Department of Radiology

School of Medicine, University of Crete

Artificial intelligence in bone marrow imaging: development of novel machine learning strategies for the diagnosis and classification of benign bone marrow pathology with the use of magnetic resonance imaging (MRI)

(Τεχνητή νοημοσύνη στην απεικόνιση του οστικού μυελού: ανάπτυξη καινοτόμων τεχνικών μηχανικής μάθησης για τη διάγνωση και ταξινόμηση καλοήθους παθολογίας του οστικού μυελού με τη χρήση Τομογραφίας Μαγνητικού Συντονισμού)

A thesis submitted for the degree of

Doctor of Philosophy

University of Crete

Michail Klontzas

May 2023

Τριμελής Επιτροπή

Επιβλέπων: **Απόστολος Καραντάνας**, Καθηγητής Ακτινολογίας

Συνεπιβλέποντες: **Κώστας Μαριάς**, Καθηγητής Επεξεργασίας Εικόνας με Έμφαση στην Ιατρική Απεικόνιση και Υπολογιστική Ιατρική, ΕΛΜΕΠΑ

Αριστείδης Ζιμπής, Καθηγητής Ανατομίας, Παν. Θεσσαλίας

Επταμελής Επιτροπή

1. Απόστολος Καραντάνας, Καθηγητής Ακτινολογίας
2. Κώστας Μαριάς, Καθηγητής Επεξεργασίας Εικόνας
3. Αριστείδης Ζιμπής, Καθηγητής Ανατομίας
4. Δημήτριος Τσέτης, Καθηγητής Επεμβατικής Ακτινολογίας
5. Κωνσταντίνος Περισυνάκης, Καθηγητής Ιατρικής Φυσικής
6. Δημήτριος Φωτιάδης, Καθηγητής Βιοϊατρικής Τεχνολογίας
7. Έλενα Κρανιώτη, Επίκουρη Καθηγήτρια Ιατροδικαστικής

Supervisor: **Apostolos Karantanas**, MD, PhD (Professor of Radiology)

Co-Supervisors: **Kostas Marias**, DiplEng, MSc, PhD (Professor of Medical Image Processing, Hellenic Mediterranean University)

Aristeidis Zibis, MD, PhD (Professor of Anatomy, University of Thessaly)

Abstract

Bone marrow edema (BME) is a non-specific finding that can accompany a wide variety of conditions affecting the bone marrow including acute trauma, acute bone marrow edema syndromes (transient osteoporosis, regional migratory osteoporosis), chronic regional pain syndrome, avascular necrosis, infection, inflammatory arthritis, osteoarthritis at advanced stages of the disease, tendinopathies, and primary and metastatic malignancies. The imaging modality of choice for the depiction of bone marrow edema is magnetic resonance imaging (MRI) with fluid-sensitive sequences. The appearance of BME on MRI can complicate the diagnosis of diseases affecting the bone marrow, creating diagnostic dilemmas that challenge general radiologists or even specialized musculoskeletal radiologists. The most important diagnostic challenges faced in everyday radiological practice are (a) the differentiation between transient osteoporosis of the hip and avascular necrosis, (b) the differentiation between subchondral insufficiency fractures of the knee and advanced osteoarthritis presenting with BME and (c) the accurate staging of avascular necrosis. Accurate diagnosis in these cases is of utmost importance since it can change the treatment from conservative (for transient osteoporosis) to surgical (for avascular necrosis) or can determine the choice between joint preserving surgery (early stages of avascular necrosis) and total hip arthroplasty (late stages of avascular necrosis. The aim of this PhD was to leverage the power of novel image analysis methods such as radiomics and deep learning to tackle the aforementioned diagnostic dilemmas. Radiomics includes the extraction of high-dimensional data from regions of interest that can be used for the detailed characterization of lesions. Artificial intelligence (traditional machine learning or deep learning) methods can be used to either analyse radiomics data or to independently perform image recognition tasks for diagnostic purposes attempting the automation of disease detection. For the purposes of this PhD, radiomics data have been utilized for the analysis of proximal femurs with either avascular necrosis or transient osteoporosis of the hip, used to train machine learning models to distinguish between the two conditions. These models achieved excellent performance in distinguishing between the two conditions, performing equally to musculoskeletal radiologists and better than a general radiologist. In addition, in

order to further automate the diagnosis between these two conditions, and to avoid bias related to the manual steps for radiomics data preparation, deep learning was used to distinguish between the two using whole images. Three convolutional neural networks (CNNs) were trained with a transfer learning methodology and finetuned with our data, in order to diagnose between transient osteoporosis and avascular necrosis. The consensus decision between the three CNNs was found to be highly accurate, performing better than two experts. Subsequently, a CNN ensemble was used to differentiate between subchondral insufficiency fractures and advanced osteoarthritis of the knee. The consensus decision of the network ensemble was compared to the diagnosis of expert radiologists. This CNN ensemble was found to be highly accurate in the differentiation between the two conditions performing better than one of the two experts. Finally, CNNs were used to distinguish between early (ARCO 1-2) and late stages of avascular necrosis (ARCO 3-4). The consensus decision of three CNNs was found to reach high performance in this diagnostic task. To further validate the model, a dataset from another country was used to assess model performance on unknown data and this validation performance was compared to the diagnosis of expert readers. Despite the performance drop in the external dataset, the CNN ensemble was still highly accurate in recognizing late AVN achieving a performance similar to the two experts. In conclusion, the work presented herein demonstrated the potential of radiomics and deep learning to assist diagnostic decisions in some of the most complicated tasks related to the presence of BME on MRI.

Περίληψη

Το οίδημα του οστικού μυελού αποτελεί ένα μη-ειδικό εύρημα το οποίο συνοδεύει μια σειρά παθήσεων που επηρεάζουν το μυελό των οστών όπως το τραύμα, τα οξέα σύνδρομα οιδήματος του οστικού μυελού (παροδική οστεοπόρωση και περιοχική μεταναστευτική οστεοπόρωση), το σύνδρομο χρόνιου περιοχικού πόνου, η ανάγγεια νέκρωση (οστεονέκρωση) της μηριαίας κεφαλής, λοιμώξεις, φλεγμονώδεις αρθροπάθειες, προχωρημένη οστεοαρθρίτιδα, τενοντοπάθειες, πρωτοπαθείς κακοήθειες και μεταστάσεις. Η απεικονιστική μέθοδος εκλογής για την απεικόνιση του οιδήματος στον οστικό μυελό είναι η Μαγνητική Τομογραφία με ακολουθίες ευαίσθητες στα υγρά. Η απεικόνιση του οστεομυελικού οιδήματος σε μαγνητική τομογραφία μπορεί να περιπλέξει τη διαφορική διάγνωση παθήσεων που επηρεάζουν τον οστικό μυελό, δημιουργώντας διαγνωστικά διλήμματα που προβληματίζουν όχι μόνο γενικούς ακτινολόγους αλλά και ακτινολόγους με εξειδίκευση στο μυοσκελετικό σύστημα. Τέτοια διλήμματα στην καθημερινή διαγνωστική πράξη αποτελούν (α) η διαφορική διάγνωση παροδικής οστεοπόρωσης του ισχίου από οστεονέκρωση της μηριαίας κεφαλής, (β) η διάκριση μεταξύ κατάγματος ανεπάρκειας του γόνατος και προχωρημένης οστεοαρθρίτιδας η οποία συνοδεύεται από οστεομυελικό οίδημα και (γ) η ακριβής σταδιοποίηση της ανάγγειας νέκρωσης της μηριαίας κεφαλής. Η ακριβής διάγνωση στις προαναφερθείσες περιπτώσεις είναι εξαιρετικής σημασίας καθώς μπορεί να καθοδηγήσει την απόφαση για χειρουργική θεραπεία (νέκρωση κεφαλής, προχωρημένη οστεοαρθρίτιδα) σε σχέση με συντηρητική θεραπεία (παροδική οστεοπόρωση, κάταγμα ανεπάρκειας γόνατος). Επίσης ακριβής σταδιοποίηση της ανάγγειας νέκρωσης της μηριαίας κεφαλής καθορίζει την απόφαση για ολική αρθροπλαστική του ισχίου σε προχωρημένα στάδια της νόσου. Ο στόχος της παρούσας διατριβής είναι η χρήση σύγχρονων μεθόδων ραδιωμικής ανάλυσης και τεχνητής νοημοσύνης για την επίλυση των προαναφερθέντων διαγνωστικών διλημμάτων. Η ραδιωμική ανάλυση αποτελεί μέθοδο με την οποία εξάγονται λεπτομερή δεδομένα από ιατρικές εικόνες τα οποία χρησιμοποιούνται στη συνέχεια για τον ακριβή χαρακτηρισμό περιοχών ενδιαφέροντος. Η τεχνητή νοημοσύνη μπορεί να προσφέρει μεθόδους για την ανάλυση δεδομένων ραδιωμικής καθώς και για την ανάλυση ολόκληρης εικόνας με μοντέλα βαθιάς μάθησης (deep learning). Για

τους σκοπούς της παρούσας διατριβής δεδομένα ραδιωμικής εξήχθησαν από εικόνες μαγνητικής τομογραφίας εγγύς μηριαίων με παροδική οστεοπόρωση ή οστεονέκρωση της μηριαίας κεφαλής. Τα δεδομένα αυτά χρησιμοποιήθηκαν για την εκπαίδευση μοντέλων μηχανικής μάθησης για τη διάκριση μεταξύ των δυο παθήσεων. Τα μοντέλα αυτά πέτυχαν άριστη διάκριση μεταξύ των δύο παθήσεων, ενώ έδωσαν διαγνώσεις με ακρίβεια όμοια με ακτινολόγους που έχουν λάβει εξειδικευμένη εκπαίδευση στο μυοσκελετικό σύστημα. Η μέθοδος αυτή φάνηκε να έχει καλύτερα αποτελέσματα σε σχέση με γενικό ακτινολόγο. Στη συνέχεια, σε μια προσπάθεια πλήρους αυτοματοποίησης της διαγνωστικής διαδικασίας και αποφυγής λαθών που σχετίζονται με τη χρονοβόρα διαδικασία εξαγωγής δεδομένων ραδιωμικής, χρησιμοποιήθηκαν τρία μοντέλα βαθιάς μάθησης που εκπαιδεύτηκαν να αναγνωρίζουν εικόνες παροδικής οστεοπόρωσης και οστεονέκρωσης. Τα μοντέλα αυτά έλαβαν συναινετική απόφαση για τη σωστή διάγνωση σε κάθε εικόνα, αποδίδοντας καλύτερα από ακτινολόγους μυοσκελετικού. Ομοίως μοντέλα βαθιάς μάθησης εκπαιδεύτηκαν να διακρίνουν μεταξύ υποχόνδρινων καταγμάτων ανεπάρκειας του γόνατος και προχωρημένης οστεοαρθρίτιδας. Ο συνδυασμός των μοντέλων αυτών απέδωσε καλύτερα ή το ίδιο σε σχέση με ειδικούς ακτινολόγους. Τέλος, βαθιά μάθηση χρησιμοποιήθηκε για τη διάκριση μεταξύ πρώιμης (ARCO 1-2) και προχωρημένης (ARCO 3-4) οστεονέκρωσης της μηριαίας κεφαλής. Η συναινετική απόφαση τριών μοντέλων βαθιάς μάθησης απέδωσε εξαιρετικά στη διάγνωση μεταξύ των δυο καταστάσεων. Προκειμένου μάλιστα να επιβεβαιωθεί η απόδοση των μοντέλων έγινε αξιολόγησή τους με τη χρήση εικόνων από κέντρο του εξωτερικού και στη συνέχεια πραγματοποιήθηκε σύγκριση με τη διάγνωση από δυο εξειδικευμένους ακτινολόγους μυοσκελετικού. Παρά τη μικρή μείωση στην απόδοση των μοντέλων όταν δοκιμάστηκαν σε εξωτερικά δεδομένα, φάνηκε να διατηρούν υψηλή συνολική απόδοση η οποία κρίθηκε όμοια με την απόδοση των δυο εξειδικευμένων ακτινολόγων. Συμπερασματικά, τα αποτελέσματα που παρουσιάζονται στην παρούσα διδακτορική διατριβή επιβεβαιώνουν ότι η ραδιωμική ανάλυση και μέθοδοι τεχνητής νοημοσύνης όπως η βαθιά μάθηση μπορούν να αποδώσουν εξαιρετικά στη λήψη συγκεκριμένων διαγνωστικών αποφάσεων οι οποίες σχετίζονται με την παρουσία οστεομυελικού οιδήματος και οι οποίες προκαλούν διλήμματα στην καθ' ημέρα διαγνωστική πράξη των γενικών και εξειδικευμένων ακτινολόγων.

Acknowledgements

When I decided to embark on a second PhD, I didn't expect that the journey would differ significantly to the first one. Working on a purely clinical project, in parallel with hospital duties and on an orthogonally different subject presented unique challenges but also great opportunities. I had the chance to delve into the exciting world of AI which has the potential to revolutionize medical imaging while tacking clinically relevant problems on musculoskeletal radiology. This opportunity has helped me mature as a researcher and radiologist and has challenged me to develop skills necessary for radiologists of the future.

This thesis would have not been completed without the invaluable help of my mentor Prof. Apostolos Karantanas who has been supportive before, during and after the completion of this thesis and has been true inspiration that drove me to become a radiologist and to love the musculoskeletal system. He has supported my career and development without reservations and has been the mentor that every junior would like to have, teaching not only research and MSK radiology but also life lessons for myself and all the junior family of our department. The support of my co-supervisors Prof Kostas Marias and Prof Aristeidis Zibis has also been extremely important for the completion of my thesis. Their advice on artificial intelligence and orthopaedic topics, their encouragement and their support for the publication of my results have been crucial for the completion of this work and is greatly appreciated.

Over the course of this PhD and during my radiology training, my "work family" and true friends were there to help with research, make every day in the department a true joy! Wednesday gatherings after MSK clinics with Yannis Stathis, George Kakkos, Kostas Spanakis, George Kavalaris, Alex Kotziamanis will be never forgotten!

Last but not least, this PhD is dedicated to my wife (Pelagia) and my son (Manolis) who have been wholeheartedly supporting my work and tolerating my absence without complaints!

Funding by the Maria M. Manassaki Scholarship and the Young Researcher's Grant of the European Society of Musculoskeletal Radiology (ESSR) is greatly appreciated.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AVN: avascular necrosis

TOH: transient osteoporosis of the hip

BME: bone marrow edema

XGboost: Xtreme Gradient Boosting

MRI: Magnetic Resonance Imaging

ARCO: Association Research Circulation Osseous

ROC: Receiver Operating Characteristics

AUC: Area Under the Curve

CNN: Convolutional Neural Network

MSK: Musculoskeletal

SVM: Support Vector Machines

List of Publications

Journal publications relevant to this thesis

1. **(Original Research) Klontzas M.E.**, Vassalou E.E., Spanakis K., Meurer F., Woertler K Zibis A.H., Marias K., Karantanas A.H. "Deep learning enables the differentiation between early and late stages of hip avascular necrosis", Under revision - European Radiology, 2023
2. **(Original Research) Klontzas M.E.**, Vassalou E.E., Kakkos G.A., Spanakis K., Zibis A.H., Marias K., Karantanas A.H.. "Differentiation between subchondral insufficiency fractures and advanced osteoarthritis of the knee using transfer learning and an ensemble of convolutional neural networks". Injury, 53(6): 2035-2040, 2022
3. **(Original Research) Klontzas M.E.,** Stathis I., Spanakis K., Zibis A.H., Marias K., Karantanas A.H.. " Deep learning for the differential diagnosis between transient osteoporosis and avascular necrosis of the hip". Diagnostics, 12(8):1870, 2022
4. **(Original Research) Klontzas M.E.,** Manikis G.C., Nikiforaki K., Vassalou E.E., Spanakis K., Stathis I., Kakkos G.A., Matthaiou N., Zibis A.H., Marias K., Karantanas A.H.. "Radiomics and machine learning can differentiate transient osteoporosis from avascular necrosis of the hip". Diagnostics, 11(9):1686, 2021
5. **(Review Paper) Klontzas M.E.**, Papadakis G.Z., Marias K. and Karantanas A.H. "Musculoskeletal trauma imaging in the era of novel molecular methods and artificial intelligence." Injury, 51(12):2748- 2756, 2020.

Peer reviewed abstracts

1. **Klontzas M.**, Vassalou E., Spanakis K., Zibis A., Marias K., Karantanas A, Staging avascular necrosis of the hip with the use of deep learning, Annual Meeting of the European Society of Musculoskeletal Radiology, June 2023
2. **Klontzas M.**, Vassalou E., Kakkos G. et al. Untangling the nature of subchondral bone marrow lesions of the knee with the use of deep learning-a multi-center cross-sectional study, European Congress of Radiology, Vienna, March 2022

Chapter 1 Introduction

Bone marrow edema (BME) is a non-specific finding that can be found in a series of benign and malignant conditions. The imaging modality of choice for the depiction of BME is Magnetic Resonance Imaging (MRI), where BME can be identified as bone marrow areas of low-signal intensity of T1-w sequences and high-signal intensity on fluid-sensitive sequences such as Short Tau Inversion Recovery (STIR) and Proton Density -weighted fat saturated sequences [1, 2]. BME can accompany a series of conditions affecting the bone marrow including acute bone marrow edema syndromes (transient osteoporosis, regional migratory osteoporosis), chronic regional pain syndrome, avascular necrosis (AVN), infection, inflammatory arthritis, osteoarthritis at advanced stages of the disease, tendinopathies, and primary/metastatic malignancies [1–3].

