pain so developing a solution to aid in solving these problems would benefit a lot of patients that are suffering from the symptom since, to the extent of the research done, it still has no standardized method for automated detection yet. Furthermore, if the patient has osteoarthritis, BME will generally get worse. X-ray imaging plays a minor role in the diagnostics of BME. This technique can be useful revealing findings that allow the diagnosis of conditions in which BME association is common. Detection of osteopenia that could be related with BME by health professionals in X-rays is extremely challenging and much less frequent and accurate than in MRI[63]; even so, if this detection was somewhat possible using machine learning models, physicians could be aware of this finding much sooner, due to X-ray imaging being usually the first imaging technique performed in the clinical setting. By having a streamlined tool for this effect, one could reduce the variability and general subjectivity of human identification. Having an accurate and reliable method to prematurely detect this finding could help health professionals in the management of these patients' diagnose with a prompt intervention. It could also help health professionals perform early intervention. In a recent study, Felson et al[21] found that 272 out of 351 (77.5%) persons with knee pain had bone marrow lesions so this type of solution may help patients from being affected by pain in their everyday life. Having more tools could also improve the understanding of how the symptom progresses and help to reduce errors made during diagnostic and treatment, improving the quality of life of patients[41].

1.3 Objectives

The general objective of this dissertation is to explore the problem surrounding bone marrow edema. To do this, a set of machine learning models will be carefully trained and tested to determine the feasibility of detecting the condition in magnetic resonance imaging scans and X-rays in an automated computer environment. With advancements made in this area, it is believed that radiologists could be assisted with tools to automatically detect and possibly quantify BME in MRI, also helping in monitoring patients' progression as well as act as an early detection methods in Rx imaging, which is not possible at the moment.

1.4 Contributions

The following contributions were made from the experiment:

- The development of a procedure for extracting annotation and image data from DICOM files
- The development of a machine learning algorithm that can detect bone marrow edema to a certain degree
- The experiments made, parameters selected and results obtained and discussed

1.5 Document Structure

This document has 5 chapters, with the following structure:

- In Chapter 1, Introduction, the problem in study, the goals as well as the motivation is described;
- In Chapter 2, State of the Art, the current knowledge of the studied matter as well as the most advanced techniques are presented;
- In Chapter 3, the experimental implementation of the solution is described and presented, along with the parameters chosen for testing;
- In Chapter 4, the metrics and performance of the algorithm is discussed;
- In Chapter 5, the inferences made from results are presented in a final manner.

Chapter 2

State of the Art

In this chapter, it will be described what exactly is Bone marrow edema (BME) , how it is usually detected, the many treatment methods available and the most appropriate technology to be used for detection and quantification of the symptom available as of the time of writing of this document.

2.1 Background

BME or bone marrow lesions (BML) is the generic name given to specific alterations found in the signal on magnetic resonance imagining (MRI) scans[11]. The term was first used in the year 1988 by Wilson et al[60] to describe the condition in patients with transient osteoporosis in the knee or hip. At the time, these patients had debilitating pain in the affected area but this symptom was spontaneously resolved in all 10 cases, which led to the name "transient marrow edema syndrome", also coined by the authors. This clinical finding can be a symptom of many different conditions, usually trauma or osteoarthritis but can also be present in patients diagnosed with a wide spectrum of illnesses. Rheumatoid arthritis, ankylosing spondylitis, psoriasis, osteomyelitis and diabetic foot are just some of the conditions where BML has also been identified[11]. Patients affected by this usually complain about pain or difficulties with motion in the affected body area, but can also show no symptoms.

2.1.1 Identifying bone marrow edema on MRI scans

Magnetic resonance imaging scans are by far the most common method to detect bone marrow edema. It is detected by high signal intensity in fluid-sensitive sequence, such as on figure 2.1, and low signal intensity in T1-weighted scans[18]. Logically, this type of image has a high correlation with the detection of edema. In figure 2.2, one can see the difference between the different image types and how edema in shown on each one.

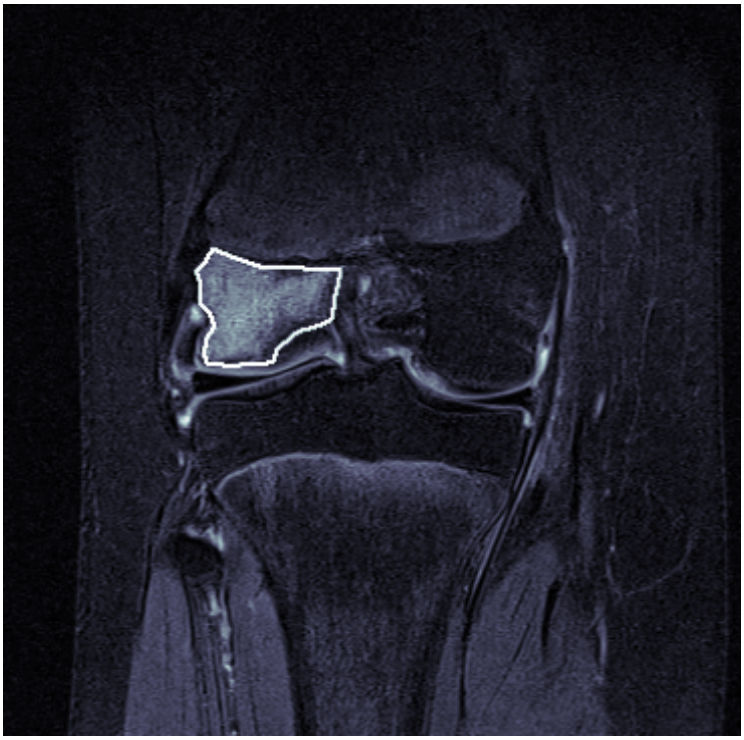


Figure 2.1: Example of a high signal in fluid-sensitive sequence with findings of edema.

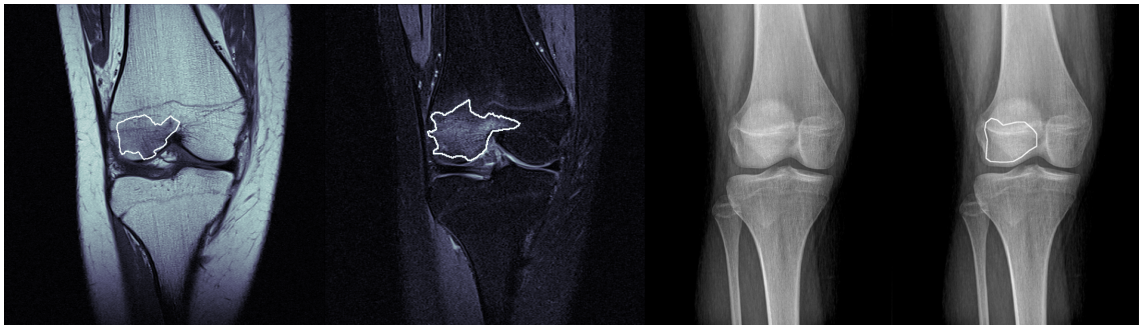


Figure 2.2: Comparison of different scans of a patient with edema. From left to right, the T1-weighted MRI image, the fluid-sensitive sequence image, the original X-ray and the X-ray with the annotated region where the edema was found in MRI.

2.1.2 Identifying bone marrow edema on X-Ray imaging

Regarding X-Ray imaging, there has not been as much research done in the detection of BME as in MRI. Some articles have even gone as far as arguing that the manifestation of the condition is not visible at all in X-ray scans[18]. While that may be true for humans, computer vision models like CNNs have been noted to have better accuracy than human beings on certain tasks. One such example is the usage of CNNs for characterizing bioleaching bacterial biofilm composition by Buetti-Dinh et al.Researchers managed to achieve an accuracy of 92.8% on the classification problem, while human experts fell behind at 51.5%[12].

2.1.3 Treatment

There are numerous ways to treat the symptoms of BME. If the patient is willing to undergo surgical intervention there is a combination of core decompression, i.e. drilling holes in the area affected by BME, along with possible injections of hydroxyapatite cement[62] or autologous bone marrow stem cells, depending on the cause of the BME[29]. For less invasive manoeuvres, extracorporeal shock wave therapy is a possible solution. This therapy is compromised of sending mechanical shocks of a certain magnitude to the area affected by the lesions. Regarding pharmaceuticals, some suggested medication includes prostacyclin analogs and bisphosphonates. Prostacyclin is a vasodilator that helps with the passage of blood in tissues with bad blood flow and bisphosphonates is a class of medication that help preserve bone density. These medications are very helpful for pain relief[18].

2.2 Machine Learning

Machine learning is the general study of computational algorithms that try to map a certain input to a certain output through complex rules. This can be compared to how a regression model in regression analysis tries to map a series of points to a mathematical function, but with more complex problems and solutions. In this field, there is gathered experience in the form of a series of input data and the corresponding output, usually collected from some sort of human process (e.g., visually identifying the breed of a dog, classifying the subject of a book, etc) or data reading from an observed phenomenon (e.g., weather readings for forecasting, stock market statistics, etc). This experience is initially split into train, test and sometimes validation sets for the different stages of the usual workflow process. The train set is fed to the machine learning model (algorithm), either entirely at the same time or with a specific strategy (e.g., through a series of batches), while the algorithm changes its inner parameters to adapt to it. After this is done, the model receives the input data of the previously mentioned test set, tries to predict its output and various metrics are calculated from the comparison of the predicted output with the observed output of this set. Some methods that train in an iterative manner, like CNNs, utilize the validation set to measure the performance of the model at each stage.

The data is one of the most important components of this procedure and has a deep effect throughout the entirety of the process. Making sure that it explores the data space in a thorough and meaningful way is key to having good results.

2.2.1 Traditional machine learning

Traditional machine learning is loosely defined by the process done by machine learning models that have existed before the increase in popularity of deep learning models. Some of these models include nearest neighbours, decision trees, logistic regression, random forests and support vector machines, among others. These types of models are generally faster during the training process than deep learning methods but are also slower in generating predictions as well as requiring

more time and domain knowledge in the data preparation stages. Due to these factors, traditional machine learning algorithms are still very relevant and widely used in the current machine learning landscape. The reason why deep learning is getting more attention is mainly due to the increasing size of data, the automatic process of selecting highly relevant features and the layered architecture capable of making analysis at different points of abstraction[16].

2.2.1.1 Support Vector Machines

Support vector machines are a type of machine learning model for supervised and unsupervised learning that is capable of solving problems for both classification and regression[9]. Developed by Vapnik and his colleagues in the 1990s, this algorithm works by constructing a set of hyperplanes in high-dimensional space to try to reach a ruleset for capable of separating data points with the aid of the kernel function, as can be seen exemplified on figure 2.3. These hyperplanes can be defined as a set of points such that the dot product is constant with a vector in the space, generating a set of orthogonal vectors[32]. Although it was initially proposed for linear classification, SVMs can perform non-linear classification with the aid of the kernel trick, allowing different kernels to be used[1]. The chosen kernel defines the type of function used by the support vector machine to separate points. By using complex kernels, the support vector machine can better model the separation between points, but also spends more time performing calculations. A schema denoting an example of digit classification can be seen on figure 2.4. Usual kernels include:

- Linear kernel
- Polynomial kernel
- Radial basis function (RBF) kernel
- Hyperbolic tangent (sigmoid) kernel
- Sigmoid kernel
- Multiquadric kernel
- Graph kernel
- Cauchy kernel

Support vector machines have been used to relative success in medical imaging tasks, e.g., Alquran et al.[4] managed to use support vector machines to detect and classify melanoma skin cancer images with a classification accuracy of 92.1%. These results are especially important due to the nature of the cancer to spread to other areas of the body, making early detection extremely important. Additionally, these models have also been used for MR image classification of breast tissue for cancer detection. Chien-Shun Lo and Chuin-Mu Wang[46] were able to detect the condition in their tests with an accuracy of over 90%. Another study collected data from over 480 patients with 23 features each with the purpose of training machine learning models, including

SVMs, to identify signs of coronary artery diseases, reaching results of testing accuracy between 72.92% and 79.17%. These results were achieved in part by using component analysis, reducing the dimensionality of the dataset to maximize the model's performance[7].