Our group and other research groups have previously shown that the morphology and the location of BME can be extremely helpful or even pathognomonic in the diagnosis of certain diseases. Such cases include the half-moon appearance of BME at the femoral neck, which can indicate the presence of intra-articular osteoid osteoma (in cases with no history of lower limb overuse) or stress reactions/fractures ("half-moon" sign) [4, 5], and the lack of BME in the inner lower quadrant of the femoral head which is characteristic of acute BME syndromes ("sparing sign") [6]. Nonetheless, despite the progress in the diagnosis of conditions accompanied by BME, its identification on routine MRI examinations can be extremely confusing for inexperienced readers, requiring extensive expertise in musculoskeletal radiology and a combination of imaging and clinical features. Even for experienced MSK radiologists, certain diagnostic dilemmas still exist that are related to the presence of BME and can complicate diagnosis of routine MRI examinations. Such dilemmas include:

1. The differentiation between transient osteoporosis and advanced stages of avascular necrosis of the hip [1, 7–9]
2. The differentiation between subchondral insufficiency fractures of the knee and advanced osteoarthritis [10–12]
3. The accurate staging of AVN, where BME usually appears in later stages of the disease [13, 14]

Accurate diagnosis in all the aforementioned cases is of utmost importance for the selection of an appropriate treatment plan which in cases of AVN and advanced osteoarthritis could involve surgery, whereas in cases of transient osteoporosis and subchondral insufficiency fractures would require conservative treatment with reduced weightbearing. In addition, accurate staging of AVN would play a role in selecting either a joint preserving surgical method (e.g. core decompression) in early stages or total hip arthroplasty in late stages of the disease [14, 15].

Radiomics analysis and artificial intelligence methods have rapidly infiltrated radiological research and clinical practice over the past decade [16]. Radiomics includes the extraction of high-dimensional mathematical image features imperceptible to the human eye from a region of interest and the analysis of that data, usually with artificial intelligence/machine learning methods, to enable diagnostic decisions [17, 18]. Artificial intelligence methods, including traditional machine learning algorithms (e.g. support vector machines, random forests, gradient boosting methods) and deep learning models, can be used to analyse radiomics data but can be also used independently to analyse images for diagnostic purposes [19–21]. These methods allow certain degrees of automation of the diagnostic process and have been shown to reach accurate diagnostic decisions especially when dealing with very specific diagnostic tasks such as fracture detection [22], even outperforming human readers.

The work presented in this PhD leverages the power of radiomics and artificial intelligence to address the aforementioned diagnostic dilemmas related to the presence of BME and is structured to reflect the use of radiomics and artificial intelligence to address multiple diagnostic dilemmas related to the presence of BME. The most important background and a critical appraisal of current literature has been presented in chapter 2, followed by a detailed description of aims/objectives and a general methodology section (chapters 3 and 4, respectively). Subsequently, starting from the differentiation between transient osteoporosis and avascular necrosis of the hip, chapter 5 demonstrates that radiomics analysis achieved high accuracy in the differentiation between the two conditions, while reaching a performance that is equal or better than human readers. In order to overcome the manual steps related to radiomics feature extraction and region of interest segmentation, deep learning has been used in chapter 6 to automate the diagnosis between AVN and transient osteoporosis of the hip. A consensus convolutional neural network (CNN) ensemble has been shown to accurately distinguish between the two conditions using whole MR images of the hip, without the need for manual segmentation and radiomics feature extraction. This novel CNN ensemble has been then used in chapter 7 to distinguish between the BME caused by subchondral insufficiency fractures of the knee and advanced osteoarthritis. The ensemble was found to be highly accurate in the differentiation between the two conditions and the performance of the deep learning pipeline was found equal or better than expert MSK radiologists. In chapter 8, a similar CNN ensemble achieved accurate distinction between early and late stages of AVN which was validated with an international image dataset and was found more accurate that expert readers. Finally, chapter 9 provides a critical conclusion, future perspectives and directions for further research.

Chapter 2 Literature Review

This chapter has been partly published in Klontzas et al. Injury 2020 [16]

Reuse of own publication data in a thesis is permitted by Elsevier.

2.1 Bone marrow edema

2.1.1 Pathophysiology and imaging

Bone marrow is the primary site of haematopoiesis in a normal adult. It consists of a bony trabecular network that creates spaces filled with cells, water, fat and protein that changes according to the age and in various disease states, with a variable proportion of red and yellow component at different ages. In a healthy adult, red haematopoietic marrow is predominant in early stages of life, being converted gradually to yellow (fatty) marrow as age increases [9, 23].

In order to achieve a deep understanding of bone marrow imaging for the recognition of normal image variants and the differential imaging diagnosis between conditions that affect the bone marrow, a basic knowledge of bone marrow histology and physiology is extremely important. The pathophysiology of bone marrow edema is still unclear, however, it has been proposed that it reflects a complex phenomenon with a combination of trauma-related sequelae including vascular rupture, hypervascularity and abundant blood perfusion, and physiological adaptation mechanisms which differ depending the underlying condition[9, 23].

The appearance of bone marrow on MR imaging is mainly dependent on the concentrations of water and fat in-between bony trabeculae. Since bone appears without signal on MR sequences, the appearance of bone marrow is dependent on the concentration of fluid (such as in cases of BME), as well as on the relative concentration of red and yellow marrow components. Suppression of the signal of fat with sequences such as fat suppressed Proton Density (PD)-weighted and Short tau Inversion Recovery sequences is extremely important because of the similar signal of water and fat in T2-w sequences. Suppression of the signal of fat in the aforementioned sequences reveals the signal of fluid (fluid-sensitive) sequences allowing the identification of edema [24].

In fluid-sensitive sequences, red marrow is distinguished from BME based on the fact that it yields less intense signal, is contained within anatomical barriers (e.g. epiphyseal plate) and is characteristically found in individuals in need for haematopoiesis (e.g. smokers, patients with anaemia, patients with high body mass index, individuals living in high altitudes etc.) and in young adults[24, 25].

Areas of BME can have certain morphological characteristics depending on the underlying condition. A representative example is the half-moon sign of BME in the femoral neck, where BME retains a half-moon morphology in cases of an underlying osteoid osteoma of the femoral neck or in cases of stress fractures of the femoral neck [5, 26]. Another example is the sparing of the medial portion of the femoral head by BME in early stages of transient osteoporosis of the hip [6]. Such morphological characteristics could be attributed to the configuration of cancellous bone struts such as the network of principal compressive and tensile trabeculae which can form boundaries for the containment of BME [4].

2.1.2 Conditions appearing with BME

2.1.2.1 Acute bone marrow edema syndromes

Acute bone marrow edema syndromes include transient osteoporosis of the hip (TOH) and regional migratory osteoporosis (RMO). These syndromes are characterized by the acute onset of pain without a history of trauma and reversibility of the symptoms only with conservative treatment. The equivalent of transient osteoporosis in the knee is called subchondral insufficiency fracture (SIF) of the knee and its pathophysiology is considered to be the same as transient osteoporosis of the hip. The exact mechanism behind acute bone marrow edema syndromes is still unclear, even though transient bone demineralization is thought to be implicated, as well as hyperaemia leading to demineralization, neurogenic compression or ischemia of vessels supplying nerve roots [27, 28].

On MRI they are characterized by the presence of extensive BME with the presence of subchondral irregular low signal intensity lines withing the area of BME representing subchondral fractures in almost 50% of the patients (**Figure *1***) [6]. The main site affected by TOH is the hip where edema can also involve the acetabulum apart from the femoral head[1–3, 9]. On plain radiographs, osteopenia may be detected but the modality of choice for the diagnosis of bone marrow edema syndromes is MRI with fluid sensitive sequences. Paramagnetic contrast administration demonstrates enhancement of the region of BME, as well as joint effusion and synovitis, even though contrast is not routinely administered since it is not necessary for the diagnosis of BME. A combination of the history (acute onset without trauma) and characteristic imaging appearance is the basis of the diagnosis. When found in the proximal femur, BME can extend to the greater trochanter and it usually spares the medial part of the femoral head . In cases where no proper treatment measures are taken, articular collapse will be noted as a consequence of the subchondral insufficiency fractures.

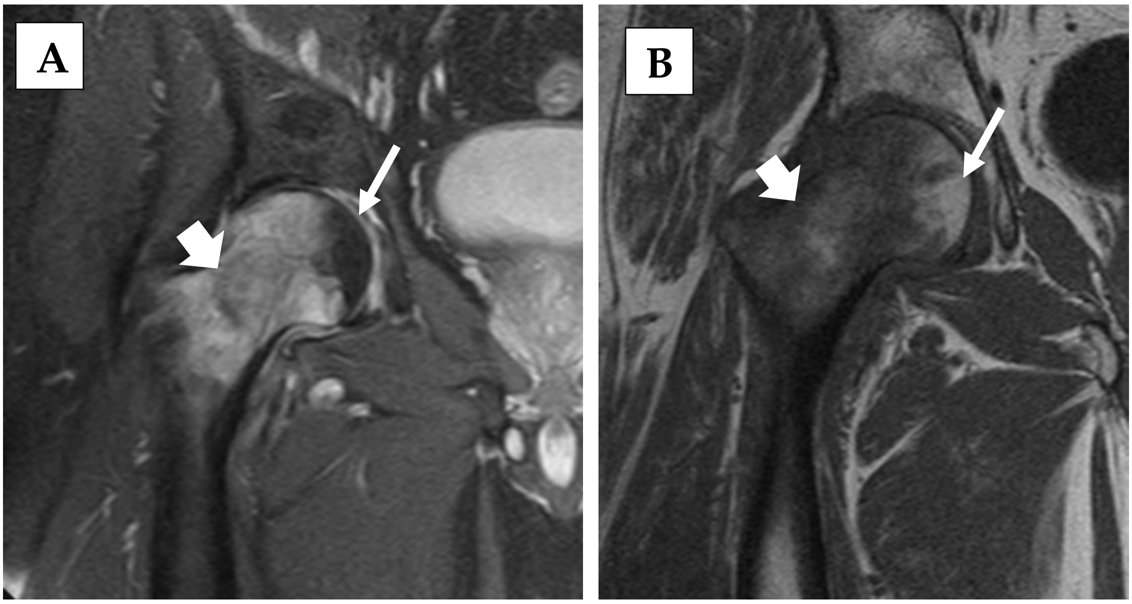


Figure 1 A 60-year-old male patient with transient osteoporosis of the right hip. Extensive bone marrow edema (thick arrows) appears with high signal intensity on coronal STIR (A) and low signal intensity on coronal T1-weighted (B) images of the proximal femur. Bone marrow edema extends to the area of the greater trochanter and spares the bone marrow of the medial part of the femoral head (sparing sign – thin arrows).

RMO is a condition where migration of the edema is noted, a while after the initial TO event [29]. Such migration can happen to other locations such as the knee or the ankle and has the same imaging and pathophysiological characteristics as TO. The chance of edema migration has been estimated to be close to 20% in a large cohort of patients with TOH, however, other studies have reported a migration chance of up to 72%[30]. A characteristic feature of patients with acute bone marrow edema syndromes is the low bone mineral density as shown on femoral neck and spinal

DEXA examinations [2, 11]. Perfusion imaging in these patients shows delayed peak enhancement at the areas of BME [31].

2.1.2.2 Avascular necrosis

Avascular necrosis (AVN) of the hip affects approximately 20,000 patients every year in the USA and is the most common cause of total hip arthroplasty (THA) at young ages, demonstrated with bilateral involvement in >70% of patients [14]. AVN results from irreversible ischemia in the subchondral area which results to necrosis of the subchondral bone that loses its mechanical properties resulting to articular collapse. Apart from fractures that disrupt the blood supply to the femoral head, a series of other risk factors have been described for the appearance of AVN, including alcohol consumption, smoking, radiation, hypercoagulation states, lipid storage diseases, autoimmune disorders and others [32].

Clinically, patients may be completely asymptomatic or present with pain around the hip, groin, buttock linked to reduced internal rotation of the hip and potential pain radiation to the knee [32, 33] and one of the previously mentioned risk factors can be usually noted.

In early stages of the disease prior to articular collapse, joint preservation techniques (core decompression, vascularized grafting, etc) are available with the potential to avoid THA which is the last resort after articular collapse and the development of secondary osteoarthritis [34, 35]. Therefore, differentiation between early and late AVN is of utmost importance for appropriate treatment selection.

Several systems have been proposed for the staging of AVN with the most popular being the Association Research Circulation Osseous (ARCO) [36], the Ficat and Arlet classification [37] and the Steinberg (University of Pennsylvania) classification[38]. The ARCO classification is the commonly used one and is suggested as the classification of choice in the 2019 international guidelines on the management of AVN [32]. According to the latest version of ARCO, AVN can be divided in 4 stages as shown on Table 1.

Table 1 The latest (2019) revision of the ARCO staging system

ARCO stage	**Image Findings**	**Details**
I	Unremarkable X-ray	“Band-like” sign and “single line” sign on fluid sensitive sequences. A cold spot can be seen on bone scan
	Findings on MRI	
II	Findings on X-ray	“Band-like” sign on T1-w sequences and “single line” sign on fluid sensitive sequences. On X-rays or CT, osteosclerosis, localized osteopenia, cystic lesions. No evident fracture, no flattening of the head.
	Findings on MRI	
IIIA		Femoral head depression up to 2 mm
IIIB		Femoral head depression >2 mm
IV		Findings indicative of osteoarthritis

As suggested by the 2019 guidelines for the management of AVN, for cases with ARCO <3 joint preserving surgical techniques (core decompression, stem cell transplantation or bone grafting/osteotomy) and/or vasodilators/anticoagulation treatment is recommended. For early ARCO 3 and above, joint replacement starts to become an option depending on the availability of vascularized bone grafting and the condition of the joint. For this reason distinguishing between early and late AVN is extremely important for the selection of appropriate treatment. This can be extremely complicated due to the presence of bone marrow edema that is in most cases present in ARCO>3 [39] but can be sometimes also seen in early disease. In addition, evaluation of articular surface collapse should be done in 3D and can be extremely complicated for inexperienced readers, especially in the absence of an evident crescent sign.

Another difficulty in the diagnosis of AVN lies in the confusion between AVN and TOH. This has been a long-lasting literature debate, especially in the previous decade [40–44]. Our group and other research groups have shown that there is no relationship between the two conditions in terms of histology [42], clinical progression to articular collapse (which is not a feature of TOH) [6], the uptake of contrast in dynamic contrast enhanced MRI studies [44] and the morphology of subchondral fracture lines [6]. The irregular thin subchondral fractures encountered in TOH have a completely different appearance to either the serpiginous band-like sign with a concavity towards the joint space that surrounds the necrotic portion of the head, or the fluid-filled crescent sign (subchondral fracture) seen in late stages of AVN. Nonetheless, despite the accumulating evidence showing no relationship between the two entities, studies continue to be published with a pronounced confusion between the two entities [45, 46]. The importance of distinguishing between the two lies in the fact that a patient diagnosed with TOH will undergo conservative treatment whereas a patient with AVN will most probably undergo some type of surgical management.

2.1.2.3 Other conditions appearing with BME

A series of other conditions are associated with BME including acute trauma, stress injuries, chronic regional pain syndrome, infections, inflammatory arthropathies, benign and malignant tumours. However, these conditions fall out of the spectrum of this PhD since no significant diagnostic dilemmas arise in relationship with BME in these conditions, since other primary findings (e.g. erosions, fracture lines, cartilage damage, masses) are used for the diagnosis of these conditions in conjuction with disease-specific clinical details such as inflammatory markers and signs of inflammation, history of trauma or overuse, and history of malignancy at another site are used to reach a final diagnosis.

2.2 Principles of radiomics

"Omics" analyses were first introduced in biological sciences with the development of fields such as genomics, transcriptomics, proteomics, metabolomics, epigenomics etc. These "omics" analyses performed a global characterization of biological systems by quantifying all measurable molecules of a defined class [47–49]. In this manner, biology shifted from probing concentrations of single metabolites or single proteins to the wholistic snapshot of a biological system at a given time point . This has been the epitome of personalized medicine since it allowed the detailed characterization of disease states, the modelling of cellular physiology and the detection of druggable targets for the development of novel therapeutic molecules [17, 50–54].

Radiomics includes the extraction of high-dimensional mathematical image features imperceptible to the human eye from a region of interest and the analysis of that data, usually with artificial intelligence/machine learning methods, to enable diagnostic decisions [17, 18]. As accurately mentioned by Gillies et al. [17] "images are more than pictures, they are data" which can be used for the detailed characterization of lesions or areas of interest. Radiomics allow precise characterization of lesions capturing fine details in the imaging appearance of a disease phenotype that is ultimately affected by changes in all other omics layers (from genome to metabolome). Radiomics can be used to pinpoint correlations between image characteristics and disease states or treatment responses being able to predict current or future biological traits or the response to a certain treatment [55–60].

Radiomics analysis involves a series of relatively well-defined steps starting from the definition of the lesion of interest and the modality of choice, proceeding with the image acquisition and pre-processing steps to prepare for feature extraction that can be done either in a traditional machine learning manner or in a deep-learning manner in cases of abundant data. These features are then used to train and validate artificial intelligence models that enable predictions of predefined outcomes (**Figure 2**).

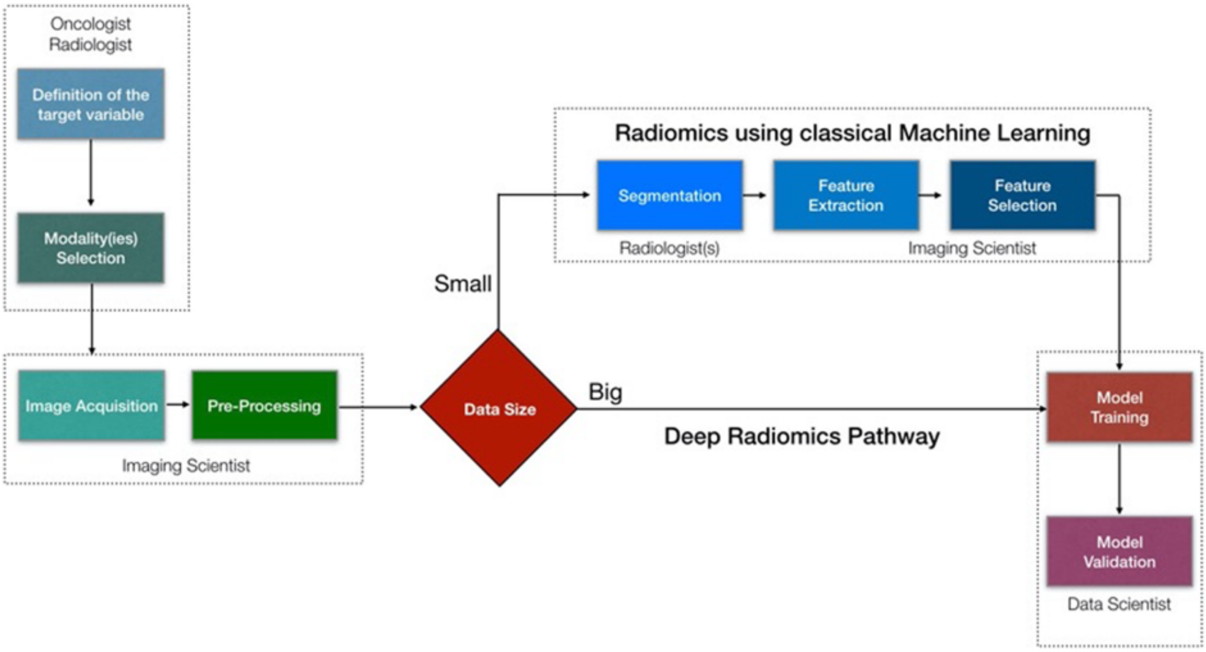


Figure 2 A multidisciplinary radiomics workflow
Initially a group of clinicians should define the clinical problem that the proposed model should deal with and make decisions on what kind of imaging modalities should be recruited. Imaging scientists needs to make sure that acquisition protocols are optimally designed producing high quality images, as well as for the pre-processing of the images. Then depending on the size of the available imaging studies we need to decide which pipeline to use. In case of big data (in the order of thousands) a deep radiomics approach can be suggested avoiding tedious and time-consuming processes like tumor segmentation by multiple radiologists. In addition, deep convolutional neural networks have been proven more efficient to model complex problems compared with traditional machine learning algorithms, as long as data availability requirement is satisfied. Finally, the data sets are allocated for training, validation and testing purposes (reproduced under CC BY license from [61]).