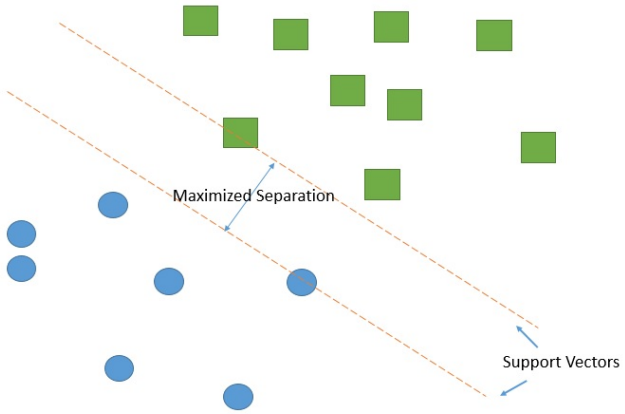


Figure 2.3: The Support Vector Machine model and the separation of classes inside the hyperspace. "Support vector machine", 20 November 2016.

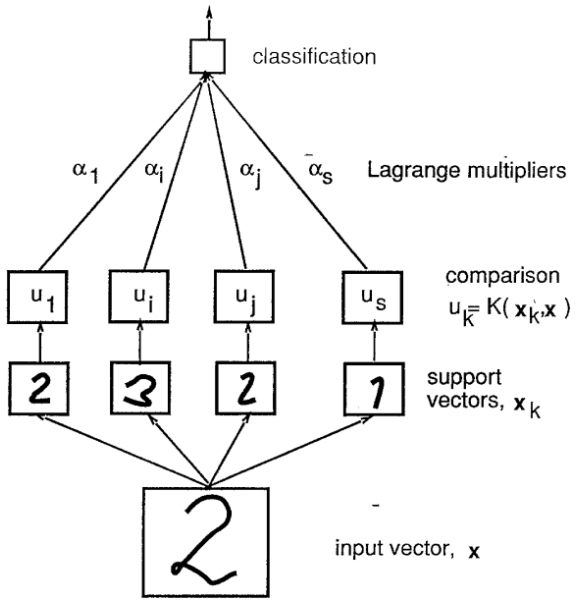


Figure 2.4: Classification of hand-written digits using support vector machines. From Cortes at al.[15]

2.2.1.2 Ensemble learning

Ensemble learning is a technique in machine learning that consists of combining multiple models that, by themselves, would be considered weak, into a better one. This technique is used mainly in the form of 3 processes. One of these processes available by ensemble learning is bagging, which is the process of creating multiple datasets and multiple learners so that a can combination of them creates a theoretically better learner. This specific technique can improve the performance of the model and reduce overfitting[13]. Boosting is yet another technique made available my ensemble learning and consists of iteratively training weak learners while giving more importance to learners that can correct the wrong classifications of the previous ones[59]. Both these processes are displayed in a summarized manner on figure 2.5. Although this can improve the overall accuracy, it also tends to overfit training data. These are just some of the techniques encompassed by ensemble learning but they can create some very useful machine learning models like the random forest, AdaBoost or gradient boost. Ensemble learning has been used to classify similar problems to the detection of bone marrow edema. A Random Forest feature subset selection wrapper and an ensemble meta-algorithm were used in a 2019 article to discriminate between low-grade and high-grade cartilaginous bone tumours based on parameters extracted from MRI scans.Gitto et al.[25] have noted that "All tumours were manually segmented on T1-weighted and T2-weighted images by drawing bidimensional regions of interest, which were used for first-order and texture feature extraction". Results yielded 85.7% and 75% accuracy on training and testing, respectively, which is promising taking into account that the musculoskeletal radiologist correctly graded 13 out of 16 (81.3%) of cartilaginous tumours from the test cohort.

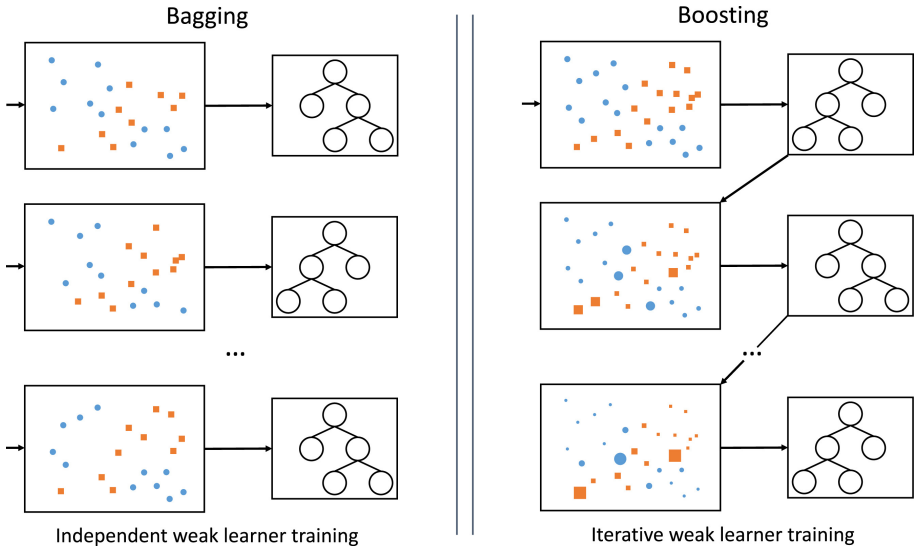


Figure 2.5: Visual representation of bagging and boosting. From González et al.[26]

2.2.1.3 Random Forest

The Random Forest model is one of the most used machine learning models in the context of medical imaging. This model type has changed the medial imaging area, especially in the context of image segmentation[37]. This system was first theorized by Ho et al.[31] in an article published in 1995 named "*Random decision forests*".

It consists of an array of decision trees with randomized values, which are fed the input data, generate their own output and then combine it in such a way that a conclusion is achieved. The easiest way to do this is by averaging the output of all the trees, in a regression problem, or taking the class that the majority of trees chose, in the case of classification[55]. Using a large number of trees helps to reduce variation when compared to regular decision trees[20]. An example of this type of model performing image classification can be seen on figure 2.6. One study was able to delimit the femur in MRI scans using this model of machine learning. By supplying this model with 839 images, they were able to achieve a point-to-curve error of less than 0.9 mm for 99% of said images. These 839 images of variate quality were manually annotated for the femur contour and then reevaluated 10 months later to assess the amount of intra-observer variability[44].

In an article by Yang et al.[61], a series of machine learning models were trained to detect bone fractures. To study this, 40561 images were feature extracted by a convolutional neural network(CNN) by exporting the 3 highest-level feature layers, fc6, fc7 and fc8, into three different types of classifiers, extreme learning machines (ELM), support vector machines (SVM) and random forests(RF). They were able to achieve fc6, fc7 and fc8 accuracy of 74.17%, 74.37% and 73.90%, respectively using the random forest model, which was the second-best result, losing only to the support vector machine model with 78.05%, 77.74%, 77.55% accuracy in the same metrics.

In summary:

Advantages:

- Has the ability to withstand irrelevant features[58]
- Is capable of handling both classification and regression problems
- Shows good accuracy even with small amounts of data
- Can output the relevance of each feature

Disadvantages:

- Computational complexity is high.
- Can overfit datasets with noise, especially in regression.
- Can struggle if class imbalance is present in the dataset.

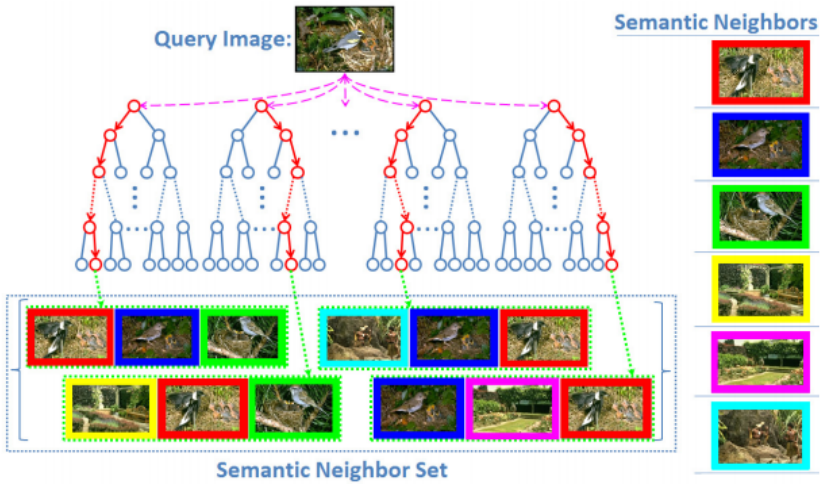


Figure 2.6: Example of random forest model performing multi-class classification. Adapted from Hao Fu, Qian Zhang, and Guoping Qiu in "Random Forest for Image Annotation"[23].

2.2.2 Deep learning

In machine learning, there is a multitude of different algorithms and approaches to solving a specific problem. Conventional machine learning methods usually have difficulties with performance if the data is not carefully prepared into an efficient format, which, in some cases, requires thorough domain knowledge and a great deal of time[42]. Throughout the years there have been substantial developments in machine learning, more specifically in artificial neural networks, which have generated a lot of interest from many different industries, one of them being medical imaging. These advancements are especially important in problems such as fraud detection where the dataset is hard to define in an efficient feature set[30][8]. These types of algorithms utilize multiple layers, each one transforming the input in some way, increasing the level of abstraction and allowing the modelling of complex functions. This area of machine learning is called deep learning and presents models which utilize techniques from representation learning which allow the models to discover efficient representations of the dataset for feature selection using prior knowledge in an automatic manner, replacing a big part of the manual feature engineering process[10]. Another reason these models provide such good results is the increasing amount of size and dimensionality of data in recent history. Due to this factor, a lot of the information contained in the data is lost without the automatic feature selection process and a complex enough system. An additional important factor in the deep learning paradigm is the capability of these networks to analyse data at a varying degree of abstraction. Abstraction can be loosely compared to how a human being would look for more specific features when trying to visually identify an object in its day to day life. For example, when trying to identify the species of an animal, a person would intuitively try to identify simple shapes and patterns such as circles or high contrast areas in a dalmatian dog. It has been more formally defined as a space where one can map the ground truth of a problem to an easier abstract representation of it, make deductions and then generate a ground solution from such[35]. Abstraction has been previously theorized to be linked with the increase in knowledge

of a specific model as the layers are progressed, with a few exceptions, mainly in Pooling and ReLu layers. This abstraction remains low in the initial layers of deep learning methods but generally increases according to depth[40]. The automatic feature selection process of deep learning algorithms is also considered a form of abstraction[35].

Some of the models used in this subject include:

- Convolutional neural networks
- Recurrent neural networks
- Autoencoders
- Deep belief networks[51]
- Deep Boltzmann machines[57]

2.2.2.1 Convolutional Neural Networks

In recent years, convolutional neural networks have been thoroughly studied, leveraging from the many benefits added to deep learning in the process such as improvements in the processing speed of computers and increases in data size and dimensionality[28]. Originally inspired by David H. Hubel and Torsten Wiesel whose work described relationships between the animal visual cortex and specific neurons[34], Yann LeCun among others proposed and later improved the CNN architecture at the end of the 1980s and the start of the 1990s. These networks are comprised of multiple layers of different types; the 3 main layer types being:

- Convolutional layers
- Pooling layers
- Fully-connected layers

Convolutional layers are the first and main type of layer used in convolutional neural networks. These layers generate features by performing a convolution operation between the image pixels and different filters according to a kernel of arbitrary shape, producing a new representation of the image called a feature map, as seen on figure 2.7. The purpose of this stage is, in a way, to find the most appropriate filter for the images so that an efficient feature set can be extracted.