The features extracted from medical images can be classified into semantic or agnostic [17, 61]. Semantic features are part of radiologists' everyday language, used to describe lesion characteristics that include size, shape, volume, length, diameter, vascularity and others. These lesions are included in radiomics, not in a qualitative manner but in a rather quantitative manner. On the contrary, agnostic features are generally not understood by every radiologist since they quantify relationships between pixels and voxels that are not simple to conceptualize. These include first order statistics, second order statistics and high-level statistics. First order features include summary numbers and histogram statistics such as mean, median, maximum/minimum, entropy, asymmetry or flatness of histogram values. These are different to texture or second-order features, that attempt to reveal relationships between voxels that are characterized by similar contrast providing a measure of lesion heterogeneity that cannot be otherwise described with semantic features. Such heterogeneity can explain why lesions that look similar on imaging have a very different biological behaviour, since the human eye cannot visualize such texture differences. A multitude of texture features can be extracted (hundreds) to obtain a detailed description of a lesion. Finally, higher order features can be obtained by applying filters to the original image that augment certain patterns. Such filters include wavelets, that transform the image by matrix multiplication with a series of radial or linear waves and Laplacian transformations of Gaussian filters that identify regions with high coarseness [17, 62]. Wavelet transformation highlights edges and irregular lines and eliminates noise profile inconsistencies, being thus valuable in the analysis of images originating from multiple scanners [63, 64].

Attempts are made to standardize the way radiomics features are extracted in order to promote reproducible research and standardized biomarker identification. Such initiatives include the Imaging Biomarker Standardization Initiative [65, 66] which is an international collaboration on standardizing the way radiomics features are extracted. Another attempt for standardization has been made by the creators of the PyRadiomics package[58] which is the most commonly used method for the extraction of radiomics data from medical images.

Following data extraction and dataset preparation, machine learning methods are employed for the training and validation of potentially useful/predictive models. These are classification or regression models which include a series of traditional machine learning algorithms such as random forests, gradient boosting models, support vector machines (SVM) and artificial neural networks. In this thesis SVM and gradient boosting methods (XGboost, CatBoost) have been used and will be further explained. Gradient Boosting algorithms represent methods that have been extremely successful in handling tabular data (e.g. radiomics).

They include methods such as XGBoost, AdaBoost, LightBoost and CatBoost which are ensemble models combining a series of weak classifiers in an attempt to optimize accuracy and robustness. The first of this model class was AdaBoost which was published in 1997 [67, 68]. The core of the gradient boosting concept includes the optimization of a loss function e.g log-loss (classification) and a series of decision trees (weak learners) which offer potential data splits for optimal classification that are subsequently added one at a time to the ensemble while minimizing loss with a gradient descent method [69]. Gradient descent represents an optimization methods that identifies local minima of loss function [70]. XGBoost is the most popular and powerful of gradient descent methods that has won a series of data science and machine learning competitions (https://github.com/dmlc/xgboost/blob/master/demo/README.md#machine-learning-challenge-winning-solutions). XGBoost minimizes a regularized loss function in an attempt to reduce model complexity and can be scaled up without significant increases in resources [71]. The backbone of XGboost uses C++ but has been also implemented in R and python packages with relatively simple commands. Multiple radiomics studies have successfully utilized gradient boosting algorithms [72–74].

Support Vector Machines (SVM) is a well-known traditional machine learning algorithm suitable for data-limited scenarios. SVM computes a geometrical way to maximize the difference between predefined classes. For this purpose it creates a hyperplane that passes between data points of each class after subjecting the dataset to a transformation with a kernel function. Subsequently the hyperplane that maximizes the distance between data classes is computed [68].

2.3 Fundamentals of deep neural networks

Deep learning has been recently applied to medical imaging, in order to avoid the manual extraction of hand-crafted radiomics features which provide a detailed but rather time consuming analysis and potentially biased characterization of images. The revolution of deep learning came with the development of powerful graphics cards (GPUs) that can handle complex calculations (convolutions) that allow neural networks to analyse images. Neural networks obtained their name by the resemblance of their structure to the complex connections between neurons in the human brain (**Figure** 3).

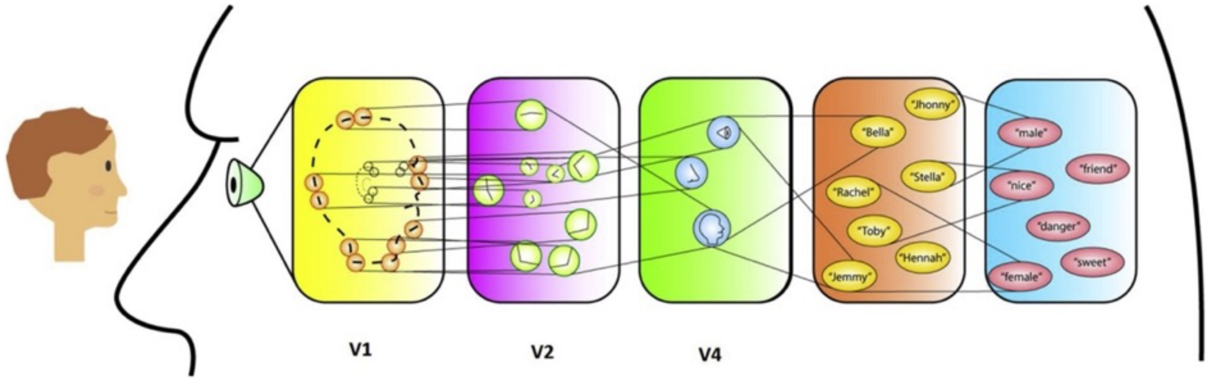


Figure 3 A neural network representation of human brain
Reproduced with permission from Saba L et al. [75]
Image courtesy of https://grey.colorado.edu/CompCogNeuro/index.php/CCNBook/Perception

The most important neural networks that revolutionized medical imaging were convolutional neural networks (CNNs) [75]. The first CNN was introduced by LeCun et al. in 1989 for the recognition of handwritten digits in Zip codes of the US Postal Service [76]. The network worked by calculating convolutions ie. merging operations for the information of a set of pixels in the image. By going through a series of convolutional layers the image is analysed to derive information that can in the end yield a prediction probability denoting whether an object belongs to a specific class. Each layer is connected to the others with weights that can be visualized as power of a synapse between two neurons. Deep learning has been feasible due to the development of backpropagation by Geoffrey Hinton et al. [77], which is the method that CNNs use to iteratively modify their weights by comparing their output to the ground truth until they "learn" the appropriate weights that allow an accurate prediction.

Several deep learning architectures have been proposed with the first being AlexNet by Krizhevsky et al. from the group of Geoffrey Hinton, which achieved classification of 1.2 million images to 1000 classes with an error of 15.2% in 2012 [75, 78]. Since then a series of CNN architectures have been developed with the most commonly used being VGG-16, Inception-ResNet V2 and InceptionV3. The architectures of these networks can be found in **Figure 4**.

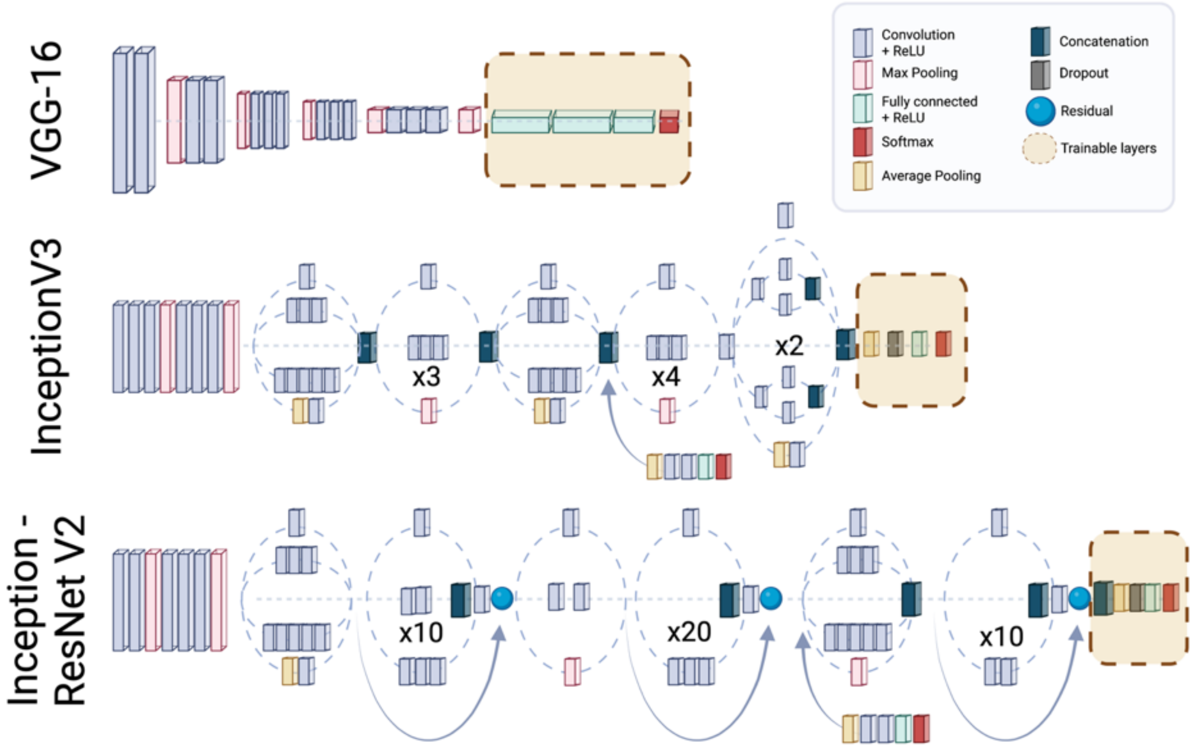


Figure 4 CNN architectures used in this thesis (created with biorender.com)

These models have achieved excellent results in classifying medical images such as the detection of pneumonia in chest x-rays (96% AUC for VGG-16) [79], fracture detection (95.4% AUC for InceptionV3) [80] and the detection of sacral fractures on radiographs (98.9% and 98.4% AUC for InceptionV3 and Inception ResNetV2 respectively) [81].

2.4 Transfer learning

Deep learning models require an abundance of image data in order to iteratively learn the appropriate weights to achieve the optimal performance. However, in cases of where diseases with low prevalence are studied or where the local population does not allow the construction of a sufficient dataset, transfer learning has been proposed to reduce the need for a large training dataset while significantly reducing computational costs [82]. In the most common form of transfer learning CNNs are trained using a large dataset to obtain the initial weights and to learn recognizing patterns common in all kinds of images (e.g. edges). Subsequently the initial weights are frozen and the final trainable layers are finetuned using the smaller local dataset to allow the CNN to learn the details of each specific application. Datasets used for transfer learning include ImageNet with >14 million images of everyday objects (including cars and animals) and the recently launched RadImageNet with 5 million medical (CT, MRI, ultrasound, PET) images from 500,000 patients. RadImageNet was recently launched (2022) and no studies (apart from the original RadImageNet publication) have yet utilized it for the training of CNNs [83].

2.5 Artificial intelligence and radiomics for the study of the musculoskeletal system

2.5.1 Radiomics for the study of musculoskeletal disease

2.5.1.1 Oncological applications

Radiomics has been used in MSK radiology mainly for oncological applications. Radiomics data have been used to distinguish between benign and malignant bone lesions such as the differentiation between sacral chordoma and giant cell tumour [60] and the differentiation between osteoblastic metastases and bone islands [84] with an AUC of almost 95% for the former and 96% for the latter. It is also interesting that when radiomics models were compared to human diagnostic performance for the differentiation between bone islands and metastases, the radiomics-based method was equal to two readers and better than a third reader [84], demonstrating the potential of the method to assist diagnostic decisions. Similar radiomics models using SPECT data, CT data and their combination achieved AUC of approximately 92% for the differentiation between benign and malignant sclerotic bone lesions [85].

Apart from osteoblastic lesions, cartilaginous tumours have also been analysed with radiomics by Gitto et al. who have established a radiomics method for the differentiation between atypical cartilaginous tumours (grade I chondrosarcoma) and grade II chondrosarcoma using MR images. The importance of development of an accurate radiomics model for such am application is that in cases of atypical cartilaginous tumours watchful waiting or intralesional curettage can be applied whereas for grade II chondrosarcoma a more aggressive wide resection is indicated [56]. Yi et al. have also shown that radiomics from T2-w images are superior to contrast enhanced T1-w images in differentiating Ewing sarcoma from osteosarcoma [86]. Similar work has been presented on the detection of malignancy in soft tissue tumors. Indeed, a radiomics-based SVM model with linear kernel achieved a 96% AUC in the detection of malignant lipomatous soft tissue tumors [59]. However, radiomics models have not only been used for disease detection but for the prediction of the response to neoadjuvant chemotherapy in patients with osteosarcoma [55]

2.5.1.2 Non-neoplastic musculoskeletal disorders

Radiomics has found limited applications for non-neoplastic disorders compared to the number of applications noted for musculoskeletal and soft tissue tumours. In rheumatological disease, radiomics has been used for the detection of inflammatory sacroiliitis [87] and for the differentiation between different types of myopathy [88]. Application of radiomics for the detection of sacroiliitis is the only case of machine learning that has been used for the diagnosis of diseases related to BME. Finally, radiomics has been applied to the analysis of the texture of paraspinal muscles [89], which did not manage to predict fatty atrophy but could correlate with muscle strength.

2.5.2 Deep learning for the study of musculoskeletal disease

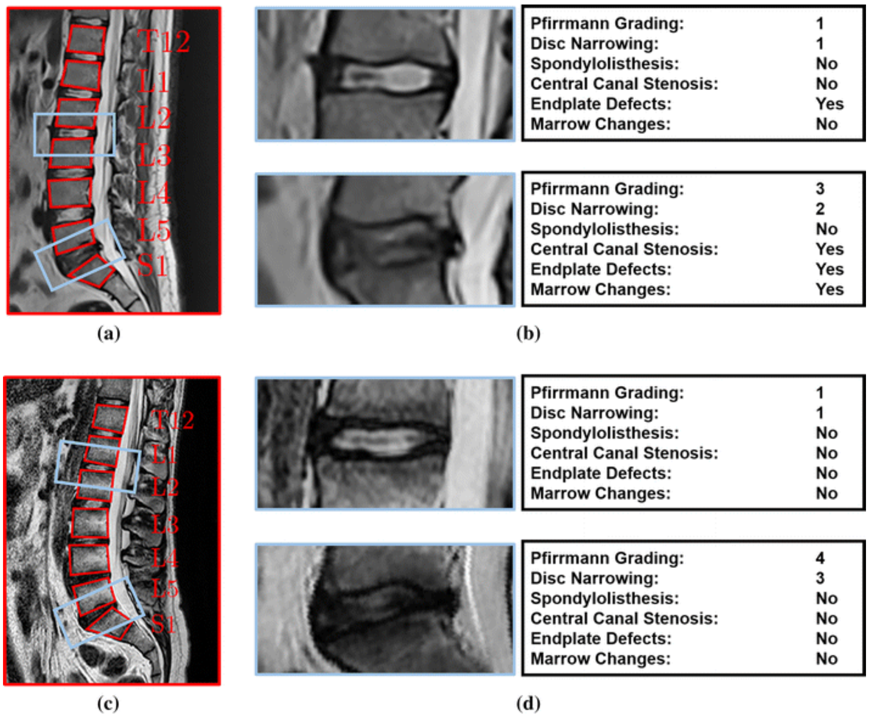
2.5.2.1 Non-neoplastic musculoskeletal disorders

With regards to deep learning, the majority of applications to the musculoskeletal system have to do with the detection of fractures on plain radiographs. Several attempts have been made with some of them achieving excellent results in large patient cohorts[22]. The field has progressed to the extent that deep learning based commercial software (e.g. Gleamer's BoneView) has been tested in large patient cohorts achieving impressive results, comparable to multiple human readers [90]. Such software can be used clinically since it has received CE mark and

FDA approval for use not only in adults but also in paediatric population with an age >2 years.

Another important application of deep learning has been the detection of meniscal and ACL tears. Fritz et al. have done a review and meta-analysis of deep learning studies for the detection of meniscal tears, showing that CNNs can achieve a high performance for the detection of medial meniscal tears but sensitivity for the detection of lateral meniscal tears was significantly lower hindering clinical application [91]. The same study showed that a series of studies have been performed for the detection of ACL tears achieving high AUCs (93% to 100%) but pooled data demonstrated that MSK radiologists performed better than most algorithms [92, 93]. Nonetheless, the authors highlighted the fact that inexperienced readers would benefit more from such deep learning algorithms[91].

Several deep learning publications exist for spinal disease. Research efforts have been focused on the automatic detection of spinal levels [94], the multi-level



Pfirrmann Grading:	1
Disc Narrowing:	1
Spondylolisthesis:	No
Central Canal Stenosis:	No
Endplate Defects:	Yes
Marrow Changes:	No

Pfirrmann Grading:	3
Disc Narrowing:	2
Spondylolisthesis:	No
Central Canal Stenosis:	Yes
Endplate Defects:	Yes
Marrow Changes:	Yes

Pfirrmann Grading:	1
Disc Narrowing:	1
Spondylolisthesis:	No
Central Canal Stenosis:	No
Endplate Defects:	No
Marrow Changes:	No

Pfirrmann Grading:	4
Disc Narrowing:	3
Spondylolisthesis:	No
Central Canal Stenosis:	No
Endplate Defects:	No
Marrow Changes:	No

Figure 5 Example of the detected region of the vertebrae and the corresponding assessments of the mid-sagittal slice of an MRI (a).

The red boxes are the detected vertebrae regions and the blue boxes are the extracted disc regions passed through to the classifier. (b) L2–L3 and L5–S1 disc volume examples from (a) and their resulting predictions computed from the disc volumes. Likewise, (d) the L1–L2 and L5–S1 disc volume examples from (c) and the predictions. Reproduced under CC BY license from Jamaludin et al. Eur Spine J 2017 [94]).

quantification of canal and foraminal stenosis [94, 95] and the detection of disc degeneration.

Finally, deep learning has been used to automate a series of non-diagnostic tasks in MSK radiology image denoising [96], examination protocolling to reduce time, segmentation of regions of interest [97], implant brand/type recognition [98, 99], body composition analysis [100] as well as the automation of measurements such as the measurement of femoral component subsidence after total hip replacement [101] and the measurement of pelvic/hip angles such as the CE angle, Tonnis angle, sourcil angle, Sharp's angle and femoral head extrusion index [102].

2.5.2.2 Oncological applications

Deep learning has been used in orthopedic oncological applications especially for the detection of pathological vertebral fractures, the identification of either sclerotic or lytic spinal metastases on CT with an AUC >80% [103]. Deep learning has achieved high performance in the differentiation between benign and malignant vertebral fractures on CT and MRI [104, 105].

The studies utilizing deep learning for the study of musculoskeletal primary tumors are limited. The only available studies to date, include the automated segmentation-grading of soft tissue sarcomas [106] which yielded moderate AUCs ~75% for the staging of sarcomas and the prediction of post-surgical recurrence of giant cell tumors based on the pre-surgical MRI which performed better than human experts [107].

Chapter 3 Aim & Objectives

The aim of this PhD project was to utilize radiomics and artificial intelligence methods to assist the diagnosis of bone marrow disease characterized by bone marrow edema. These state-of-the-art methods were used to address important clinically relevant questions where significant expertise in MSK and bone marrow imaging is required for the accurate image-based diagnosis. Ultimately, the aim of this PhD was to demonstrate the capability of radiomics and artificial intelligence methods to assist the reporting of MSK examinations where the presence of BME can raise diagnostic or treatment dilemmas. Specific objectives included:

1. To use radiomics and traditional machine learning methodology for the differentiation between transient osteoporosis (TOH) and avascular necrosis (AVN) of the hip. MRI-based radiomics from femoral head segments was used to differentiate between the two conditions (**chapter 5**).
2. To utilize deep learning models for the differentiation between TOH and AVN. Use of deep learning overcomes the need for femoral head segmentation and automates the diagnostic decision. Towards this objective, three convolutional neural network (CNNs) architectures were trained using transfer learning and finetuned with our custom dataset and the ensemble consensus decision of all three networks was recorded as the final decision of the algorithm (**chapter 6**).
3. To distinguish between BME related to subchondral insufficiency fractures of the knee (SIF) or advanced osteoarthritis, using deep learning. Transfer learning was also used to train three CNNs and their consensus decision was used as the output of the algorithm (**chapter 7**).
4. To distinguish between early (ARCO<3) or late (ARCO ≥3) stages of AVN using deep learning with a similar transfer learning methodology (**Chapter 8**).

Chapter 4 Materials & Methods

This chapter presents methodology used throughout this thesis, in more than one chapter. Methodology unique to each individual chapter is presented at the beginning of the respective chapter.

4.1 Patients & ethics

4.1.1 Avascular necrosis and transient osteoporosis of the hip

Anonymized consecutive MRI examinations of patients with avascular necrosis of the hip (AVN) and transient osteoporosis of the hip (TOH) were retrospectively collected from the archive of the second opinion bone marrow imaging clinic of our hospital (run by a MSK radiologist with 40 years of experience in MSK imaging) between July 2014 and March 2020. The aforementioned MRI examinations were performed in a multitude of MR scanners (1.5T or 3T) of multiple vendors. Exclusion criteria included: trauma, infection, tumours, inflammatory arthropathies, follow-up <1-year, prior surgery on the hip of interest [1].

4.1.2 Subchondral insufficiency fractures of the knee and advanced osteoarthritis

Consecutive MRI examinations with subchondral insufficiency fractures (SIF) or advanced osteoarthritis (OA) of the knee were retrospectively collected from the RIS-PACS system of our hospital and two collaborating clinics by retrospectively examining 1756 knee MRI examinations. All patients with SIF or OA accompanied by bone marrow edema were included. Exclusion criteria included: (i) history of recent knee trauma or intervention, (ii) clinical and laboratory diagnosis of infectious or inflammatory arthropathy and (iii) MRI findings of bone marrow reconversion extending to the epiphysis, (iv) neoplasms around the knee. Cases where SIF and OA coexisted and cases with a fluid filled subchondral fracture surrounded by BME, with or without articular collapse, were also excluded.

[1] Detailed inclusion criteria and ground truth establishment methods can be found in each individual chapter.

4.1.3 Ethics

Approval of the project was obtained by the University Hospital of Heraklion Ethical Committee (Ref. No. 360/08/29-04-2020) and informed consent was waived due to the anonymized retrospective nature of the study.

4.2 Radiomics pipeline

In order to address vendor/scanner variability between STIR MR images, grey level harmonization was achieved by means of histogram normalization and by establishing a fixed bin width according to the PyRadiomics guidelines for MRI-based radiomics (https://pyradiomics.readthedocs.io). In addition, voxel spacing standardization was applied by voxel dimension resampling to 1x1x1 mm. Proximal femurs (neck/head proximal to the intertrochanteric line) were manually segmented with paint tool of 3D slicer (v 4.11 for Windows, slicer.org) and radiomics features were extracted with the use of PyRadiomics implementation of 3DSlicer. Single reader segmentation is justified by the clear-cut boundaries of bone tissue which do not allow boundary confusion with adjacent soft-tissue structures.

Extracted radiomics features included first order, shape-based, gray level co-occurrence matrix (glcm), gray level run length matrix (glrlm), gray level size zone matrix (glszm), neighbouring gray tone difference matrix (ngtdm) and gray level dependence matrix (gldm) features, as well as their wavelet and Laplacian of Gaussian transformations. A total 849 radiomics features were extracted from each segment and used for further analysis.

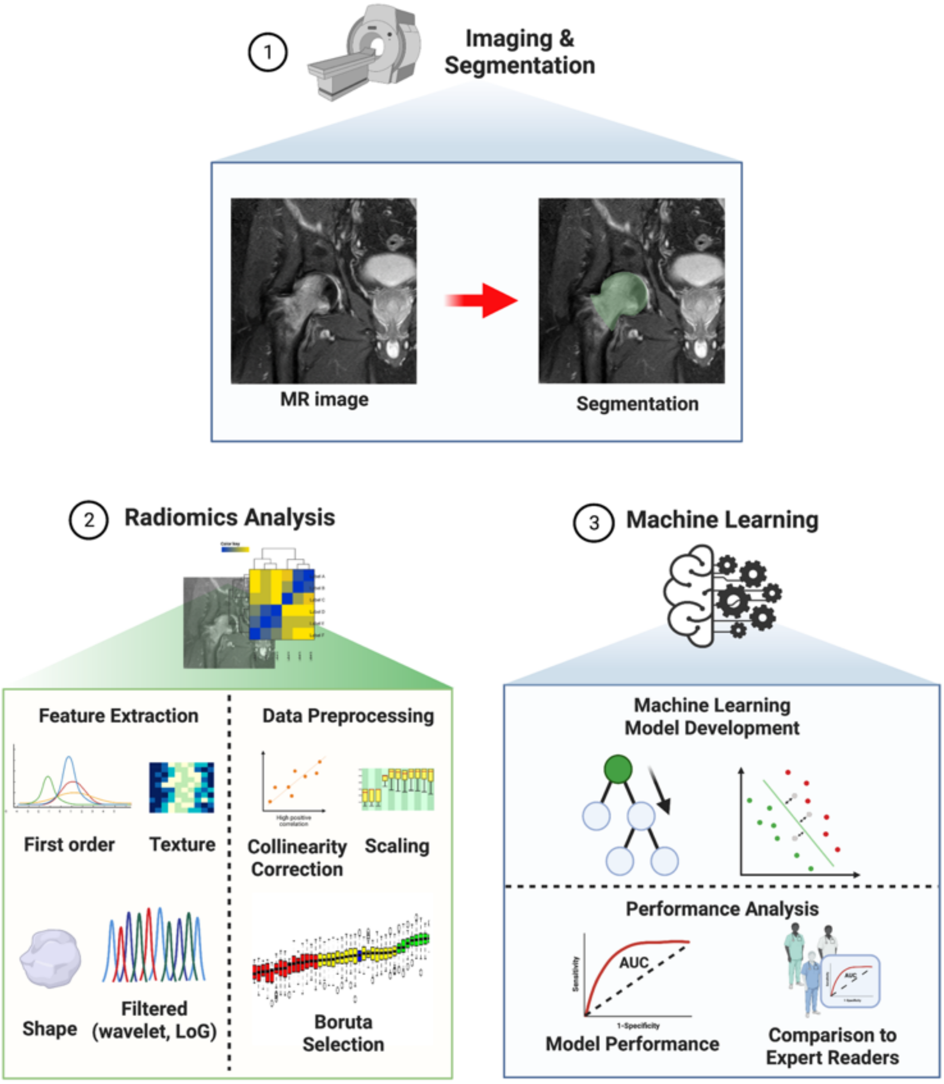


Figure 6 MRI-based radiomics pipeline

MR images are obtained and the region of interest is segmented (1). Subsequently, radiomics features are extracted and the dataset undergoes preprocessing (2) prior to use for machine learning purposes. The final dataset is used for the training and validation of machine learning models, which are evaluated for their performance and compared to expert readers.

Dataset curation included collinearity correction by removal of all features with a Pearson correlation coefficient >0.7. Relevant radiomics features were subsequently selected with the use of Boruta tree-based feature selection algorithm with a p-value threshold at 0.01, to enable the construction of meaningful signatures while limiting the possibility of overfitting (**Figure 6** MRI-based radiomics pipeline. The dataset was then split to training and testing and scaled based on the formula $RF_{scaled} = \frac{RF-\mu_{RF}}{SD_{RF}}$ prior to use for machine learning model development.

4.3 CNN training with transfer learning

Convolutional neural network architectures (VGG-16, Inception-ResNetV2, InceptionV3) were trained with a transfer learning methodology in order to tackle the limited size of our dataset as proposed by [82]. Prior to use in deep learning all images were resized to 150x150 px to fit the input specifications of the aforementioned CNN architectures. In addition, datasets were augmented by means of horizontal image flipping rotation (10^{o} clockwise/anticlockwise) to expose the network to a higher amount of training data. Firstly, the initial layers of each network were trained with the use of ImageNet dataset (public dataset of >14 million images) to allow the networks to recognize generic features that cannot be learned with the use of a small clinical dataset. Subsequently, network weights were frozen and the final trainable layers of each network were finetuned with the use of our datasets. CNN training was performed for 50 -100 epochs with an early stopping function in order to avoid overfitting.

Deep learning was performed with Python v.3.8, the Keras framework and the TensorFlow backend on a Windows 10 Pro workstation with 32 GB RAM, Intel i7-10700F @2,9 GHz CPU and NVIDIA GeForce RTX 2060 Super 8GB GPU.

4.4 Statistical analysis

Descriptive statistics were used to analyze patient demographics, presented as frequencies and mean ± standard deviation (SD). Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated for each machine learning classifier and expert reader for the detection of AVN against TOH. CNN performance was evaluated using precision, recall and f1-scores for each individual CNN and the ensemble.

Receiver operating characteristic (ROC) curves were constructed with the use of the pROC R package and classifier performance was assessed with the respective area under the curve (AUC) and 95% confidence intervals for the AUC calculated by bootstrapping. Expert reader performance (AUC) was compared to the machine learning classifiers with the use of DeLong's method [108]. Statistical analysis was performed with the use of R (v. 4.03, https://www.R-project.org/) and the non-parametric Mann–Whitney U test was used to compare the ages of patients between groups. Significance was defined with a p-value lower than a = 0.05.

Chapter 5 – Radiomics and machine learning enable the differentiation between avascular necrosis and transient osteoporosis of the hip

This chapter has been published in Klontzas et al. Diagnostics 2021 [72]. Copyright remains with the authors under CC BY license.

5.1 Introduction

Transient osteoporosis (TOH) and avascular necrosis of the hip (AVN) represent two disease entities that can be associated with bone marrow edema (BME) of the proximal femur [1, 2]. BME is the hallmark of TOH, whereas in AVN it appears at late stages of the disease [1]. This has caused a longstanding confusion in radiological and orthopaedic literature, where AVN was initially thought to be a continuation of TOH. This confusion is facilitated by the fact that subchondral fractures can be found in both entities [1, 6, 14]. Nonetheless, it is widely now accepted that they represent distinct disease entities which have a completely different pathophysiology [42] and there is no progression of TOH to AVN [6, 43]. Indeed, the appearance of subchondral fractures in TOH is completely different to the appearance of the "crescent" sign or "band-like" sign seen in AVN [14]. Differentiation between the two is of utmost importance since the management of TOH, which is a self-limiting condition, is entirely conservative (reduced weightbearing) whereas the treatment of AVN is mainly surgical. Even though accurate differentiation between the two entities is extremely important since it can lead to unnecessary surgery, great confusion still exists in literature [45, 46] and accurate diagnosis requires a combination of clinical and imaging findings and experience of the reader in MSK radiology.

To overcome the difficulty in radiological diagnosis between the two entities, radiomics analysis was employed to achieve a comprehensive analysis of image features and patterns that are indiscernible to the human eye for the differentiation between the two entities. Radiomics data were extracted from segments of the proximal femur in MR images of patients with either AVN or TOH. Omics data analysis was then performed and three machine learning classifiers were constructed to differentiate between the two conditions. Finally, the clinical value of the proposed classifiers was assessed by comparing the performance of the classifiers to the diagnostic performance of MSK and general radiologists as well as radiology residents.

5.2 Chapter-specific Methodology

5.2.1 Patients

A total of 213 consecutive hips were retrospectively included, comprising 104 hips (n=67 patients) with AVN and 109 hips with TOH (n=107 patients). Recruitment was performed as previously described.

5.2.2 MR imaging and ground truth establishment

Diagnosis was established with the combination of imaging and clinical data including (1) the nature of pain onset, with acute pain onset indicating TOH and insidious pain onset indicating AVN, (2) the presence of risk factors predisposing to AVN. In order to establish the diagnosis of TOH, all potential TOH patients were followed-up for at least 1 year to ensure spontaneous resolution of symptoms with conservative treatment. Images were obtained at several 1.5T and 3T MR scanners across the country and were evaluated at the second opinion bone marrow imaging clinic of our department. MR imaging evaluation for ground truth establishment included a minimum of (1) coronal T1-w, (2) axial fat-suppressed PD/T2-w, (3) coronal short tau inversion recovery (STIR), (4) a high-resolution 3D gradient echo sequence for the global evaluation of the hip joint. In cases, where the existing MR protocol was insufficient, the examination was repeated in our department. Images were evaluated by a MSK radiologist with 40 years of experience together with all available clinical data and in collaboration with the referring orthopaedic surgeon.

Diagnosis of AVN was established with the identification of the "band-like" sign as low-signal intensity line on T1-w images, as well as the "single line" as high signal intensity line on fluid-sensitive sequences (fs PD/T2-w and STIR). BME and subchondral fractures (high signal intensity lines on fluid sensitive sequences) indicated advanced stages of AVN (ARCO≥3). The diagnosis of TOH was established based on the identification of extensive BME at the proximal femur. Secondary findings supportive of the diagnosis included the presence of the "sparing sign" [6], synovitis and joint effusion. Ultimately, the diagnosis was confirmed by the resolution of symptoms and BME over the course of the follow-up only with conservative treatment. For the purposes of radiomics analysis mid-coronal STIR images were used since STIR images suffice for the diagnosis of both conditions in routine practice given the fact that they are the sequences of choice for the evaluation of BME.

5.2.3 Radiomics analysis and machine learning model development

Image preprocessing, radiomics feature extraction and dataset preparation was performed as described in chapter 4. In order to utilize radiomics data for the differentiation between TOH and AVN, three machine learning classifiers, XGBoost, CatBoost and Support Vector Machines (SVM) were trained using 70% of the dataset and tested using 30% of the data. Classifiers were built using 10-fold crossvalidation in the training dataset and hyperparameter tuning was performed using random search. The testing dataset could be considered as an external validation set since examinations came from a multitude of MR scanners (**Figure 7**). Models were trained and tested with the use of R programming language (v.4.03, www.R-project.org) with the use of “xgboost”, “catboost” and “e1071” packages.