The pooling layer's job is to downsample the feature map by generating a numeric value for each independent matrix of pixels of specific size, thus increasing abstraction and discarding information, as seen on figures 2.8 and 2.9. This process is usually done by selecting either the maximum value of the matrix or the average value of the matrix, the former being the most popular strategy[27]. After layers of these 2 types process the feature set enough, the model connects the feature map with multiple dense fully-connected layers of neurons, similar to a traditional artificial neutral network, into the output layer with the shape of the chosen output.CNN's are cable of creating very efficient representations of the data, while also preserving the spatial relationships

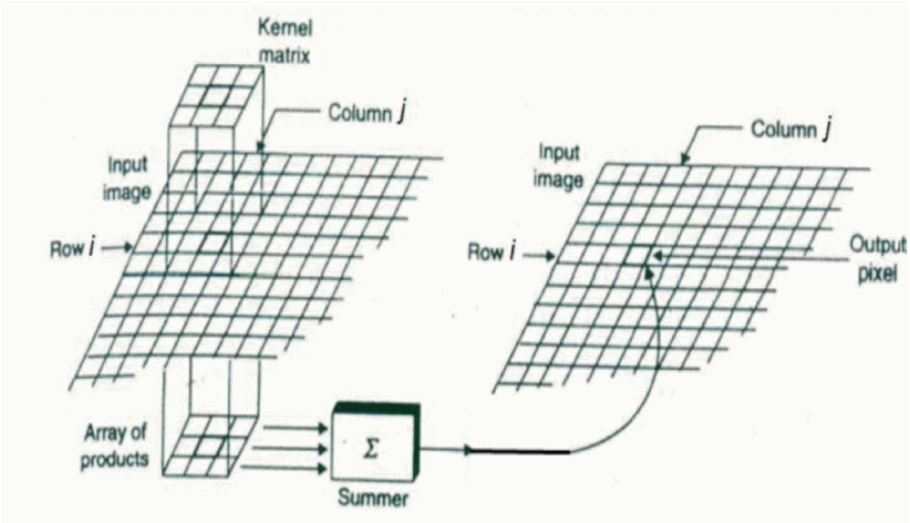


Figure 2.7: The convolution process applied on convolutional layers of CNNs. Adapted from Albawi et al.[3]

found in it[48]. The issue with training the neural networks is that they heavily rely on the amount of data that is fed to them, which may be an issue in the medical context, especially when researching new or rare illnesses[43]. Furthermore, these machine learning models require heavy amounts of computing power, as well as domain knowledge to use them effectively[24]. In summary:

Advantages:

- The ability to extract features automatically from image/video, these features are often highly correlated with the problem
- The ability to outperform conventional machine learning methods on various types of image-related problems
- Capacity to preserve the spatial relationships of the images fed through the model (can identify objects anywhere in the image)

Disadvantages:

- They require a large amount of computing power but often take advantage of the graphical processing unit's capabilities to perform mass calculations
- They usually require a large enough dataset to outperform other machine learning methods
- They require the domain knowledge to use them effectively

Image scans used by health professionals to visually identify anomalies in the body of a patient, like magnetic resonance or x-ray imaging, end up being a big data storage for deep learning algorithms to learn useful patterns, like breast cancer diagnosis[24] among many others. An article published no April of 2019 managed to develop a CNN model capable of achieving 79.5% accuracy in classifying whether or not an MRI scan contained a bone marrow edema lesion in the

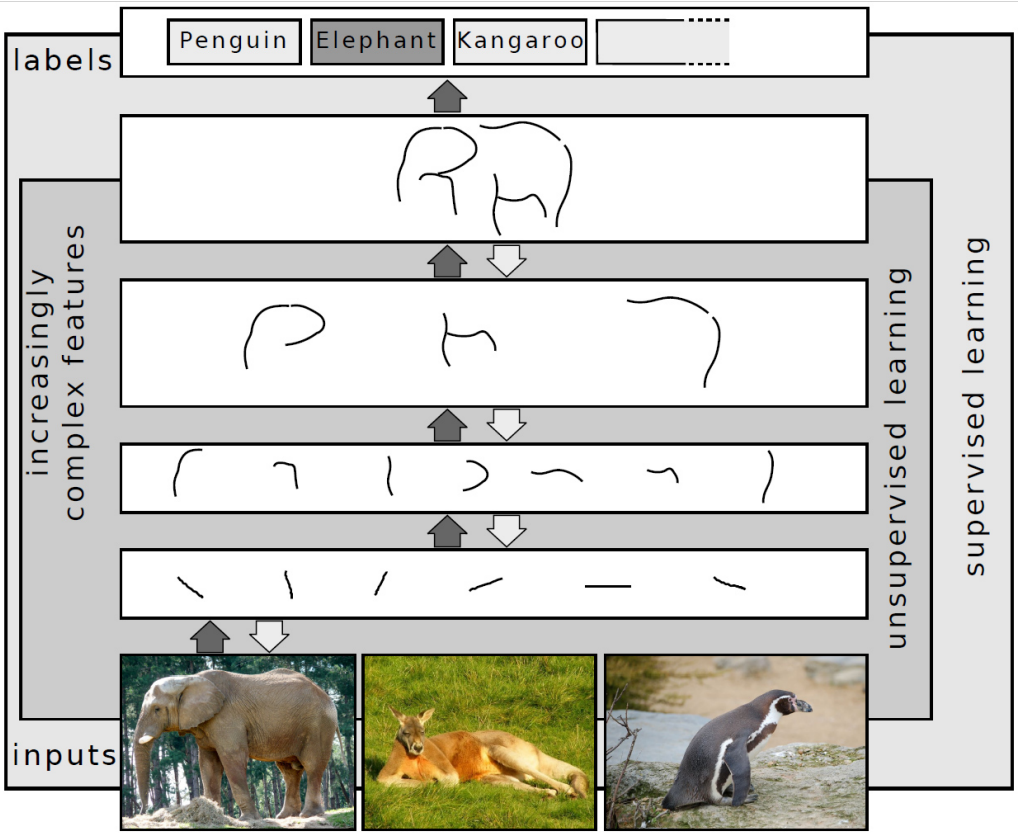


Figure 2.8: The abstraction process of a deep learning model. Image from Sven Behnke, Representing Images on Multiple Levels of Abstraction in Deep Learning, November 2012.

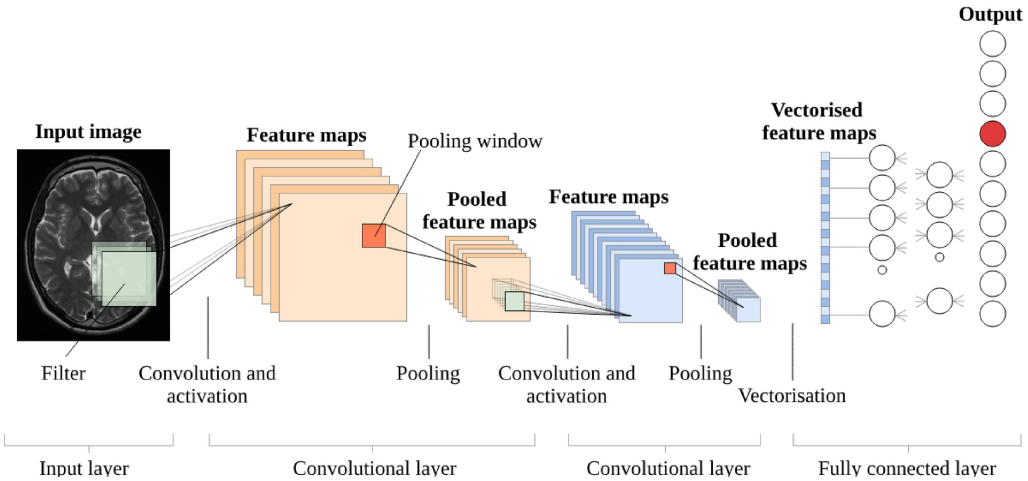


Figure 2.9: Visual representation of the inner workings of convolutional neural networks. From Lundervold et al.[48]

knee. To get this result, it was fed 1435 MRI scans from the knee region[5]. In another study, CNNs were used for outlining bone marrow edema lesion on knee MRI scans. The Dice similarity coefficients (DSC) and Pearson's R2 were reported as being 0.70 and 0.94, respectively, on a dataset of 600 subjects and 1782 scans and were 90% faster than a human doing the same task[56].

2.2.3 Metrics and the confusion matrix

To more accurately understand how a model behaves, one needs to understand metrics and their importance. A multitude of different metrics are used to evaluate these algorithms, some of them being accuracy, precision, recall, specificity, balanced accuracy, F1 score and the Matthews correlation coefficient (MCC) for classification problems or mean squared error and mean absolute error for regression problems, among many others[22]. Furthermore, the confusion matrix allows the analysis of the output of classification problems in a more intuitive way and is compromised of the following 4 numeric values: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN)[49], as seen on figure 2.10. A lot of the aforementioned metrics are based on the values obtained by this matrix, as can be seen in the following equations:

Accuracy:

$$Accuracy = \frac{(TN+TP)}{(TN+TP+FP+FN)}$$

Precision:

$$Precision = \frac{TP}{(TP+FP)}$$

Recall:

$$Recall = \frac{TP}{(TP+FN)}$$

Specificity:

$$Specificity = \frac{TN}{(TN+FP)}$$

Balanced accuracy:

$$Balanced\ accuracy = \frac{(Recall+Specificity)}{2}$$

F1 score:

$$F1\ score = 2 * \frac{(Recall \times Specificity)}{(Recall+Specificity)}$$

MCC:

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(FP+TP) \times (FN+TP) \times (FP+TN) \times (FN+TN)}}$$

Mean squared error:

$$Mean\ squared\ error = \frac{\sum_{i=1}^{n}(y_t - y_p)^2}{n}$$

Mean absolute error:

$$Mean\ absolute\ error = \frac{\sum_{i=1}^{n} \mid y_t - y_p \mid}{n}$$

, where n is the number of data points, y_t is the observed output and y_t is the predicted output

		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 2.10: Schema denoting the multiple components of the confusion matrix. From Luque et al.[49]

Furthermore, the receiver operating characteristic (ROC) curve allows the visualization of a binary classifier's diagnostic performance throughout all its classification threshold levels. This curve is used to intuitively identify the ideal value for the trade-off between specificity and sensitivity. The area under the curve (AUC) of the ROC can be used to measure how well a machine learning model can distinguish between two classes[19]. To be able to analyze the relationship between two different variables, Spearman's rank correlation coefficient was used. This measurement is used to evaluate how well these two variables can be fit into a monotonic function and can evaluate relationships that need not be linear. It is more formally defined as a measurement of rank correlation[52][6]. It is expressed by the following mathematical formula:

Spearman's rank correlation coefficient:

$$Spearman's\ rank\ correlation\ coefficient = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

, where d_i is the difference of ranks and n is the amount of observations

Chapter 3

Methodology and Experimental Setup

Over the course of a few months, our work has been primarily divided into 2 problems, detecting and annotating BME in MRI scans as well as annotating the expected region to be affected in the X-Ray based on the MRI findings. Detection in MRI images is overall simpler than in X-ray since MRI is a very complete and complex method with several types of sequences that captures images at different levels of the studied region, while an x-ray compacts all that information into a single image. This makes it so that the BME region is overlapped with bone and skin, making human detection extremely challenging. It also doesn't contain the depth dimension found in MRI scans, making less data available. Alongside this, edema doesn't seem to be detected by the nature of the X-ray images.