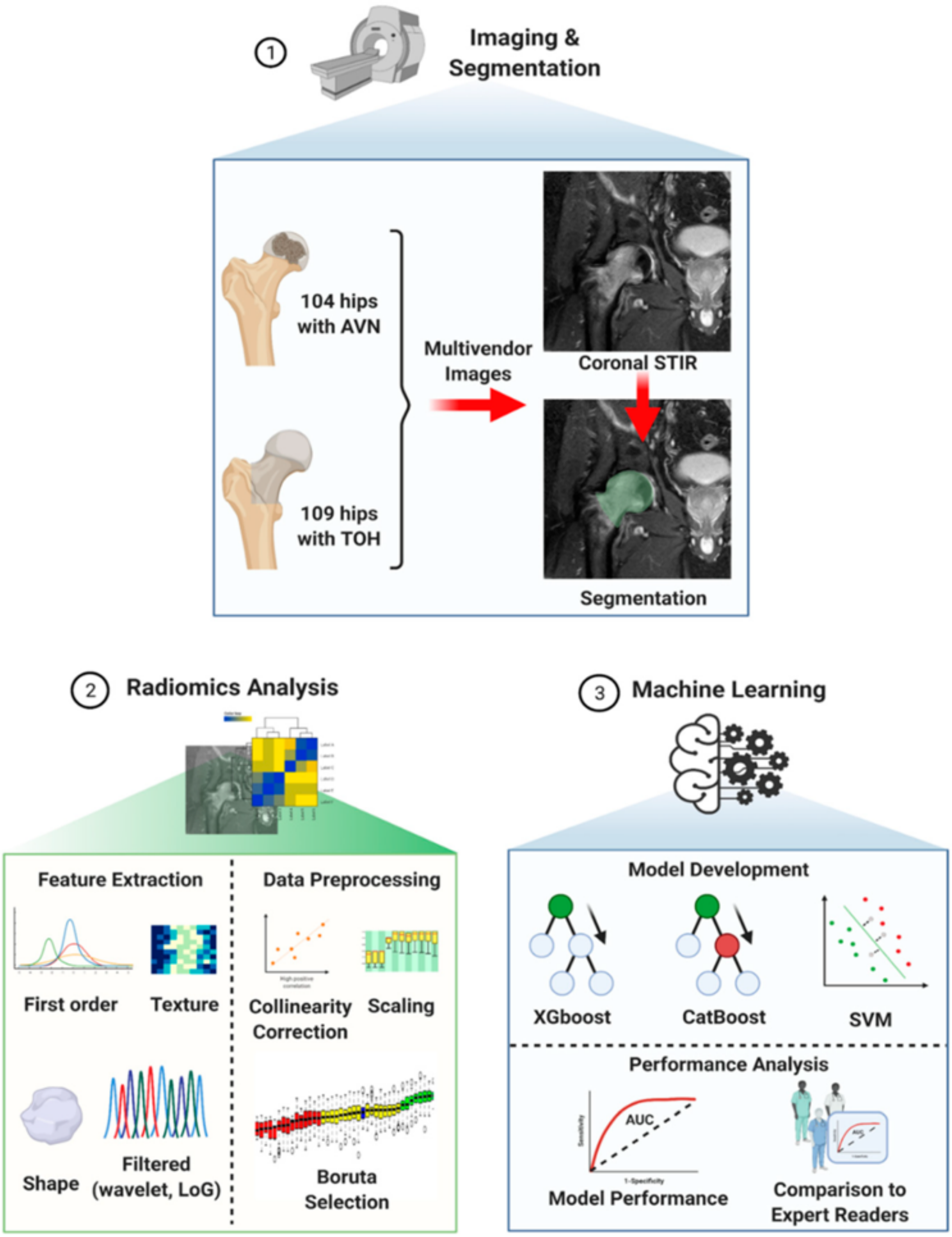


Figure 7 Computational pipeline for radiomics analysis and machine learning model development.

The process starts with image acquisition and segmentation of the femoral head and neck (1) followed by radiomics analysis (2) consisting of feature extraction and data preprocessing in preparation for subsequent model development (3). Three machine learning algorithms (XGboost, CatBoost and SVM) were trained and validated with multivendor data and their performance was compared to that of expert readers. TOH: Transient Osteoporosis of the Hip; AVN: Avascular Necrosis; STIR: Short Tau Inversion Recovery; LoG: Laplacian of Gaussian; SVM: Support Vector Machine (created with BioRender.com, reproduced under CC BY license from [72]).

5.2.4 Comparison to expert readers

Radiologists in training (a 4th and a 5th year resident with a special interest in MSK imaging), a general radiologist and two MSK radiologists (7 and 5 years of MSK experience, respectively) evaluated all images utilized for model development and provided a diagnosis blinded to the ground truth and to the predictions of radiomics models. Radiologists at all levels of experience (residents to MSK experts) were recruited in order to understand how the results of the algorithm compare to real life reporting and were presented with randomly shuffled images of our test set. In cases where no consensus could be reached by all three MSK-oriented readers (two MSK radiologists and a final year resident with MSK interest) were recorded as challenging and were used for further benchmarking of the developed algorithms.

5.3 Results

5.3.1 Dataset demographics

Patients with TOH had a mean age of 45.77 ± 10.3 years which was found similar to the age of AVN patients which was found to be 43.74 ± 14.77 years (*P*=0.464). The dataset contained a total of 119 left and 94 right hips of 61 female and 113 male patients. Details on patient demographics can be found in **Table 2**.

5.3.2 Radiomics and machine learning

Table 2 Patient demographics

	Total	AVN hips	TOH hips
Number of hips	213	104	109
Age	44.76 ± 12.53 years	43.74±14.77 years	45.77±10.3 years
Side	94R - 119L	56L - 48R	63L - 46R
Sex*	61F - 113M	38F - 29M	23F - 84M

*: number of patients; AVN: avascular necrosis; TOH: transient osteoporosis of the hip; F: female; M: male; R: right; L: left

Boruta tree-based algorithm was used for feature selection yielding a subset of 38 meaningful radiomics features (31 wavelet and 7 original) which were subsequently used for machine learning model development (**Figure 8**). XGBoost was achieved the highest performance of all three models for the differentiation between TOH and AVN reaching an AUC of 93.7% (95% CI from 87.7 to 99.8%), whereas CatBoost achieved slightly lower performance with an AUC of 92.1% (95% CI from 85.4 to 98.8%) and SVM achieved the lowest AUC of 87.4% (95% CI from 79.1 to 95.6%) (**Figure 9** and **Table 3**). Given the excellent performance of XGBoost, it was selected as the model of reference and features important to model performance were extracted. A

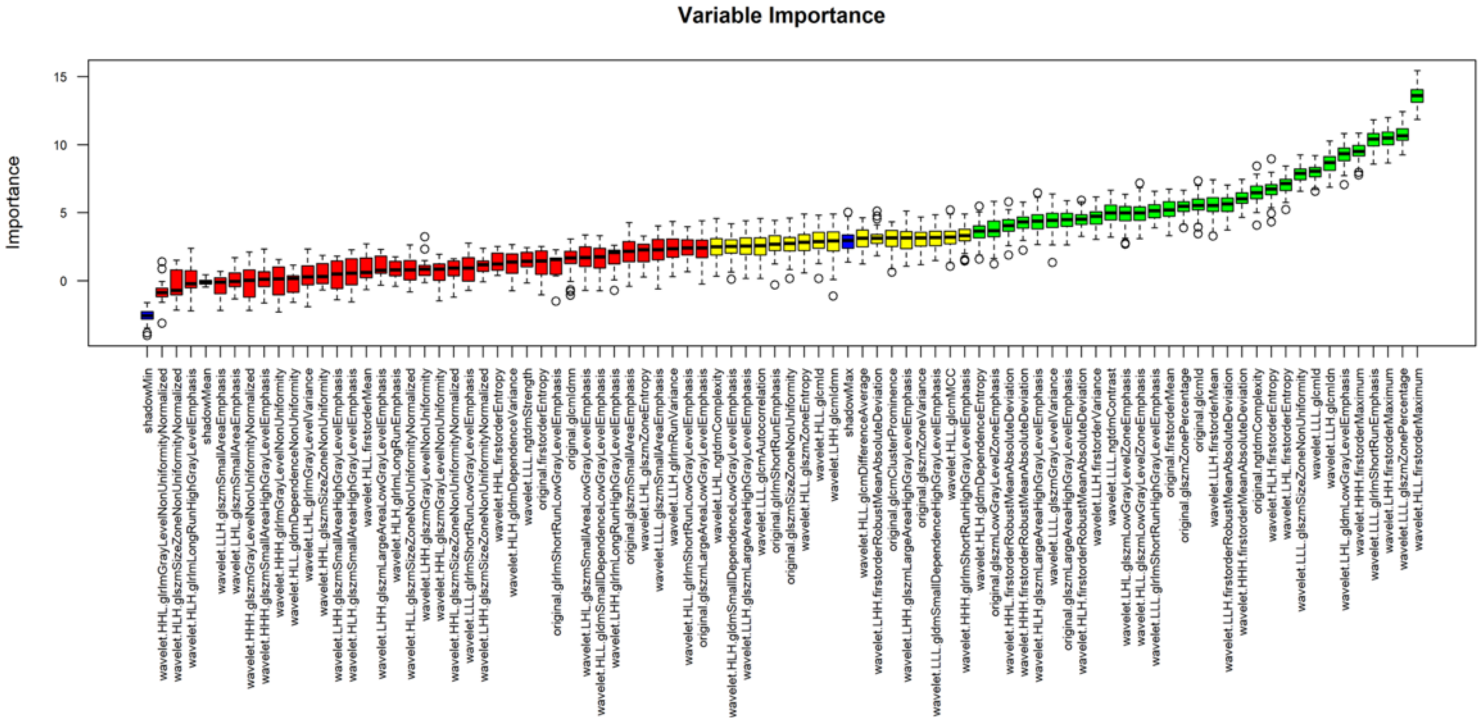


Figure 8 Identification of important features with the use of Boruta feature selection.

Following collinearity correction and scaling, Boruta was applied as an artificial intelligence algorithm to select relevant features for unbiased development of machine learning classifiers. The Z-score boxplot presents rejected (red), tentative (yellow) and accepted (green) features. P<0.01 was used as a cut-off for the selection of accepted features. Blue boxes represent Z-scores of shadow features acting as internal controls for the selection of important variables. Subsequent machine learning was performed using accepted (green) features (reproduced under CC BY license from [72]).

total of 31/38 radiomics features were important determinants of

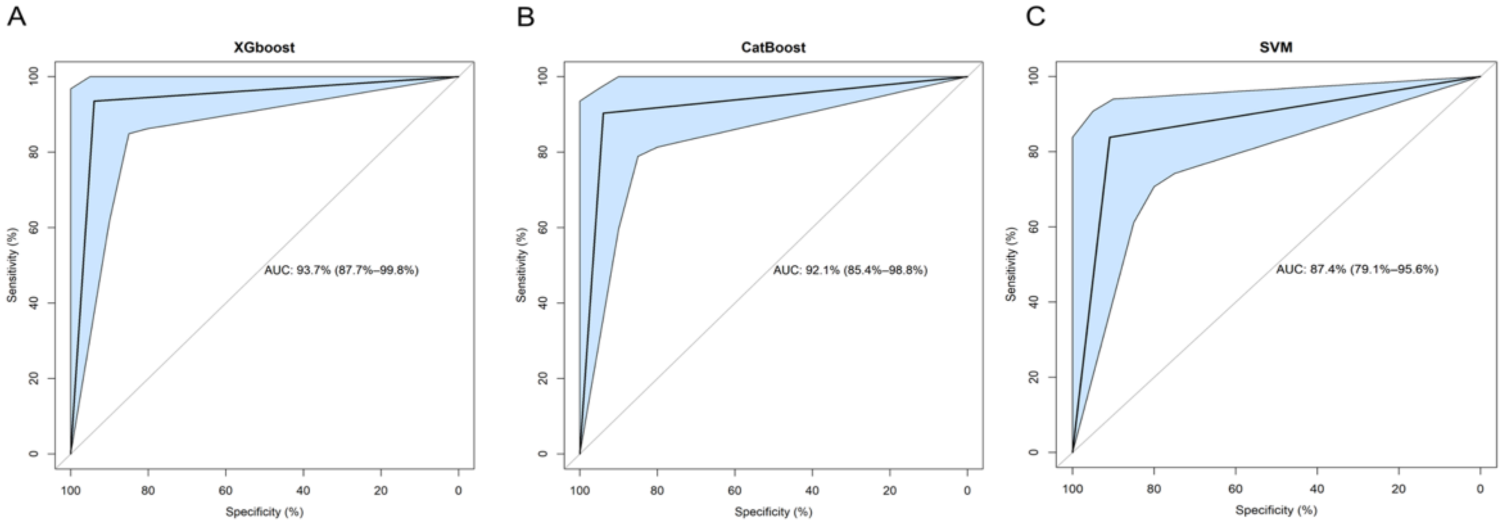


Figure 9 Receiver Operating Characteristics (ROC) curves of machine learning models.

XGboost (A), CatBoost (B) and Support Vector Machines (SVM) (C). Light blue areas represent the respective 95% confidence intervals calculated with bootstrapping. AUC: Area Under the Curve (reproduced under CC BY license from [72])

XGBoost performance with 3/38 (entropy, short-run emphasis and wavelet filtered maximum) were the most important determinants of the model's capacity to distinguish between AVN and TOH (**Figure *10***).

Table 3 Performance of the three machine learning algorithms

Performance Measure	XGB	CB	SVM
AUC (95% CI)	93.74% (87.7 - 99.8%)	92.1% (85.4 - 98.8%)	87.4% (79.1 - 95.6%)
Sensitivity	93.55%	90.32%	83.87%
Specificity	93.94%	93.94%	90.91%
PPV	93.55%	93.33%	89.66%
NPV	93.94%	91.18%	85.71%
P-value		<0.001	

AUC: Area Under the Curve; CI: Confidence interval; XGB:XGboost; CB: CatBoost; SVM: Support Vector Machines; PPV: positive predictive value; NPV: negative predictive value

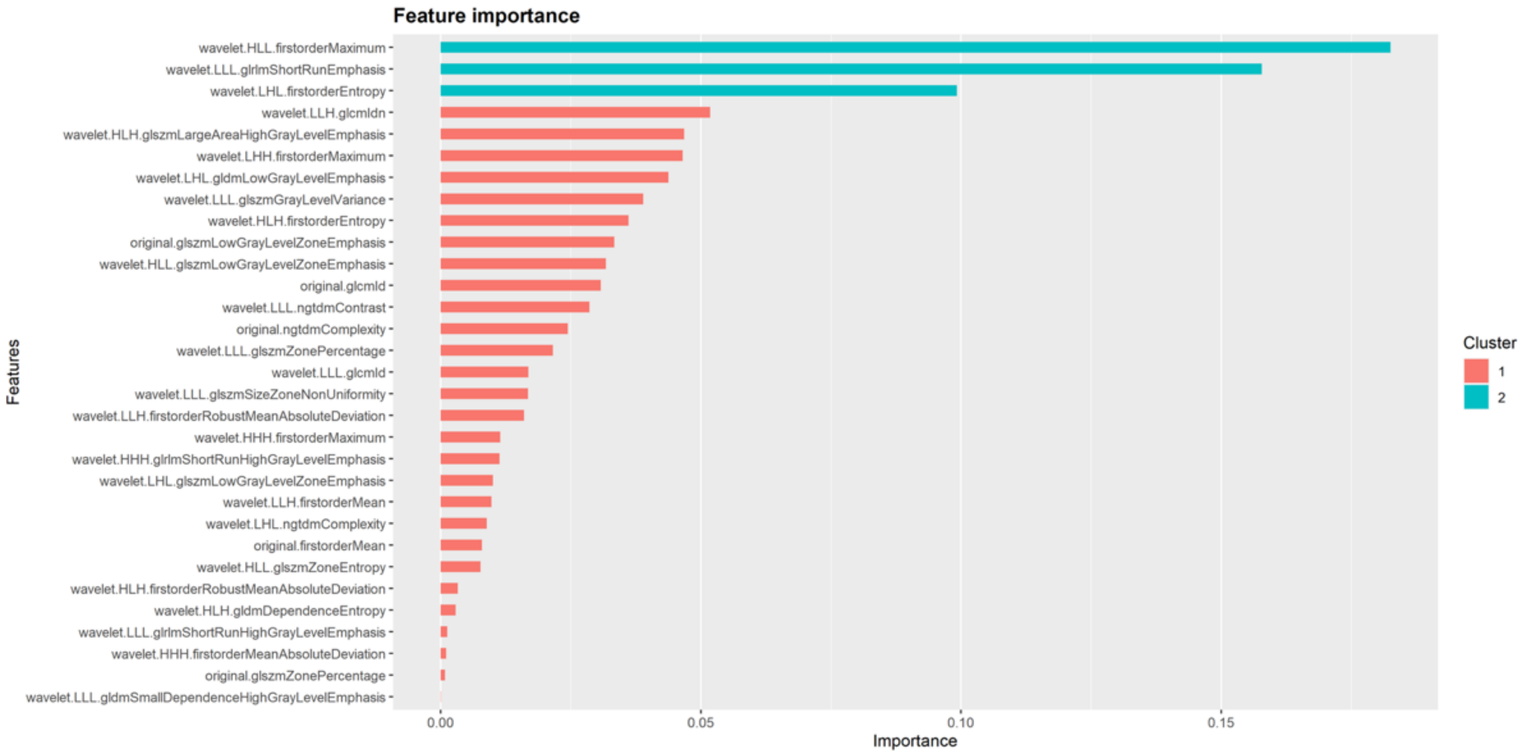


Figure 10 Radiomics features identified as important for the performance of XGboost.

Important features belong to two clusters based on their degree of importance. Cluster 2 contains three features which represent the most important determinants of XGboost performance in differentiating between TOH and AVN (reproduced under CC BY license from [72])

5.3.3 Comparison of radiomics models to expert readers

The clinical value of the developed radiomics-based method was assessed by comparing the best performing radiomics model (XGboost) to the performance of radiologists at the whole spectrum of training (residents - experts) and expertise (general radiology – MSK radiology). One of the two MSK experts (MSKR2) achieved the highest performance with an AUC of 90.6% (95% CI from 86.7% to 94.5) with a sensitivity of 89.42% and a specificity of 91.82%, compared to the second MSK radiologist (MSKR2) who achieved a slightly lower performance with an AUC of 88.3% (95% CI from 84% to 92.7%). A similar performance to fellowship-trained MSK radiologists was achieved by residents with a special interest (>6 months of training during specialty) in MSK radiology achieving AUCs of 88.9% and 87.2% for the 4th and the 5th year resident, respectively. Performance of the general radiologist was found significantly lower than the performance of XGboost (P=0.017) whereas no other reader reached a significantly different performance compared to the model (**Figure 11** and **Table 4**). Finally, in order to benchmark the model against complicated cases, XGboost was tested in cases where no consensus could be reached by expert readers, reaching an AUC of 91.7% (95% CI 75.3–100%) (**Figure 12**).

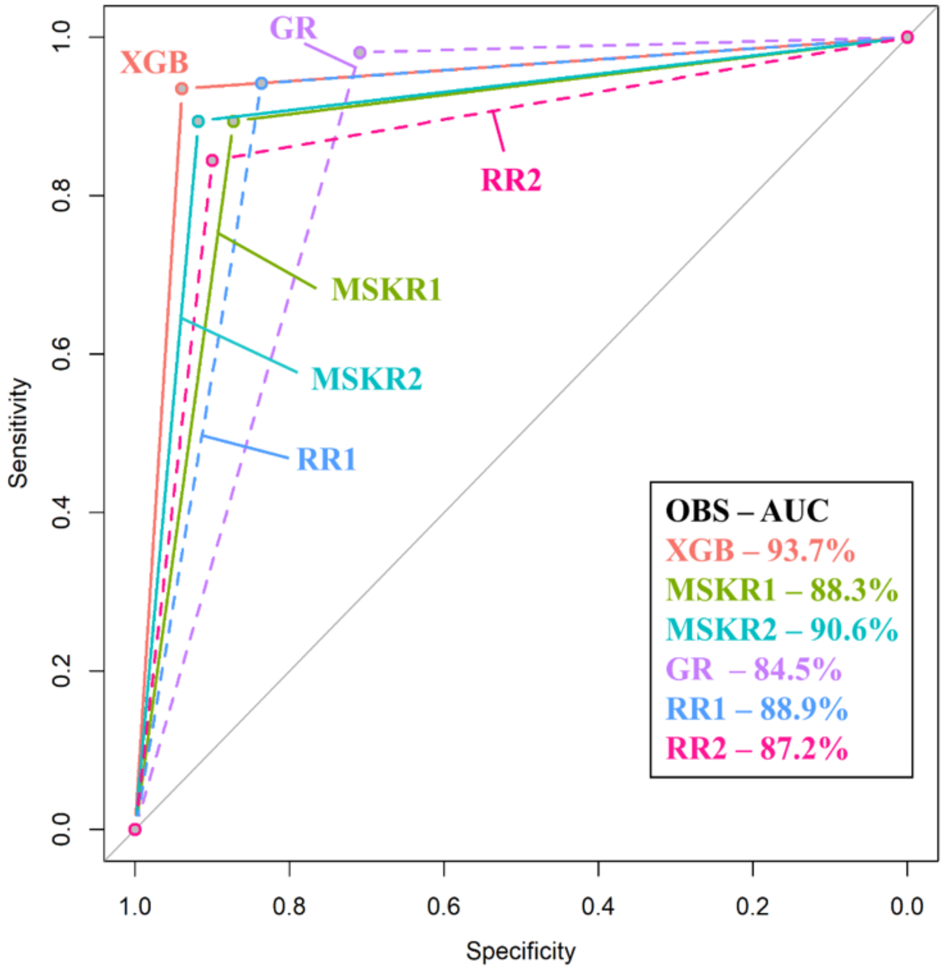


Figure 11 Comparison between Receiver Operating Characteristics (ROC) curves of XGboost and expert readers.

ROC curves of XGboost and musculoskeletal radiologists are plotted as continuous whereas the ROC curves of residents and the general radiologist are plotted as dashed lines. XGboost (pink line) is shown to have the best performance which was significantly higher than the performance of a general radiologist (GR – purple line). XGB: XGboost; MSKR: Musculoskeletal Radiologist; GR: General Radiologist; RR: Radiology Resident; OBS: Observer; AUC: Area Under the Curve (reproduced under CC

Table 4 Comparison of XGboost to expert readers

Performance Measure	XGB	MSKR1	MSKR2	GR	RR1	RR2
AUC (95% CI)	93.74% (87.7 - 99.8%)	88.3% (84 - 92.7%)	90.6% (86.7 - 94.5%)	84.5% (80 - 89%)	88.9% (84.8 - 93.1%)	87.2% (82.7 - 91.7%)
Sensitivity	93.55%	89.42%	89.42%	98.08%	94.23%	84.47%
Specificity	93.94%	87.27%	91.82%	70.91%	83.64%	90%
PPV	93.55%	86.92%	91.18%	76.12%	84.48%	88.78%
NPV	93.94%	89.72%	90.18%	97.50%	93.88%	86.09%
	P-value*	0.15	0.39	**0.017****	0.19	0.08

AUC: Area Under the Curve; CI: Confidence interval; CR:Consultant radiologist; RR: radiology resident; PPV: positive predictive value; NPV: negative predictive value; *P-value of the comparison of each reader to XGB; **: statistically significant value

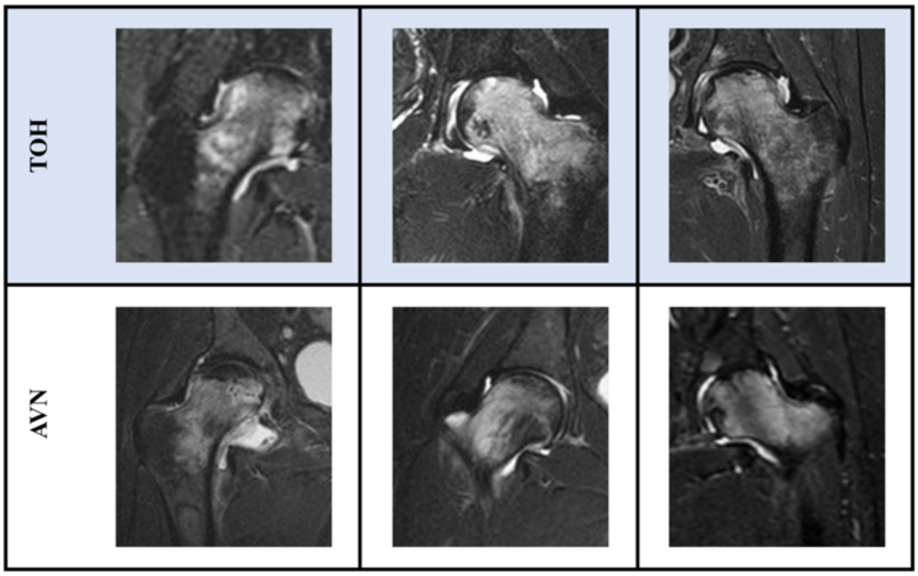


Figure 12 Examples of cases where differential diagnosis between avascular necrosis (AVN) and transient osteoporosis of the hip (TOH) can be complicated (reproduced under CC BY license from [72]).

5.4 Discussion

In this chapter I have created an MRI-based radiomics model that achieved differentiation between hips with AVN and TOH. The model reached a performance that was equal to MSK trained radiologists and radiology residents and higher that a general radiologist and achieved excellent performance when benchmarked against a subset of complicated cases.

Distinguishing between AVN and TOH is a challenging task for radiologists and several attempts have been made to distinguish between the two[43]. The diagnostic challenge raises by the fact that subchondral fractures and BME can be present in both conditions. However, as we have previously shown in a cohort of 155 hips with AVN with a follow-up between 1-10 years, subchondral insufficiency fractures that are demonstrated in almost 50% of TOH patients have a different appearance compared to the "crescent sign" in AVN [6, 43, 109] . In addition, it has been shown that bone marrow edema is present in advanced stages of AVN as a complication of articular collapse [39]. Dynamic contrast-enhanced MRI has been proposed as a way to distinguish between the two conditions but, alas, with a limited sample size that cannot sufficiently prove the value of the method for all stages of AVN [44]. For this purpose, the method presented herein achieved a performance of approximately 95% AUC, scoring similarly to MSK experts and higher that a general radiologist. This pipeline has the potential to be used as a tool to assist the diagnostic decisions of reporting radiologists with the ultimate aim to reach optimal treatment decisions and avoid unnecessary surgery for patients with TOH that could be misdiagnosed for AVN.

A series of radiomics features were found important for the predictive capacity of the presented XGboost model. The majority of these features were wavelet transformations of original features. This can potentially reflect the image appearance of characteristics specific to each one of the diseases, including the presence of the "crescent sign" of AVN, the pattern of BME, serpiginous AVN changes and the low signal irregular lines of subchondral fractures seen in TOH [14]. These characteristics consist of edges and irregular lines which are significantly highlighted when an image undergoes wavelet transformation at distinct scales [63, 64]. Another important advantage of wavelet transformation is the elimination of noise profile inconsistencies that may arise between different MR scanners of our datasets. These inconsistencies have been greatly addressed by image equalization prior to radiomics analysis, nonetheless, wavelet transformations offer an additional filtration layer to address potential remaining inconsistencies.

The work of this chapter has strengths and limitations. Strengths of this work include the relatively large number of hips examined in our dataset, given the low prevalence of the two diseases, especially TOH. Additionally, the use of a multi-vendor dataset increases the value and generalizability of our results since the models have been exposed to data from multiple scanners. Another strength of our approach is the demonstration of the diagnostic capacity of the proposed model compared to a series of radiologists at all levels of training and expertise. Nonetheless, this work carries certain limitations. First of all, the retrospective nature of the study, which was inevitable due to the low prevalence of the studied disease states. Manual segmentation could also present a potential limitation of our work. However, variability in the segmentation of structures (femoral head/neck) surrounded by thick cortical bone with clear boundaries significantly reduces the possibility of segmentation errors. Finally, use of only coronal STIR images for the extraction of radiomics data could be a limitation of our approach since in everyday clinical practice a series of sequences are available to the reporting radiologists. Nonetheless, fluid-sensitive sequences (STIR, fs PD/T2-w) are sufficient in everyday diagnostic practice, when a radiologist faces the dilemma to distinguish between TOH and AVN, since such sequences are the gold standard for the depiction of BME which is a common characteristic of both disease entities.

Chapter 6 – Deep learning for the differentiation between avascular necrosis and transient osteoporosis of the hip

Results presented in this chapter have been published in Klontzas et al. Diagnostics, 2022 [110].

6.1 Introduction

As demonstrated in chapter 5, radiomics analysis allows the accurate differentiation between TOH and AVN. Nonetheless, the process is limited by the need for accurate segmentation of the region of interest, and the presence of the additional time consuming and computationally expensive steps of radiomics feature extraction, selection and dataset curation for machine learning.

The purpose of the work presented in this chapter was to overcome potential bottlenecks in the pipeline of chapter 5. For this purpose, deep learning was employed, utilizing whole images, skipping the manual steps of radiomics analysis and automating the differentiation between AVN and TOH. Three well-established convolutional neural network (CNN) architectures were trained with a transfer learning methodology and finetuned with the dataset of chapter 5. Subsequently, an ensemble consensus decision was recorded by combining the results of all three CNNs and the ensemble decision was compared to the decision of an MSK radiologist and an MSK imaging fellow.

6.2 Chapter-specific Methodology

6.2.1 Data preparation and deep learning model training and testing

Image preparation for deep learning was performed as described in chapter 4. In order to account for the limited sample size and group imbalance, the dataset was augmented to a total of 420 images (210 TOH and 210 AVN) as suggested by Candemir et al[82] (see chapter 4 for details on data augmentation).

Transfer learning was performed with the use of three well-established CNN architectures, VGG-16, Inception V3 and Inception-ResNetV2. The initial weights of these networks were obtained from the ImageNet dataset with >14 million images and then the final trainable layers were finetuned with 70% of our data. The rest (30%) were used for validation of the networks. Adopting a transfer learning methodology is an important strategy to tackle data-limited scenarios for deep learning model training [82]. Networks were trained for a maximum of 50 epochs with an early stopping function in order to avoid overfitting. Visual assessment of training/validation accuracy and training/validation loss plots was used as an extra measure to avoid overfitting. The decision of each individual network was subsequently recorded and a consensus CNN ensemble decision was reached as the decision reached by at least two out of three networks (**Figure** *13*).

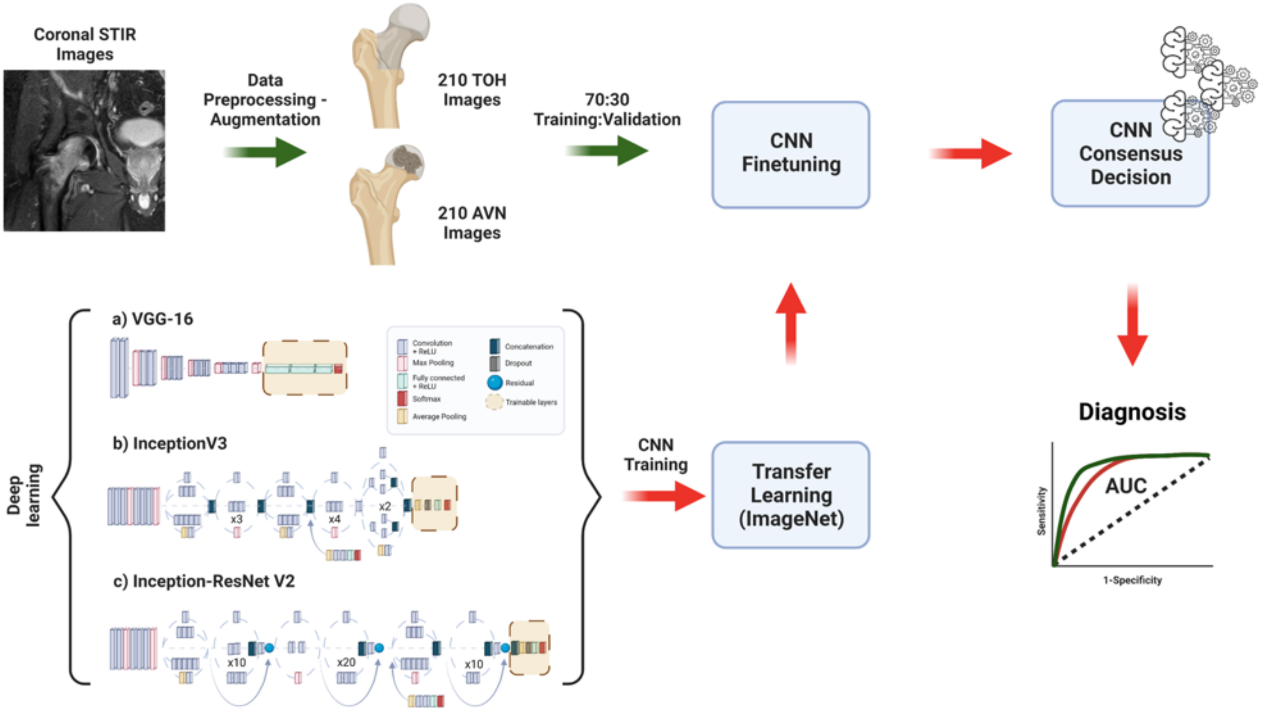


Figure 13 Flow diagram describing methodology followed for data augmentation, deep learning model training with transfer learning, and the development of a model ensemble for the diagnosis of TOH vs. AVN (created with BioRender.com and reproduced under CC BY license from [110]).

6.2.2 Comparison of CNN ensemble to expert readers

CNN ensemble performance on the validation set was benchmarked against the diagnosis of one MSK radiologist with 7 years of experience and one MSK radiology fellow, who have both work in our specialized center for bone marrow imaging.

6.3 Results

6.3.1 Deep learning model training and validation

All three CNNs achieved an AUC between 96% and 97.6% as demonstrated in . Even though, Inception-ResNetV2 achieved the same performance compared to the model ensemble the ensemble achieved a 0% type I error for the detection of TOH, a 100% precision for the detection of AVN and 100% recall for the detection of TOH. Interestingly, the type I error of the ensemble (AVNs diagnosed as TOH) was lower than all individual CNNs. Interestingly, the majority of TOH cases were correctly identified as TOH with no mistakes for the ensemble, VGG-16 and Inception-ResNetV2 and only one mistake (false negative) by InceptionV3 (**Figure** *14* and **Table 5**. Performance metrics of individual convolutional neural networks and the respective network ensemble.) .

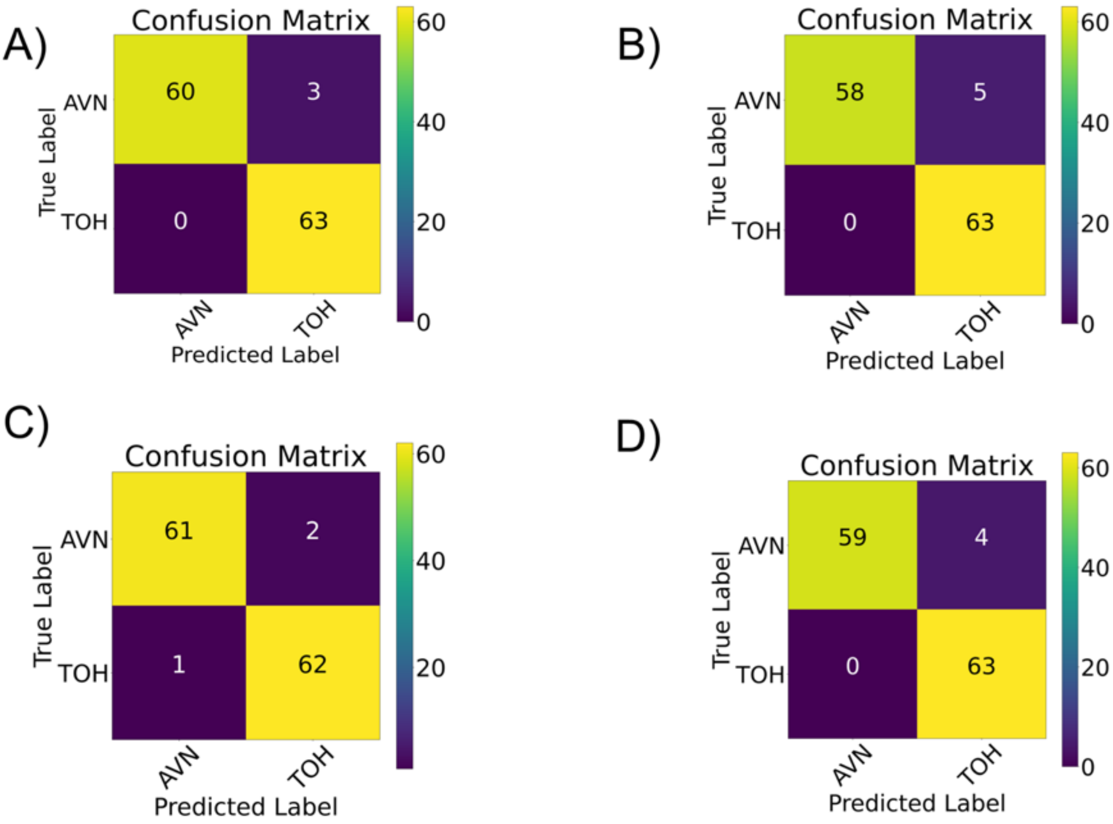


Figure 14 Confusion matrices for the individual and ensemble CNN decisions
Matrices represent CNN ensemble (A), VGG-16 (B), InceptionV3 (C) and Inception-ResNetV2 (D). TOH: transient osteoporosis of the hip; AVN: avascular necrosis of the femoral head.