3.1 Data Preparation

In the initial phases of the implementation, most of the work focused on gathering the necessary data and treating it for future use in the development of the algorithms. Images of knee X-ray, see figure 3.3 as well as T1-weighted and fluid-sensitive sequence images of MRI scans were collected from 72 different patients under the age of 18, provided by the University Hospital Center of São João. Of these 72 patients, 36 had BME detected by health professionals and were manually annotated with regions of interest on the respective slices with zones identified with BME, as can be seen on figure 3.1. The latter 36 non-edema patients were annotated in the bone region so that the algorithms could learn healthy bone composition as well, as seen on figure 3.2. A patch-removal strategy was employed in order to have enough data so that the random forest classifier and especially the CNN classifier can have enough images to work with and reduce the impact of outliers, among other factors. Each patient had a folder and a label specifying if the patient has edema in their scans or not. Having that, the patient folder contained multiple items, including an executable to view the scans, an XML file with details about the exam and the scan file distribution and the binary data of these scans in the Digital Imaging and Communications in Medicine (DICOM) format.

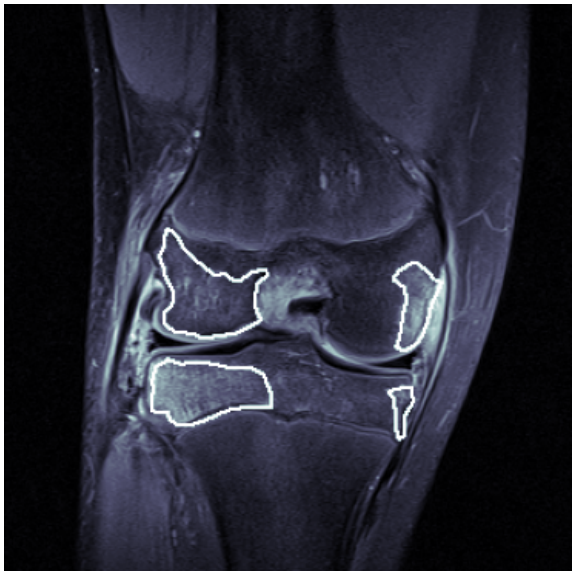


Figure 3.1: A single slice of an MRI scan with multiple annotated edema findings.



Figure 3.2: Image of an MRI scan with two non-edema regions annotated.

The reading of these images is made using a python script with the help of the **pydicom** package, which provides an interface for us to read the encoded data from the DICOM format. This format first proposed in 1985 works by distributing information object definitions (IODs) to the multiple objects collected by the scan thus linking them together. The XML file called 'CONTENT.XML' contains the patient details, the exam details and the description of the multiple images resultant from the scans. This information allows us to isolate types of scans (MRI, X-ray) and types of files (annotation file, scan file). As can be seen on listing 3.1, this file can be used to discover the location of various scans. After discovering all relevant files, scans and annotations are subsequently opened using pydicom and matched according to their Service-Object Pair (SOP)



Figure 3.3: Example of an X-ray image.

identifiers (ID).The topology of the DICOM format, according to SOP IDs can be seen on figure 3.4. The pydicom API allows access to the DICOM attributes of scan and annotation files, which are multiple variables containing information regarding the files. A summarized version of the attributes used can be seen on table 3.1 The annotation images can then be extracted from the files and used as image masks to isolate the edema and non edema regions of interest in the scan files using the data available from reading the scan and annotation images using the pydicom API, as can be seen summarized in figure 3.5. The corresponding image for the file opened is obtained through either the *pixel_array* field for the case of scans or the *overlay_array* method for the case of annotations generated from opening the DICOM file using pydicom. With both images as well as the annotation origin, a mask is applied such that only the pixels inside of the ROI are visible, in order to isolate the edema or non-edema bone regions apart from the rest of the image. Subsequently, small patches of arbitrary size are extracted from the masked image and added to the pool of images to feed to the models later on. This process can be seen in a more intuitive way in figure 3.6.

```
<study id="AA0FB43E">
    <study_data>
      <description />
      <accession_id>F245A1B5ZU</accession_id>
      <exam_id>F242B1B5ZU</exam_id>
      <body_part>LEFT KNEE</body_part>
      <modality>MR</modality>
      <instance_uid>5.0.632.46.1.8267238467.485472</instance_uid>
    </study_data>
    <series id="00001156">
        <series_data>
            <modality>MR</modality>
            <number>17</number>
            <description>knee_mri</description>
        </series_data>
        <image id="FB0C54A9">
            <datetime>21/07/2019 14:26:42</datetime>
            <utc_date>07/21/2019</utc_date>
            <utc_time>14:26:42</utc_time>
            <dicom_file>
                <src>../DICOM\0000F5A9\AAABD847\AA0FB43E\00003FDC\EEDEC017</src>
                <index>0</index>
            </dicom_file>
        </image>
    </series>
</study>
```

Listing 3.1: An example of a section of a CONTENT.XML file

Multiple parameters were considered for the data preparation process in order to assure that every model has the best environment possible to learn. Some of the parameters considered were patch size, which was, whether patches would have overlapping between each other or not, the very hyper-parameters of the models. The patch size considered was of size (10,10) and (18,18), due to being the most adequate for the average size of the edema. A normalization process was also applied, where a perfect circle annotated in the muscle region would be identified alongside the edema or non-edema region, averaged and subtracted from every pixel in the region of interest, which was tested and showed results supporting the method. This annotated circle can be seen on figure 3.7. Since the annotations are closed curves, there are patches in the cropped edema region that overlap with non-edema regions; thus it was decided to use a cutoff of 10% of non-edema region for considering whether a certain patch is valid or not. An image of a patch that goes out of bounds of the edema can be seen on figure 3.8

After all this is done, there are multiple patients with scans that have multiple images, broken down into patches of "edema" and "non edema" regions which can now be fed into our machine learning models.

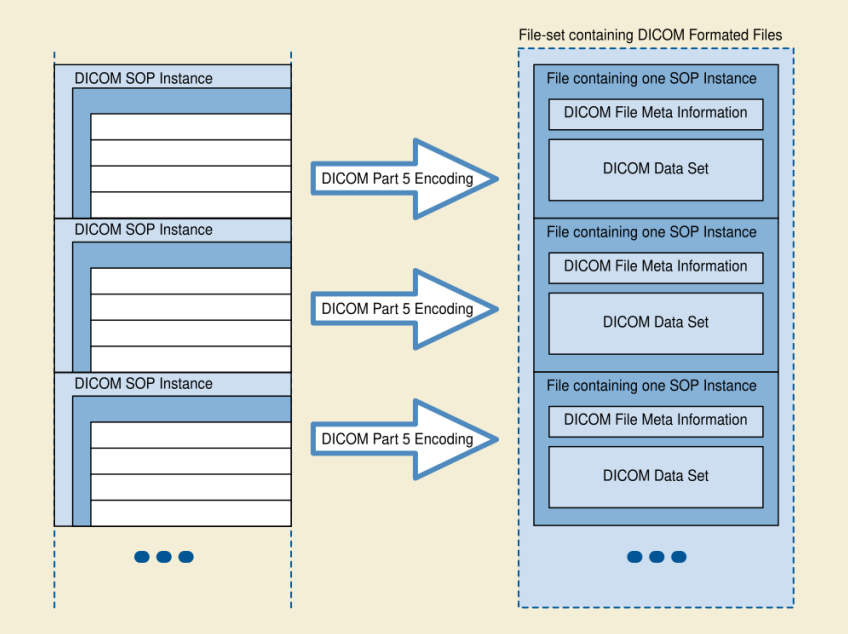


Figure 3.4: The topology of the DICOM format. Shows the usage of SOP instances.

Attribute name	DICOM tag	Description
Referenced Series Sequence	(0008,1115)	A data structure containing the referenced SOP ID found on annotation files. This ID should match the SOP ID found on the corresponding scanned image.
SOP Instance UID	(0008,0018)	The SOP ID of the scan files. This ID should match the one found on the annotation file.
Overlay Data	(60xx,3000)	The array containing the value of each pixel of the annotation ROI, i.e. the annotation image. This array, along with the origin, defines the region of interest.
Overlay Origin	(60xx,0050)	The origin of the annotation file. This attribute is used to place the annotation image correctly on the scanned image.

Table 3.1: The attributes of the DICOM format used to read the scan and annotation files, along with their tag and a brief description.

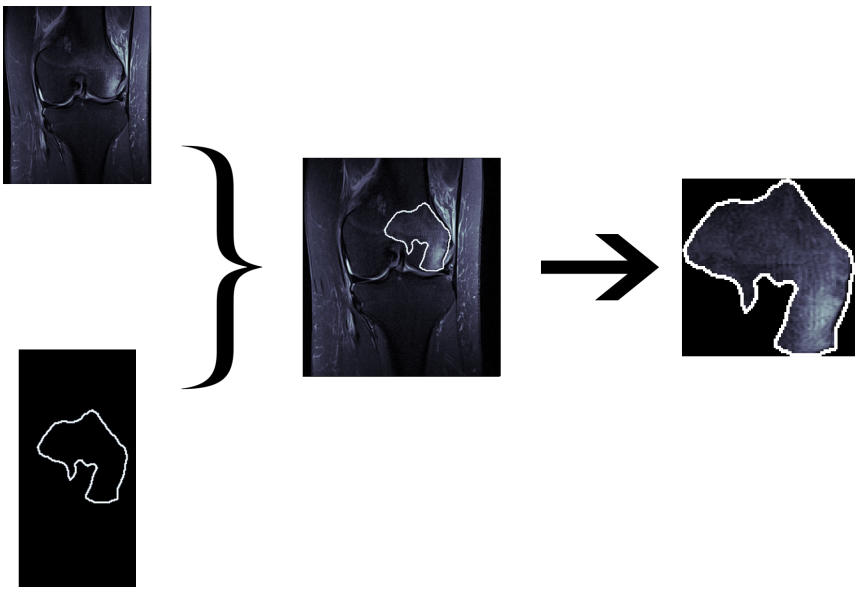


Figure 3.5: The process of reading DICOM image files. This procedure consists of reading both the image as well as the annotation file, followed by the merge of these images and the subsequent crop of the region of interest area.

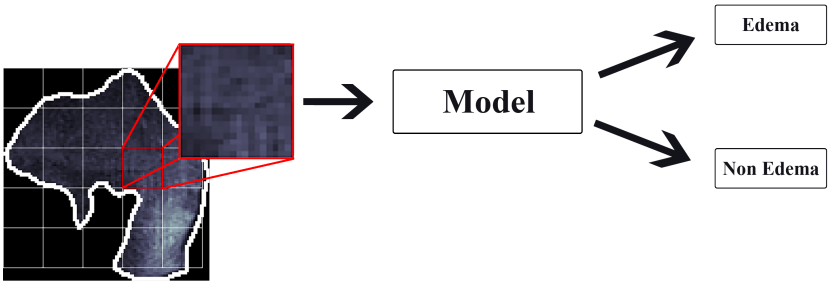


Figure 3.6: Schema representing the dataflow of patches to the machine learning models.

3.1.1 Class imbalance

Due to the fact that edema regions are generally small, even though the number of patients is balanced, the resulting number of patches can be seen in figure 3.9 as being very unbalanced, resulting in additional difficulties for the algorithm and generally worse performance. To combat this, both undersampling and oversampling of the patches was performed, the latter being the strategy with most efficiency. This process, however, is not perfect and can also affect the model negatively. This further complicates the already short amount of data.