Table 5. Performance metrics of individual convolutional neural networks and the respective network ensemble.

	AUC	Group	Precision	Recall	f1-score
Model ensemble	**97.62%**	AVN	1	0.95	0.98
		TOH	0.95	1	0.98
VGG-16	**96.03%**	AVN	1	0.92	0.96
		TOH	0.93	1	0.96
InceptionV3	**96.82%**	AVN	1	0.94	0.97
		TOH	0.94	1	0.97
Inception-ResNet-V2	**97.62%**	AVN	0.98	0.97	0.98
		TOH	0.97	0.98	0.98

AUC: Area Under the Curve; AVN: Avascular Necrosis; TOH: Transient Osteoporosis of the Hip

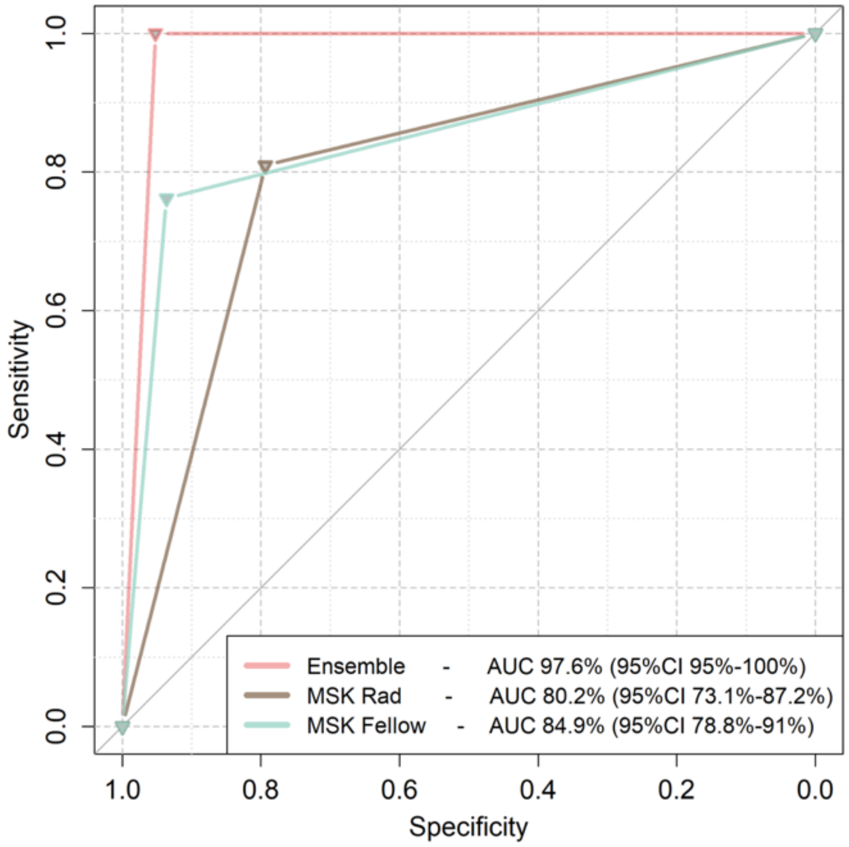


Figure 15 Receiver Operating Characteristics (ROC) curves of the model ensemble and MSK imaging experts

Model ensemble curve is plotted as a pink line, the MSK radiologist curve is plotted with brown color and the MSK fellow curve is plotted with a turquoise color. MSK Rad: Musculoskeletal

6.3.2 Comparison between CNN ensemble and expert readers

In order to further benchmark the CNN ensemble, its diagnostic performance was compared to the diagnosis of expert readers who were presented with the images of the validation dataset. The model ensemble achieved a higher performance than both experts ($P<0.001$) reaching an AUC of 97.6% (95%CI from 95% to 100%), compared to 80.2% (95%CI from 73.1 to 87.2%) and 84.9% (95%CI from 78.8% to 91%) for the MSK radiologist and fellow respectively (**Figure 15**).

6.4 Discussion

As discussed extensively in chapter 5, differentiation between TOH and AVN is of utmost importance for the selection of the appropriate treatment plan and the decision between conservative (TOH) and surgical treatment (AVN). In this chapter, I have introduced a deep learning methodology that overcomes potential shortcomings of radiomics analysis and provides automated differentiation between the two disease entities using a multivendor/multi-institutional dataset. The CNN ensemble presented in this chapter achieved excellent predictive capacity which was found significantly higher than expert MSK radiologists.

The work presented in chapter 5 set the basis for the comprehensive analysis of femoral head images with radiomics that provides a highly accurate distinction between TOH and AVN [72]. Nonetheless the CNN ensemble reached an AUC of 98% which was higher than expert readers, which was not the case with the XGboost radiomics model which achieved a maximum AUC of approximately 94% which was not significantly higher than expert performance. Interestingly enough, reader performance in this work was found between 80.2% and 84.9% which is poorer than the performance of readers in the chapter 5. This can be attributed first to the differences in reader experience, as well as to intrinsic dataset differences arising from the augmentation process.

CNN performance (both ensemble and individual CNNs) is comparable to CNN performance with other medical imaging datasets. Examples include X-ray datasets where voting CNN ensembles were used for the detection of tuberculosis [111]. The authors created an ensemble of Inception and Xception models and employed several augmentation/preprocessing steps. A similar voting strategy has been used by Tsiknakis et al. [112] who have combined Xception, InceptionV3, ResNet and Inception-ResNet architectures to perform an image classification task that included twenty classes of images. Their ensemble reached an AUC of almost 100% demonstrating the value of voting ensembles in solving classification problems. Importantly, even though similar AUCs can be reached with the use of single CNNs, ensemble voting strategies offer robust results minimizing type I and type II errors.

This work has certain strengths and weaknesses. One of the most important strengths is the use of a multi-vendor dataset for the training and validation of CNNs. It has been establish that between-scanner batch effects can limit the applicability of models trained with data of only one scanner. Therefore, using data from multiple scanners has been suggested as a strategy to increase the generalizability of CNN predictions [113]. Another strength of our approach is the comparison to MSK experts that demonstrates the clinical value of our models. With regards to the limitations of our approach, these are similar to the limitations described in chapter 5.

Chapter 7 – Deep learning for the differentiation between subchondral insufficiency fractures of the knee and advanced osteoporosis

Results presented in this chapter have been published in Klontzas et al. Injury, 2022 [114].

7.1 Introduction

One of the most common diagnostic dilemmas arising when reporting knee MRI is the differentiation between bone marrow edema (BME) (also called bone marrow lesions – BML) in the context of subchondral insufficiency fractures (SIF) and BME in the context of osteoarthritis (OA) [1, 2, 9, 11, 115, 116]. On the one hand, SIF was previously known as spontaneous osteonecrosis of the knee, but is currently widely accepted that no necrosis is involved in the pathophysiology of the disease. Indeed, SIF is thought to result from micro-trabecular insufficiency fractures that are associated with extensive BME extending from the subchondral zone of the bone deep inside the bone marrow [12, 117, 118]. On the other hand, OA is characterised by cartilage damage and BME is encountered late in the course of the disease and is thought to be correlated with the presence of pain and progressive disease [119, 120]. BME in OA appears in a limited bone marrow area characteristically at outer joint sites, especially where cartilage loss is noted [119, 120]. Nonetheless, differentiation between BME related to SIF and BME related to OA can be extremely complicated, especially when no low-signal fracture line is seen in SIF [117] or when SIF and OA coexist. Differentiation between SIF and OA is extremely important, since it defines the treatment plan which is conservative in cases of SIF (reduced weight-bearing) or surgical (total knee arthroplasty) in cases of advanced OA.

The aim of the work presented in this chapter was to use transfer learning to train a convolutional neural network (CNN) ensemble with a consensus voting strategy as presented in chapter 6, that would differentiate SIF from OA. The performance of the proposed strategy was compared to the performance of MSK radiologists. Such a deep learning methodology would be a valuable tool to assist complicated diagnoses by inexperienced and experienced readers.

7.2 Chapter-specific Methodology

7.2.1 Patients – MRI diagnosis

Patient recruitment has been described in detail in chapter 4. A total of 212 knees with SIF and 102 knees with advanced OA were included in the study.

All MRI examinations were blindly evaluated by a senior MSK radiologist with 40 years of experience, and two MSK radiologists with 5 and 7 years of experience in musculoskeletal imaging respectively. Ground truth was established in cases with agreement of all three radiologists. In cases with conflict, final ground truth diagnosis was considered the one with the agreement of at least two experts.

For the purposes of ground truth establishment, all examinations included at least (1) PD-w fat-suppressed sequences at three planes, (2) one coronal T1-w sequence and (3) a sagittal gradient echo sequence. SIF ground truth diagnosis was established using a combination of imaging and clinical data including: (a) the identification of BME extending from the subchondral area to the epiphysis/metaphysis, (b) identification of focal thickening of the subchondral bone or low-signal intensity area in the immediate subarticular bone, (c) potential identification of low signal intensity irregular lines within areas of BME at variable depths deep to the articular surface (indicating subchondral fracture lines), (d) depiction of a fluid-filled cleft in the subchondral bone (depiction of c and d is not necessary for the diagnosis). Clinical indications of SIF included the acute onset of pain and complete symptom resolution only with reduced weightbearing [12]. Clinical indications for the diagnosis of OA on were based on the recommendations of European League Against Rheumatism (EULAR) including: (a) pain linked to activity, limited morning stiffness and joint functional impairment, (b) enlargement of the joint, (c) movement restriction and (d) crepitus. Image-based (MRI) diagnosis of OA included the following findings: (a) the presence of osteophytes, (b) cartilage damage, (c) bone marrow edema, (d) degenerative meniscal tears/subluxation [121] (**Figure *16***).

CNNs were trained with the use of mid-coronal PD-w fat suppressed images, which represent fluid sensitive images and are considered to be the gold standard for the depiction of BME.

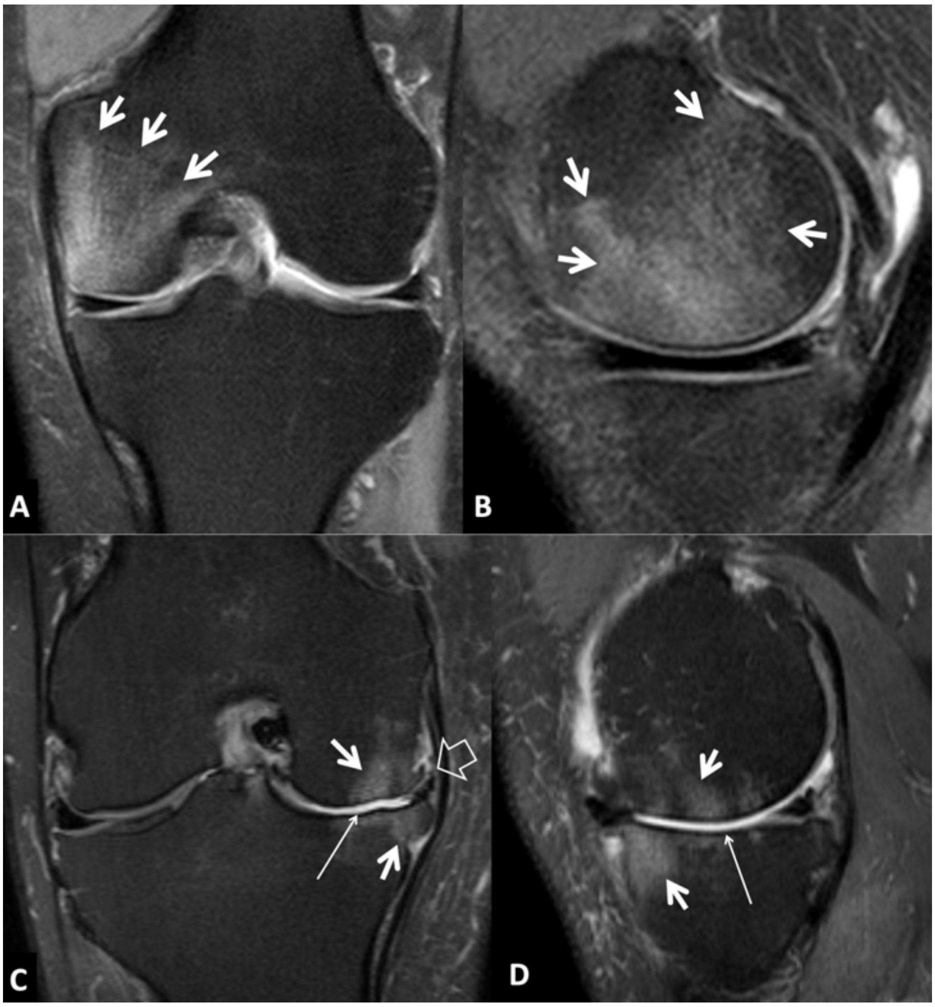


Figure 16 Imaging features of typical subchondral insufficiency fracture (A,B) and osteoarthritis (C,D) cases.

Coronal (A) and sagittal (B) fat-suppressed intermediate weighted MR images of a 53-year-old male patient with a history of 1-month pain without any injury, demonstrate bone marrow edema (arrows) in keeping with the presence of subchondral insufficiency trabecular microfractures. The articular cartilage is intact. In comparison, coronal (C) and sagittal (D) fat-suppressed intermediate weighted MR images of a 62-year-old female patient with a history of 9-month medial compartment pain, show bone marrow edema (arrows), articular cartilage erosion in the medial femoral and tibial condyles (thin arrows) in keeping with the presence of degenerative osteoarthritis. There is also meniscal degeneration and a marginal osteophyte (open arrow).

7.2.2 Data preparation and CNN ensemble development

Data preparation for deep learning was performed as described in chapter 4. In order to tackle the limited number of examinations, data augmentation was performed with horizontal flipping and clockwise/anti-clockwise rotation (10°) reaching a total of 500 SIF and 500 OA images for training and 87 images in each group for validation of the algorithm (85%/15% data split) (**Figure** *17*).

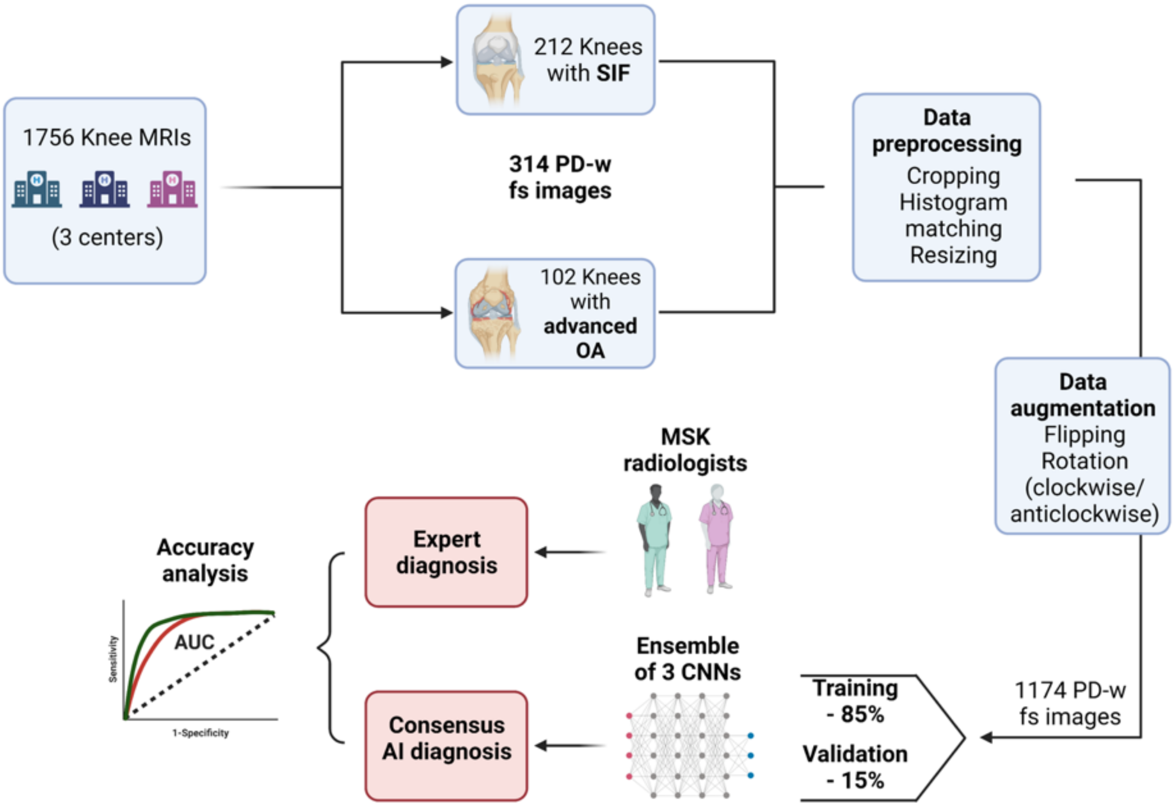


Figure 17 Schematic describing the convolutional neural network architectures used in this study and the transfer learning process used for the development of the deep learning ensemble.
ReLU: Rectified Linear Unit; PD-w fs: proton density weighted with fat suppression; OA: osteoarthritis; SIF: subchondral insufficiency fractures (created with BioRender.com) (reproduced with permission from Klontzas et al Injury 2022 [114]).

The same CNN architecture ensemble (VGG-16, InceptionV3 and Inception-ResNetV2) as the one presented in chapter 6 was trained with transfer learning and finetuned with our data and predictions were derived for each of the individual models as well as for their consensus decision (**Figure** *18*). Images of the validation dataset were also evaluated by two MSK radiologists to compare the performance of the CNN ensemble to real life diagnostic scenarios.