3.1.2 Age in BME detection

Age is a relevant factor when considering the problem of BME. In children, a crucial component of their skeleton called the epiphyseal plate might interfere with BME edection. This component is very hyperintense in fluid-sensitive sequences - T2 or proton-density (PD) weighted images with fat-suppression, or short tau inversion-recovery (STIR) - and appears in regions that may

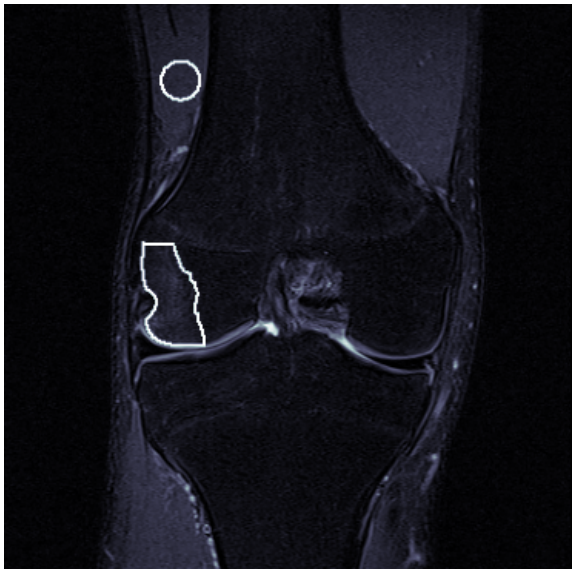


Figure 3.7: Image of an MRI scan with annotated circles for normalization and an edema region.

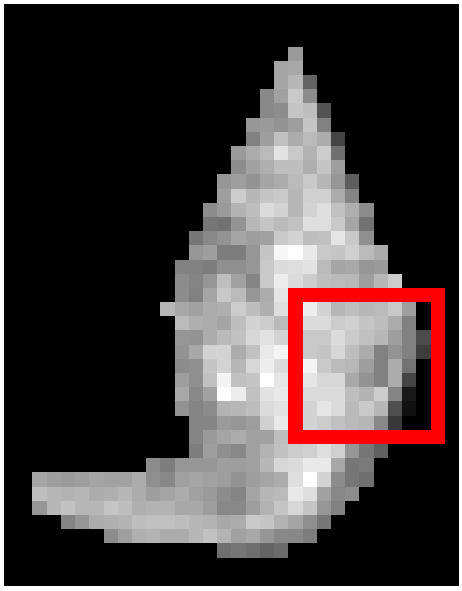


Figure 3.8: Image of a patch erogenously capturing pixels outside of the edema region

overlap with edema regions. Considering this along with the fact that BME is also detected by very hypersensitive regions in the previously mentioned images, confusion may arise in the BME detection process. Moreover, this component shrinks in size and loses hyperintensity as the child grows, increasing the difficulty of the detection from a machine learning perspective[54]. Due to the affect of age in the detection of BME, the ages of the patients used in the dataset were collected. A box plot of the ages of patients with or without edema findings can be seen on figure 3.10. From this statistic it can be seen that most of the patients are between 11 and 17 years old, which is a range that can still be affected by the aforementioned problems; hence there may be some correlation between patient age and model accuracy.

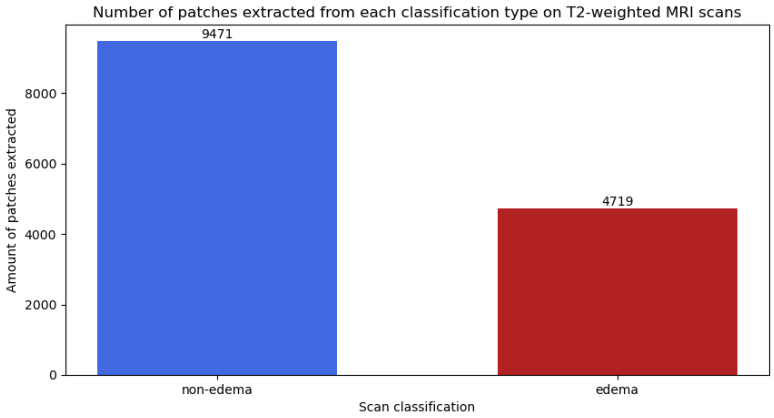


Figure 3.9: The distribution of patches generated between edema and non-edema patients.

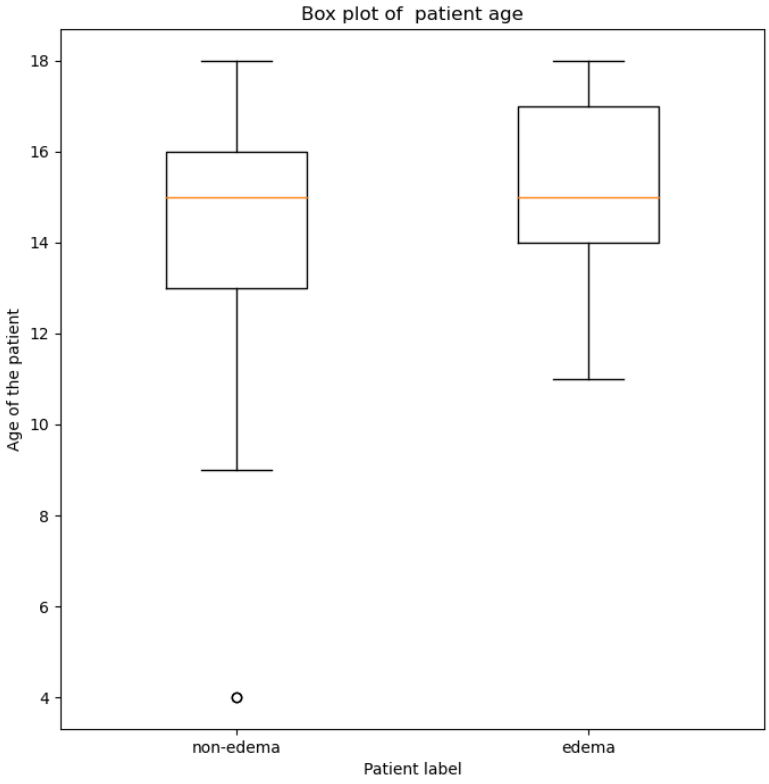


Figure 3.10: The box plot of patients with or without edema findings.

3.2 Algorithm Validation

While implementing the algorithms, we will also focus on how to correctly evaluate them. Knowing how to correctly measure the developed models' performance is crucial in many different

stages. It is important to monitor the performance of a proposed solution so that it can be adjusted in a way that improves the results. Unfortunately, there is not one perfect approach, and there are often benefits to evaluating a solution using different metrics depending on the goal of the project. In classification, we will mainly use the confusion matrix and the values generated from it, the accuracy, the balanced accuracy, the precision, the recall, the F1 score and the MCC score.

3.3 Models

Due to the research made on the State of the Art chapter, the proposed solution will focus on 2 types of machine learning algorithms, random forests and convolutional neural networks due to their accuracy, convenience and versatility. Random forests work well with limited data and irrelevant features and have shown great results on many different computer vision tasks and scenarios but require a good deal of data preparation to extract good features. CNN's have shown very good results in modern problems although they require a large amount of data to achieve higher accuracy. It is not easy to make a prediction on how well these techniques will work on the gathered data as it depends on a number of factors such as the amount of noise in the dataset, and the size of the dataset, how well the data space is explored, etc. These models receive a series of inputs(images) and outputs("edema" or "non-edema" classifications) and try to infer rules to generalize the problem. In this case, they were trained and tested in order to be able to identify if the patch of a scan contains edema or not.

Regarding the CNN model, it was decided that the Keras API would be the interface of choice, due to its flexibility and consistency, as well as its ease of use. The model topology seen on table 3.2 and figure 3.11 was used with focus on a balance of speed and accuracy. These types of models discover their own features from the training data through the means of its weight-sharing architecture and filter application, which makes the training process much more convenient when compared with traditional approaches.

Layer (type)	**Output Shape**	**Param #**
conv2d (Conv2D)	(None, 10, 10, 32)	896
max_pooling2d (MaxPooling2D)	(None, 5, 5, 32)	0
conv2d_1 (Conv2D)	(None, 5, 5, 64)	18496
max_pooling2d_1 (MaxPooling2D)	(None, 2, 2, 64)	0
conv2d_2 (Conv2D)	(None, 2, 2, 128)	73856
max_pooling2d_2 (MaxPooling2D)	(None, 1, 1, 128)	0
flatten (Flatten)	(None, 128)	0
dropout (Dropout)	(None, 128)	0
dense (Dense)	(None, 512)	66048
dense_1 (Dense)	(None, 128)	65664
dense_2 (Dense)	(None, 2)	258

Table 3.2: Text-based representation of the chosen CNN topology listing layers, the output shape of the layers and the number of parameters each layer has.

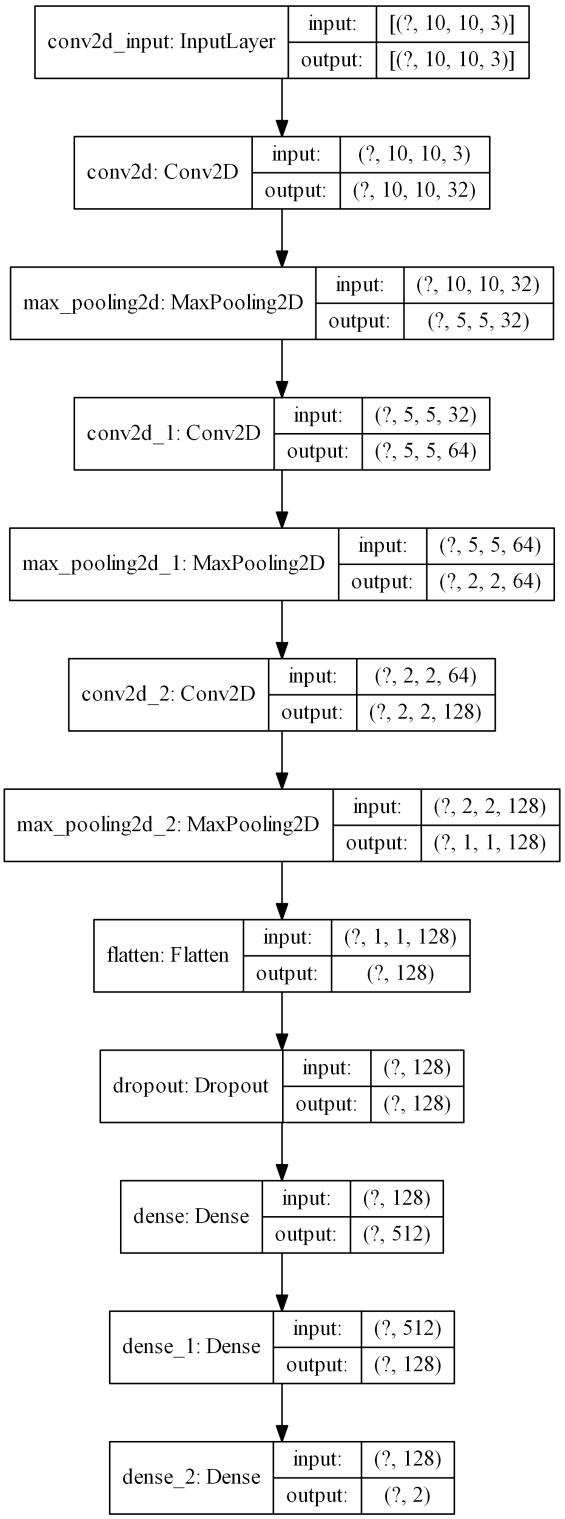


Figure 3.11: Visual representation of the chosen CNN topology

Regarding the random forest model, it was decided that the sklearn API was to be used .This type of model works by having a network of decision trees, where each one of these trees votes on the class it thinks the input data is from and picks the majority, in the case of classification. When

focusing on regression this network uses statistics like mean or average of the output of each tree to infer a decision. A k-fold cross validation process was applied to the ensure that an optimal setup of hyper parameters would be used. This procedure indicated that a random forest with a minimum amount of 100 trees, the number of features to ponder when looking for the best split to be the square root of the total number of features, no upper bound on tree length, the minimum number of samples to split a node of 2 and bootstrapping was optimal. These parameters, the interval of values pondered for each of them and the optimal value found can be seen on table 3.3. This process involves the partitioning of the data into k groups of training and testing, as can be seen on figure 3.12, such that every data point serves as either test or training data and that each point is part of the testing set in exactly one of the k iterations. After this, the algorithm runs for k iterations, where it trains and tests a model at each iteration with its specific train and test set, extracts metrics and saves the model with the best chosen metric such as accuracy. In order to extract relevant features from each patch image, the pyradiomics[1] package was used, which is an open-source project aimed to extract radiologic features from medical images.

Parameter name	**Interval**	**Chosen value**
number of estimators	$\{190k+100 \mid k \in [0,10]\}$	100
max_features	$\{'sqrt'\}$	'sqrt'
max_depth	$\{9.(63)k+4 \mid k \in [0,11]\} \cup None$	None
min_samples_split	$\{2,5,10\}$	2
bootstrap	$\{'True','False'\}$	True

Table 3.3: The random forest parameters and the multiple values considered. The k-folds cross validation process iterates over the parameter values inside the chosen intervals and locates the best combination according to a specific metric such as accuracy.

Each one of these types of models was trained and tested 20 times using random partitioning for train, test and, in the case of the CNN, validation sets, and then the results averaged out, due to the volatility of the low amount of scans gathered. This experiment utilized 54 of patients gathered for training and the remaining 18 for testing. In the case of having a validation set, 5 of the 54 patients utilized for training are removed from this set and used for validation effects.

The main procedure followed by health professionals in the detection of this condition involves looking at MRI scans of a patient in the affected area by the use of visual analysis.

On these types of images, both CNNs and random forest classifiers provide a way to detect bone marrow edema in our tests with some level of accuracy. In this case, the CNN is the more modern approach, having the ability to generate its own feature set by using image filters and convolutional operations[17], while the random forest is recognized as the more traditional machine learning workflow, using a multitude of decision trees to vote for specific classes (in this case, either "edema" or "not edema").

The detection in X-ray images suffers from the fact that it consists of the overlapping of multiple different components of the human body, namely bone, skin and cartilage, into a single image,

[1] https://pyradiomics.readthedocs.io/en/latest/

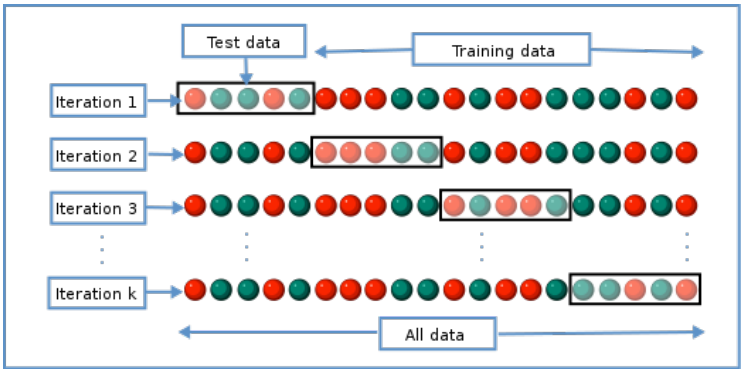


Figure 3.12: The k-fold cross validation method works by partitioning the train and test data such that every data point serves as either test or training data and that each point serves as test in exactly one iteration. Gufosowa, 18 September 2019.

providing limited analysis. This property makes it so that the edema region is probably obfuscated to a high degree and thus, health professional are unable to detect it in these images. Furthermore, this means that less data is inherently available to be used since the depth dimension of the MRI scan is not available in this case. Alongside this, edema simply doesn't seem to be detected by X-ray. It is intended to find out if this process can be more successful than the near impossible human detection.

Chapter 4

Results and Discussion

During the experiments done, it was found that oversampling was the most appropriate method to try to rectify the class imbalance found after patch extraction, generating about 1% balanced accuracy gain over undersampling and about 6% over no class balance. T1-weighted and fluid-sensitive sequence MRI images we considered and tested, the latter one generating about 4% more accuracy over the former. It is theorized that this happens due to the increased amount of textures found on T1-weighted images, which acts mostly as noise to the machine learning algorithms. A patch size of either (10,10) or (18,18) was used, with the former being the chosen value due to achieving better results. Moreover, using a much bigger size such as (18,18) showed that a large number of edema regions would not get any patches extracted since, they were too small compared to the patch size, whereas using a smaller size would reduce the predictive quality of patches. Normalization showed results that indicate that this process has very low relevance to the problem. This is suspected to be due to the fact that the data preparation process of the CNN model and the pyradiomics module applies filters that are very similar to the normalization process or the generated features are not very affected by it.

The best result achieved in our tests was using a CNN of patches of fluid-sensitive sequence MRI images with size of (10,10) with normalization applied, achieving 72.1% of balanced accuracy, as seen on table 4.1. The graph of the area under the curve and the confusion matrix of this test can be seen on figure 4.1 and 4.2, respectively. These two images were extracted from the results of the 20 execution concatenated into a single set. These results seems congruent with previous experiments[56] and helps better establish the baseline performance on the understudied problem of edema detection in radiology. As previously thought, X-ray scan images are not a sufficiently strong predictor for the presence of BME, mainly due to its poor visibility. The following results were obtained for X-ray as well as fluid-sensitive sequence MRI images:

The best classifier produced in the tests done was subsequently utilized to classify patches in separate images, producing the results found in figures 4.3, A.1 and A.2, for case of edema and figures 4.4, A.3, A.4 for the case of non edema. The edema patch appear to have a high percentage of matches, but the non-edema patches show signs of confusion.

Image type	Model	Normalize	Balanced accuracy	Precision	Recall	F1	MCC	AUC
MRI	CNN	yes	72.10%	72.30%	71.66%	0.7198	0.4421	0.79
MRI	CNN	no	71.74%	71.83%	71.54%	0.7168	0.4348	0.78
MRI	RF	yes	69.11%	69.61%	67.84%	0.6871	0.3423	0.75
MRI	RF	no	68.39%	69.53%	65.49%	0.6745	0.3685	0.74
X-ray	CNN	no	56.32%	58.67%	42.79%	0.4949	0.1314	0.58
X-ray	RF	no	56.03%	59.13%	42.21%	0.4852	0.1374	0.58

Table 4.1: Results obtained in the multiple tests done

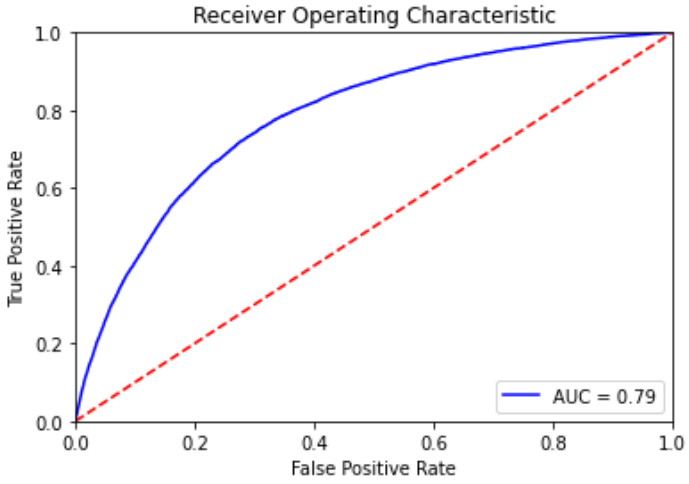


Figure 4.1: The AUC curve of an execution of the CNN model.

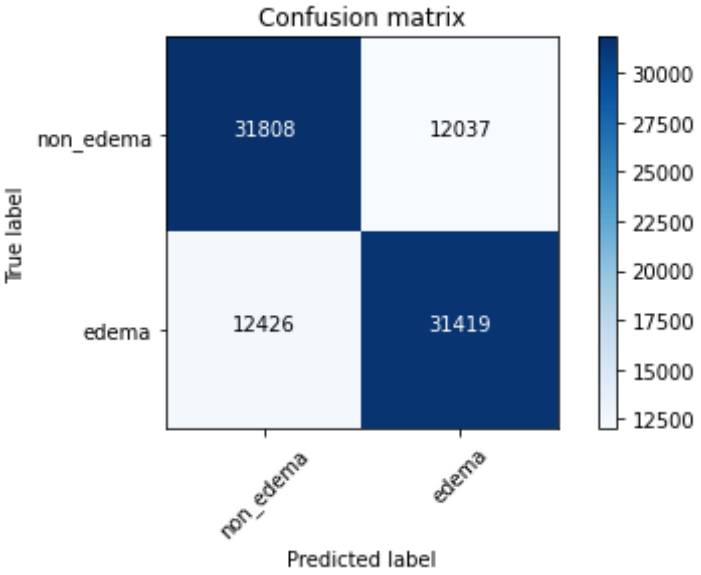


Figure 4.2: The confusion matrix of an execution of the CNN model.

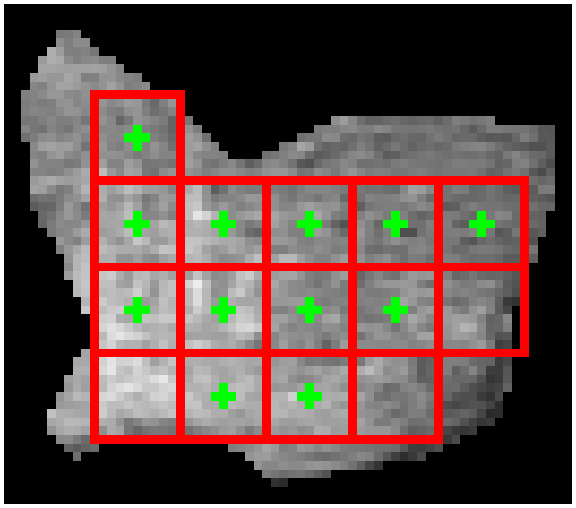


Figure 4.3: First example of an execution of the CNN model on an independent patch.

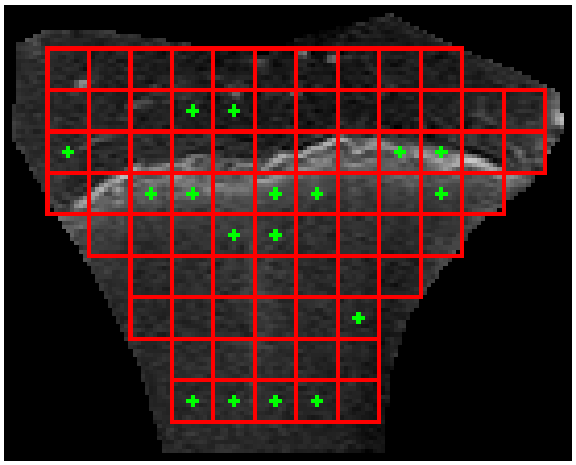


Figure 4.4: First example of an execution of the CNN model on a non-edema patch.

4.1 Age in BME detection

Age is a relevant factor when considering the problem of BME. To analyze the possible effect this had on the models used, statistics such as the mean and the standard deviation of the age of patients' patches were compared against the models' accuracy values using the Spearman's rank correlation coefficient. Using 20 executions of a CNN model, the rank correlation between either the mean or the standard deviation of the ages of patches used and either the training accuracy, the validation accuracy or the testing accuracy was calculated and resulting in the table 4.2. These values lead to believe that there is a moderate positive correlation between the model's accuracy and the ages of the patients, supporting what was previously said in chapter 3 about the epiphyseal plate. More specifically, the correlation value between the testing accuracy and the age mean of 0.48, along with its low p-value of 0.034 show that there is moderate positive correlation along with strong statistical significance in this calculation[2]. The scatter plot for these two variables can be seen in figure 4.5. This may pose a major issue with BME detection using machine learning

First variable	Second variable	Correlation	P-value
Training accuracy	Age mean	0.20	0.387
Validation accuracy	Age mean	0.45	0.046
Testing accuracy	Age mean	0.48	0.034
Training accuracy	Age standard deviation	-0.47	0.037
Validation accuracy	Age standard deviation	-0.09	0.691
Testing accuracy	Age standard deviation	-0.24	0.307

Table 4.2: Values generated by the calculation of the Spearman's rank correlation coefficient.

models and is suspected to be the cause of some of the false positives encountered in the tests performed.

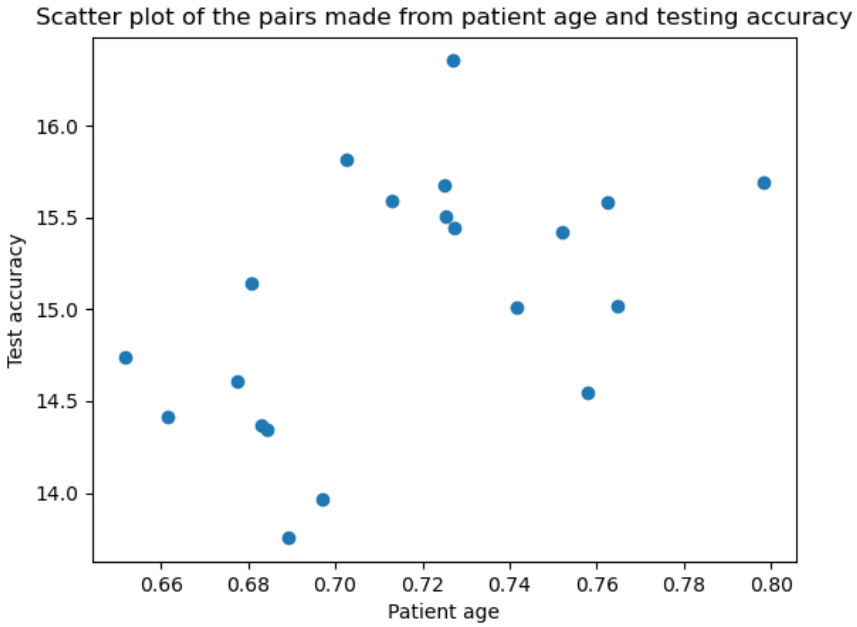


Figure 4.5: The scatter plot of testing accuracy and patient mean age.

4.2 Discussion

The previous results were an inherent consequence of the experimental nature of the research done. The experiments done show promising results but also show that the BME detection is a complicated problem, due to a multitude of reasons. Firstly, MRI images are highly variant making the amount of patients gathered was sub-par compared to what seems to be necessary to fully explore the data space in an efficient and meaningful way, confirmed by the high variability between different train and test sets. Additionally, these images are affected by the machine parameters used by health professionals in their scans. The normalization process tried to mitigate this effect but showed low relevance to the accuracy score of the models. The age of the patients is

also suspected to diminish the accuracy of the results due to physiological changes in the bone, as seen previously; however, due to the low amount of patients, it is uncertain how much of a factor this may have been. Another issue with the BME detection problem is the imprecise definition of BME in health professionals, making the training data generated from human classification somewhat subjective and less reliable to model. Even with what was considered to be a small amount of data, the CNN architecture showed to be on par or better in terms of accuracy when compared with the random forest model receiving pyradiomics-generated features, reinforcing the usefulness of deep learning in the subject. showed to be These results seem to coincide with was expected from analysing similar experiments done and also seems to be congruent with the results produced from applying the algorithms to individual images after the training was done.

Chapter 5

Conclusions

Bone marrow is a radiologic finding that is usually associated with common conditions such as trauma or osteoarthritis, making it a field of study that could potentially impact the lives of millions of people. With the purpose of exploring the problem of detecting this finding in MRI and X-ray images, a CNN and a random forest model were trained and tested from 72 patients under the age of 18; 36 of those having findings and 36 having no findings of edema. The images from these patients were manually annotated in the corresponding regions of interest and broken down into small patches which were subsequently fed into the aforementioned models. The best accuracy achieved was 72.1%, resultant from using a CNN model with fluid-sensitive sequences - T2 or proton-density (PD) weighted images with fat-suppression, or short tau inversion-recovery (STIR). These results showed that some machine learning models are capable of detecting BME to certain degree; however, they also show that this problem requires additional research and that better results may be achieved.

5.1 Future work

These results show some level of correlation between the input images and the observed edema, i.e., the machine learning algorithms seem to be able to learn how to classify BME but would require more thorough study to achieve great results. Some ways considered to possibly improve on the experiment are as follows:

- One way to better the work done would be to train a separate machine learning model to segment the bone from the rest of the scanned images in order to be able to use the neighbouring patches of the edema regions of interest.

- With the addition of many more patients, the patch strategy could be changed to using the entire segmented bone section and training a CNN to receive these types of images, in hopes of it locating the edema regions on different locations of the bone, since CNN generally perform better with larger datasets.

- Reducing the standard deviation and increasing the mean of the patient age could also improve results, as the opposite has been suspected to reduce the model performance due to the epiphyseal plate.
- To counter the imprecise definition of BME, one could make more strict restrictions for the human classification data used in the machine learning models. This would be done to reduce the uncertainty that could arise with certain cases of this classification process.
- Using circles in the non-edema part of the bone instead of the muscle region for the normalization process may produce better results as this area seems to be more correlated with the problem and could help reduce some of the confusion of the algorithms.
- Experimenting with other machine learning algorithms could reveal a better setup for the experimente

Appendix A

Apendix

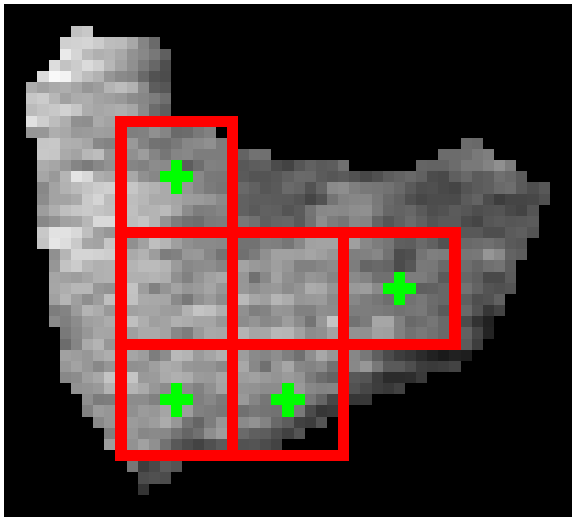


Figure A.1: Second example of the best classifier on an edema image.

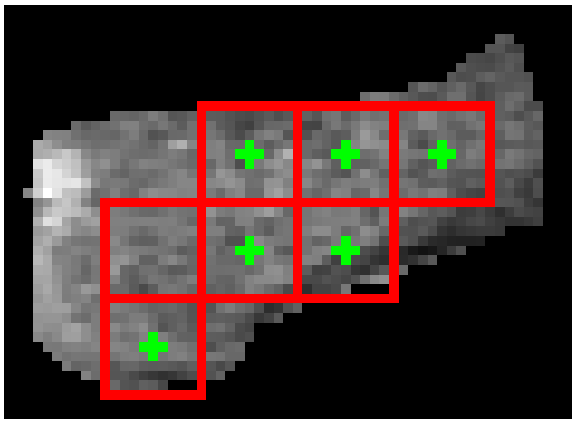


Figure A.2: Third example of an execution of the CNN model on an edema patch [Third example of an execution of the CNN model on an edema patch.]

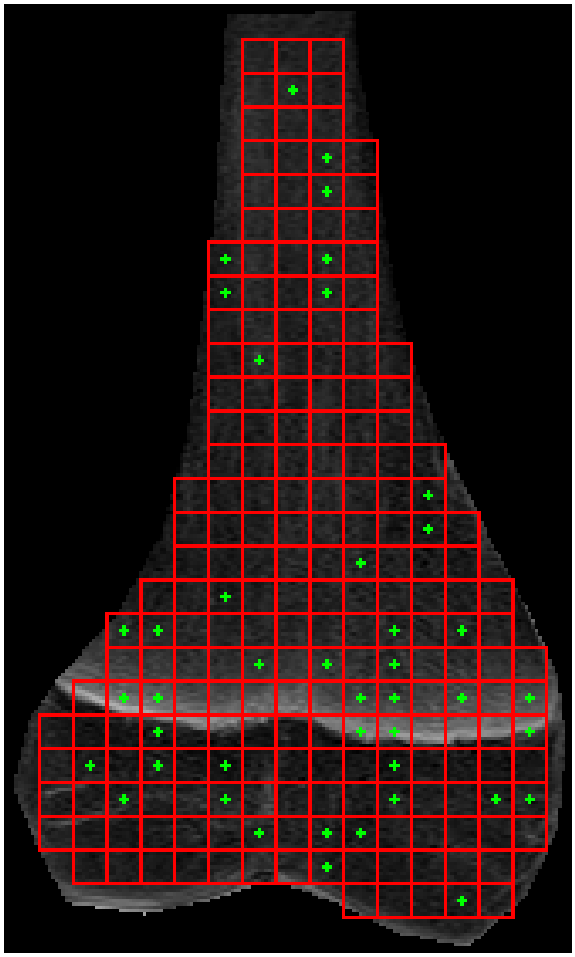


Figure A.3: Second example of an execution of the CNN model on a non-edema patch.

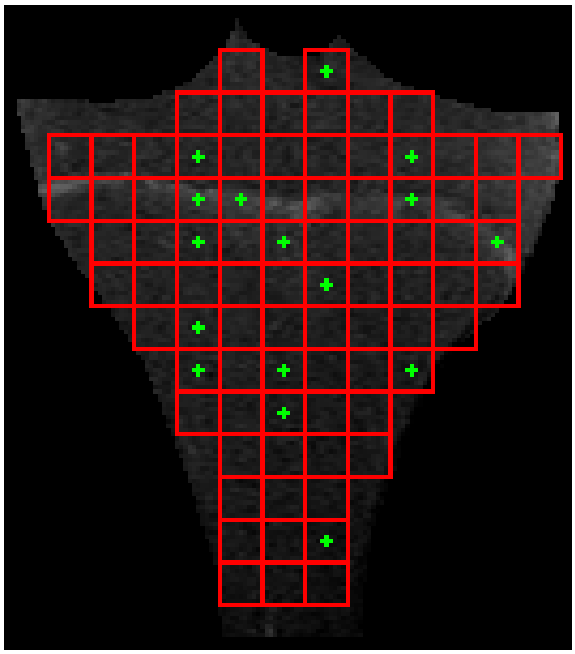


Figure A.4: Third example of an execution of the CNN model on a non-edema patch.

References

[1] M. Aizerman. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[2] Haldun Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93, September 2018.

[3] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6, Aug 2017.

[4] Hiam Alquran, Isam Abu Qasmieh, Ali Mohammad Alqudah, Sajidah Alhammouri, Esraa Alawneh, Ammar Abughazaleh, and Firas Hasayen. The melanoma skin cancer detection and classification using support vector machine. In *2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–5. IEEE, 2017.

[5] B. Astuto Arouche Nunes, I. Flament, R. Shah, M. Bucknor, T. Link, V. Pedoia, and S. Majumdar. MRI-based multi-task deep learning for cartilage lesion severity staging in knee osteoarthritis. *Osteoarthritis and Cartilage*, 27:S398–S399, April 2019.

[6] R Artusi, P Verderio, and EJTIjobm Marubini. Bravais-pearson and spearman correlation coefficients: meaning, test of hypothesis and confidence interval. *The International journal of biological markers*, 17(2):148–151, 2002.

[7] Ismail Babaoğlu, Oğuz Fındık, and Mehmet Bayrak. Effects of principle component analysis on assessment of coronary artery diseases using support vector machine. *Expert Systems with Applications*, 37(3):2182–2185, March 2010.

[8] Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, and Björn Ottersten. Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, 51:134–142, 2016.

[9] Asa Ben-Hur, David Horn, Hava Siegelmann, and Vladimir Vapnik. Support Vector Clustering. *Journal of Machine Learning Research*, 2:125–137, November 2001.

[10] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[11] 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*, 10:1178623X17703382–1178623X17703382, April 2017. Publisher: SAGE Publications.

[12] Antoine Buetti-Dinh, Vanni Galli, Sören Bellenberg, Olga Ilie, Malte Herold, Stephan Christel, Mariia Boretska, Igor V. Pivkin, Paul Wilmes, Wolfgang Sand, Mario Vera, and Mark Dopson. Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. *Biotechnology Reports*, 22:e00321, June 2019.

[13] Ken Chen, Yaqi Peng, Shengyong Lu, Binbin Lin, and Xiaodong Li. Bagging based ensemble learning approaches for modeling the emission of PCDD/Fs from municipal solid waste incinerators. *Chemosphere*, page 129802, January 2021.

[14] Patrick M. Connor, David M. Banks, Alan B. Tyson, James S. Coumas, and Donald F. D'Alessandro. Magnetic Resonance Imaging of the Asymptomatic Shoulder of Overhead Athletes: A 5-Year Follow-up Study. *The American Journal of Sports Medicine*, 31(5):724–727, September 2003. Publisher: SAGE Publications Inc STM.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[16] Shaveta Dargan, Munish Kumar, Maruthi Rohit Ayyagari, and Gulshan Kumar. A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning. *Archives of Computational Methods in Engineering*, 27(4):1071–1092, September 2020.

[17] Arden Dertat. Applied Deep Learning - Part 4: Convolutional Neural Networks, November 2017.

[18] Erik F Eriksen. Treatment of bone marrow lesions (bone marrow edema). *BoneKEy Reports*, 4, November 2015.

[19] Jerome Fan, Suneel Upadhye, and Andrew Worster. Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine*, 8(1):19–20, January 2006. Publisher: Cambridge University Press.

[20] Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering*, 2(1):602–609, December 2014. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21642583.2014.956265.

[21] D. T. Felson, C. E. Chaisson, C. L. Hill, S. M. Totterman, M. E. Gale, K. M. Skinner, L. Kazis, and D. R. Gale. The association of bone marrow lesions with pain in knee osteoarthritis. *Annals of Internal Medicine*, 134(7):541–549, April 2001.

[22] Peter A Flach. The geometry of roc space: understanding machine learning metrics through roc isometrics. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 194–201, 2003.

[23] Hao Fu, Qian Zhang, and Guoping Qiu. Random Forest for Image Annotation. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, Lecture Notes in Computer Science, pages 86–99, Berlin, Heidelberg, 2012. Springer.

[24] Fei Gao, Teresa Wu, Jing Li, Bin Zheng, Lingxiang Ruan, Desheng Shang, and Bhavika Patel. SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis. *Computerized Medical Imaging and Graphics*, 70:53–62, December 2018.

[25] Salvatore Gitto, Renato Cuocolo, Domenico Albano, Vito Chianca, Carmelo Messina, Angelo Gambino, Lorenzo Ugga, Maria Cristina Cortese, Angelo Lazzara, Domenico Ricci, Riccardo Spairani, Edoardo Zanchetta, Alessandro Luzzati, Arturo Brunetti, Antonina Parafioriti, and Luca Maria Sconfienza. MRI radiomics-based machine-learning classification of bone chondrosarcoma. *European Journal of Radiology*, 128:109043, July 2020.

[26] Sergio González, Salvador García, Javier Del Ser, Lior Rokach, and Francisco Herrera. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 64:205–237, December 2020.

[27] Benjamin Graham. Fractional Max-Pooling. *arXiv:1412.6071 [cs]*, May 2015. arXiv: 1412.6071.

[28] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, May 2018.

[29] Christian Hendrich, Engelmaier Franz, Gerhart Waertel, Rolf Krebs, and Marcus Jäger. Safety of autologous bone marrow aspiration concentrate transplantation: initial experiences in 101 patients. *Orthopedic Reviews*, 1(2):e32, October 2009.

[30] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012. Conference Name: IEEE Signal Processing Magazine.

[31] Tin Kam Ho. Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, page 278, USA, August 1995. IEEE Computer Society.

[32] Martin Hofmann. Support vector machines-kernels and the kernel trick. *Notes*, 26(3):1–16, 2006.

[33] Siegfried Hofmann. The painful bone marrow edema syndrome of the hip joint. *Wiener Klinische Wochenschrift*, 117(4):111–120, 2005.

[34] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, 1968. _eprint: https://physoc.onlinelibrary.wiley.com/doi/pdf/10.1113/jphysiol.1968.sp008455.

[35] Roman Ilin, Thomas Watson, and Robert Kozma. Abstraction hierarchy in deep learning neural networks. pages 768–774, May 2017.

[36] Dimitrios C. Karampinos, Stefan Ruschke, Michael Dieckmeyer, Maximilian Diefenbach, Daniela Franz, Alexandra S. Gersing, Roland Krug, and Thomas Baum. Quantitative MRI and spectroscopy of bone marrow. *Journal of Magnetic Resonance Imaging*, 47(2):332–353, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.25769.

[37] Ender Konukoglu and Ben Glocker. Chapter 19 - Random forests in medical image computing. In S. Kevin Zhou, Daniel Rueckert, and Gabor Fichtinger, editors, *Handbook of Medical Image Computing and Computer Assisted Intervention*, The Elsevier and MICCAI Society Book Series, pages 457–480. Academic Press, January 2020.

[38] Kyung-Hoi Koo, In-Oak Ahn, Rokho Kim, Hae-Ryong Song, Soon-Taek Jeong, Jae-Boem Na, Yong-Sik Kim, and Se-Hyun Cho. Bone Marrow Edema and Associated Pain in Early Stage Osteonecrosis of the Femoral Head: Prospective Study with Serial MR Images. *Radiology*, 213(3):715–722, December 1999. Publisher: Radiological Society of North America.

[39] Peter R. Kornaat, Milko C. de Jonge, and Mario Maas. Bone marrow edema-like signal in the athlete. *European Journal of Radiology*, 67(1):49–53, July 2008.

[40] Robert Kozma, Roman Ilin, and Hava T. Siegelmann. Evolution of Abstraction Across Layers in Deep Learning Neural Networks. *Procedia Computer Science*, 144:203–213, January 2018.

[41] Elizabeth A. Krupinski. Current perspectives in medical image perception. *Attention, perception & psychophysics*, 72(5), July 2010.

[42] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[43] Zhuoling Li, Minghui Dong, Shiping Wen, Xiang Hu, Pan Zhou, and Zhigang Zeng. CLU-CNNs: Object detection for medical images. *Neurocomputing*, 350:53–59, July 2019.

[44] C. Lindner, S. Thiagarajah, J. M. Wilkinson, The arcOGEN Consortium, G. A. Wallis, and T. F. Cootes. Fully Automatic Segmentation of the Proximal Femur Using Random Forest Regression Voting. *IEEE Transactions on Medical Imaging*, 32(8):1462–1472, August 2013. Conference Name: IEEE Transactions on Medical Imaging.

[45] Carolina Freitas Lins, Carlos Ernesto Garrido Salmon, and Marcello Henrique Nogueira-Barbosa. Applications of the Dixon technique in the evaluation of the musculoskeletal system. *Radiologia Brasileira*, 54:33–42, January 2021. Publisher: Publicação do Colégio Brasileiro de Radiologia e Diagnóstico por Imagem.

[46] Chien-Shun Lo and Chuin-Mu Wang. Support vector machine for breast mr image classification. *Computers & Mathematics with Applications*, 64(5):1153–1162, 2012.

[47] Greg Lovell, Howard Galloway, Will Hopkins, and Ann Harvey. Osteitis Pubis and Assessment of Bone Marrow Edema at the Pubic Symphysis With MRI in an Elite Junior Male Soccer Squad. *Clinical Journal of Sport Medicine*, 16(2):117–122, March 2006.

[48] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, May 2019.

[49] 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. *Pattern Recognition*, 91:216–231, 2019.

[50] Arya Minaie, Katherine Woolley, Ryan Smith, Olohirere Ezomo, Richard Feinn, Bernadette Mele, Thomas Martin, Juan Garbolosa, and Karen M. Myrick. Detecting Bone Marrow Edema With Magnetic Resonance Spectroscopy: A Brief Report. *The Journal for Nurse Practitioners*, 16(8):e129–e132, September 2020.

[51] Amir Mosavi, Sina Ardabili, and Annamaria R Varkonyi-Koczy. List of deep learning models. In *International Conference on Global Research and Education*, pages 202–214. Springer, 2019.

[52] MM Mukaka. A guide to appropriate use of Correlation coefficient in medical research. *Malawi Medical Journal : The Journal of Medical Association of Malawi*, 24(3):69–71, September 2012.

[53] Karen M. Myrick, Richard Feinn, Juan Garbalosa, Thomas Martin, Bernadette Mele, and David Wallace. Magnetic Resonance Imaging Findings of Bone Marrow Edema in a Division I NCAA Women's Soccer Team During a Competitive Season. *The Journal for Nurse Practitioners*, 13(6):425–432, June 2017.

[54] Jie C. Nguyen, B. Keegan Markhardt, Arnold C. Merrow, and Jerry R. Dwek. Imaging of Pediatric Growth Plate Disturbances. *RadioGraphics*, 37(6):1791–1812, October 2017. Publisher: Radiological Society of North America.

[55] Mahesh Pal. Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222, 2005.

[56] F. Preiswerk, M. Sury, J. Wortman, and J. Duryea. Application of deep learning for the quantitative assessment of bone marrow lesions (BMLs). *Osteoarthritis and Cartilage*, 27:S388–S389, April 2019.

[57] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 448–455, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

[58] Deepshikha Shrivastava, Sugata Sanyal, Arnab Kumar Maji, and Debdatta Kandar. Chapter 17 - Bone cancer detection using machine learning techniques. In Sudip Paul and Dinesh Bhatia, editors, *Smart Healthcare for Disease Diagnosis and Prevention*, pages 175–183. Academic Press, 2020.

[59] Steven Simske. Chapter 1 - Introduction, overview, and applications. In Steven Simske, editor, *Meta-Analytics*, pages 1–98. Morgan Kaufmann, January 2019.

[60] A. J. Wilson, W. A. Murphy, D. C. Hardy, and W. G. Totty. Transient osteoporosis: transient bone marrow edema? *Radiology*, 167(3):757–760, June 1988.

[61] Feng Yang, Guohui Wei, Hui Cao, Mengmeng Xing, Shuxuan Liu, and Jing Liu. Computer-Assisted Bone Fractures Detection Based on Depth Feature. *IOP Conference Series: Materials Science and Engineering*, 782:022114, April 2020. Publisher: IOP Publishing.

[62] Pei Yang, Chuanbao Bian, Xin Huang, Anli Shi, Chunsheng Wang, and Kunzheng Wang. Core decompression in combination with nano-hydroxyapatite/polyamide 66 rod for the treatment of osteonecrosis of the femoral head. *Archives of Orthopaedic and Trauma Surgery*, 134(1):103–112, January 2014.

[63] W. Zbijewski, A. Sisniega, J. W. Stayman, G. Thawait, N. Packard, J. Yorkston, S. Demehri, J. Fritz, and J. H. Siewerdsen. Dual-energy imaging of bone marrow edema on a dedicated multi-source cone-beam CT system for the extremities. In Christoph Hoeschen and Despina Kontos, editors, *Medical Imaging 2015: Physics of Medical Imaging*, volume 9412, pages 206 – 211. International Society for Optics and Photonics, SPIE, 2015.