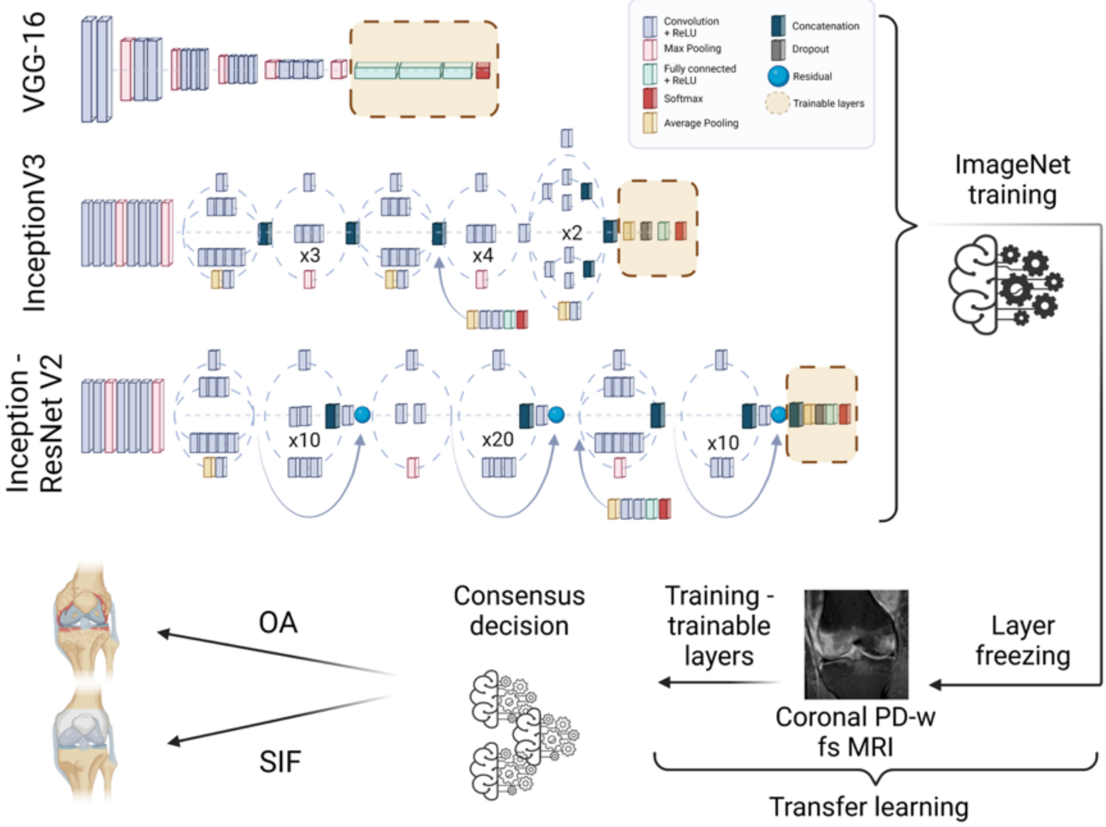


Figure 18 Schematic describing the convolutional neural network architectures used in this study and the transfer learning process used for the development of the deep learning ensemble.
ReLU: Rectified Linear Unit; PD-w fs: proton density weighted with fat suppression; OA: osteoarthritis; SIF: subchondral insufficiency fractures (created with BioRender.com) (Reproduced with permission from Klontzas et al. Injury 2022 [114]).

7.3 Results

7.3.1 Patient characteristics

Patients included in this study had a mean age of 62.67 ± 10.8 years for the SIF group and 63.87 ± 11.84 years for the OA group. A total of 60 female and 42 male patients with OA were included in the study compared to a total of 93 male and 119 female patients with SIF.

7.3.2 Performance of individual CNNs and the CNN ensemble and comparison to radiologists

The highest performance among individual CNNs was reached by InceptionV3 that achieved a total AUC of 93.68%, compared to VGG-16 which achieved the smallest AUC of 82.18%. Inception-ResNetV2 demonstrated a slightly worse precision compared to InceptionV3, reaching however a similar AUC (92.53%). The performance of the CNN ensemble was the highest compared to each individual CNN (with an AUC of 95.97%). Detailed performance metrics can be found in **Table 8** and **Figure *19***.

In order to benchmark the performance of the ensemble against expert performance, images of the validation set were blindly evaluated by two MSK radiologists . One of the two experts scored lower (AUC of 82.76%) than the CNN ensemble (P<0.001) whereas the other achieved a score (AUC 91.95%) that did not differ significantly compared to the consensus ensemble decision (P>0.05) (**Table 7**, **Figure *21*** and **Figure *20***).

Table 6 Performance metrics of individual models and the neural network ensemble.

	AUC	Group	Precision	Recall	f1-score
Model ensemble	**95.97%**	SIF	0.98	0.94	0.96
		OA	0.94	0.98	0.96
VGG-16	**82.18%**	SIF	0.89	0.91	0.84
		OA	0.77	0.74	0.81
InceptionV3	**93.68%**	SIF	0.94	0.93	0.94
		OA	0.93	0.94	0.94
Inception-ResNet-V2	**92.53%**	SIF	0.90	0.95	0.93
		OA	0.95	0.90	0.92

AUC: Area Under the Curve; SIF: Subchondral Insufficiency Fractures; OA: Osteoarthritis

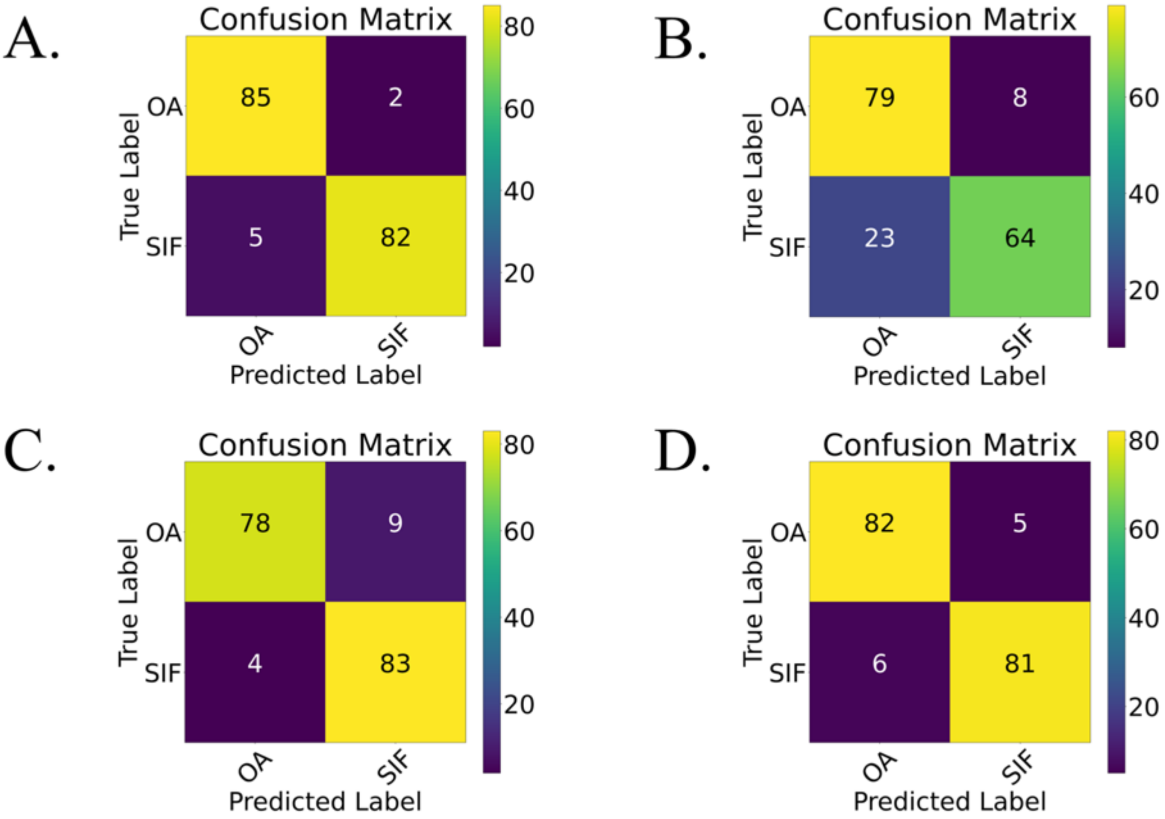


Figure 19 Confusion matrices from the validation of the neural network ensemble (A) and each individual neural network (B – VGG-16, C – Inception-ResNetV2, D - InceptionV3)
OA: osteoarthritis; SIF: subchondral insufficiency fractures (reproduced with permission from Klontzas et al. Injury 2022 [114]).

Table 7. Performance metrics of expert readers compared to the model ensemble.

	AUC	**Group**	**Precision**	**Recall**	**f1-score**
Model ensemble	**95.97% (93.1% - 98.9%)**	SIF	0.98	0.94	0.96
		OA	0.94	0.98	0.96
MSKR1	**91.95% (87.9% - 96%)**	SIF	0.92	0.92	0.92
		OA	0.92	0.92	0.92
MSKR2	**82.76% (77.5% - 88%)**	SIF	0.76	0.97	0.85
		OA	0.95	0.69	0.80

AUC: Area Under the Curve; SIF: Subchondral Insufficiency Fractures; OA: Osteoarthritis

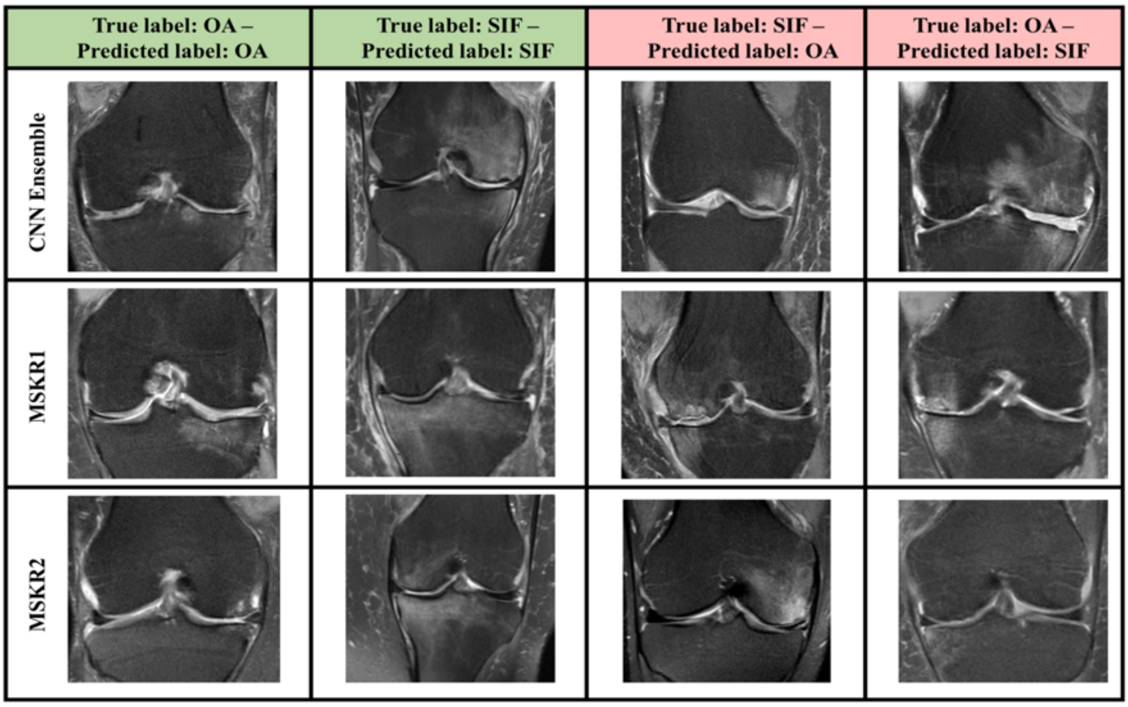

True label: OA – Predicted label: OA
True label: SIF – Predicted label: SIF
True label: SIF – Predicted label: OA
True label: OA – Predicted label: SIF
CNN Ensemble
MSKR1
MSKR2

Figure 21 Performance of the CNN ensemble compared to MSK radiologists. (A) Examples of correctly classified (green) and misclassified (red) cases of subchondral insufficiency fractures and OA by the convolutional neural network (CNN) ensemble and each one of the two musculoskeletal radiologists (MSKR1 & MSKR2).

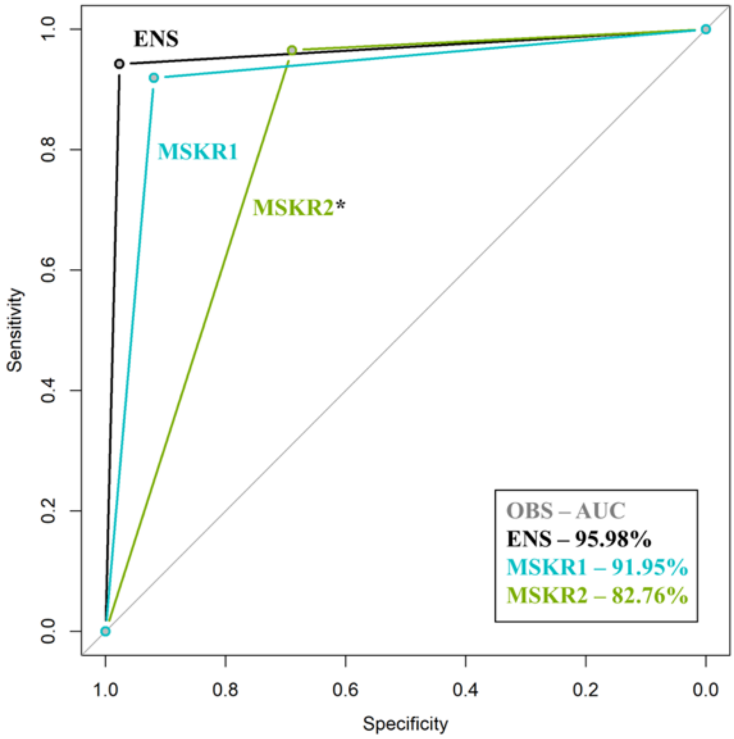


Figure 20 Comparison between Receiver Operating Characteristics (ROC) curves of the CNN ensemble and expert readers

CNN ensemble ROC curve (black line) is shown to have significantly higher performance than the one of the MSK radiologists (MSKR2 – green line) and equal to the other (MSKR1 – blue line). MSKR: Musculoskeletal Radiologist; OBS: Observer; AUC: Area Under the Curve. *: P<0.05 for the comparison to the performance of the ensemble.

7.4 Discussion

In this chapter, I have developed a CNN ensemble that accurately differentiates between BME related to SIF compared to BME related to advanced OA. The proposed deep learning methodology achieved a performance that was equal to or higher than expert MSK radiologists.

SIF was traditionally considered to be related to necrosis, being previously known as “spontaneous osteonecrosis of the knee”. Nonetheless it has been nowadays established that necrosis is not a feature in the pathophysiology of the disease [1, 29], which occurs as a result of micro-trabecular insufficiency fractures of the knee accompanied by abundant BME that is nicely depicted by fluid-sensitive sequences on MRI [117]. The importance of our method lies in the fact that it can become an important tool assisting the differential diagnosis between SIF and advanced OA that presents with BME. This can be extremely important in cases where: (1) subchondral fracture lines are not evident in SIF (42% of patients) [117], (2) unicompartmental osteoarthritis may be complicated by SIF, (3) subchondral sclerosis due to osteoarthritis may mimic the appearance of subchondral fractures seen in SIF [12, 117, 118, 122]. In these cases, reaching the correct diagnosis can determine the need for surgical treatment (total knee arthroplasty) which is the treatment of choice for advanced knee OA.

Our work has certain strengths and limitations. Strengths include the training of our algorithms on a multi-institutional/multivendor dataset that can increase the generalizability of our results and the large number of SIF patients which represents the largest SIF cohort presented in literature to date (to the best of our knowledge).

Limitations of our approach may be the lack of histological confirmation of SIF which, however, is neither feasible not ethically justified in clinical practice. The lack of an external validation dataset could be another limitation of our results which is, to a certain extent, alleviated by the use of a multi-institutional MRI cohort. Finally, the fact that cases where SIF and OA co-existed were excluded from our dataset could be a limitation of our study. However, in such cases where advanced OA is a prominent feature treatment will be surgical irrespective of the co-existence of SIF.

Chapter 8 – Deep learning for the staging of avascular necrosis of the hip

Results presented in this chapter are currently under review in "European Radiology". Part of the work of this chapter has become feasible by funding from the Young Researcher's grant awarded to Dr Klontzas by the European Society of Musculoskeletal Radiology.

8.1 Introduction

In cases where AVN is left untreated, it progresses to joint collapse and secondary osteoarthritis with THA being the only treatment option. However, in early stages of the disease prior to articular collapse, joint preservation techniques (core decompression, vascularized grafting, etc.) are available with the potential to avoid THA [34]. Therefore, differentiation between early and late AVN is of utmost importance for appropriate treatment selection.

A variety of AVN staging systems exist, with the system of the Association Research Circulation Osseous (ARCO) [36] being the most commonly used and the one recommended in the latest (2019) international guidelines on the management of AVN [32]. The latest version of ARCO defines four main stages, with joint preservation techniques being available for the two first stages (ARCO < 3) whereas hip replacement being the recommended treatment for terminal disease (ARCO 3-4). Nonetheless, distinguishing between ARCO 2 (early) and ARCO 3A (late) is an extremely challenging task, requiring significant expertise in musculoskeletal radiology and a combination of imaging findings including indications of loss of femoral head sphericity and the presence of a subchondral fracture[32, 123, 124].

Artificial intelligence has been previously used for the diagnosis of AVN on plain radiographs [125] and MRI [126], to identify factors increasing the risk for collapse[127], and to differentiate late AVN from other causes of proximal femoral bone marrow edema such as transient osteoporosis [72, 110]. Attempts have been also recently made to quantify the necrotic volume and surface area in an attempt to associate this with the stage of AVN [15]. However, quantification of the necrotic part volume is not part of any clinically relevant classification system and volume cut-offs have not been set to levels that will define optimal treatment.

The aim of this chapter was to develop a deep learning methodology to differentiate between early (ARCO 1 & 2) and late (ARCO 3 & 4) stages of AVN. For this purpose, convolutional neural networks (CNNs) have been trained with a transfer learning methodology and finetuned with the use of a cohort of patients with AVN. The algorithm was internally tested and then subjected to external validation on an international cohort of AVN patients and its performance was benchmarked against the performance of musculoskeletal radiologists. Development of such a model would be invaluable in assisting clinical decisions between joint preservation surgery and total hip arthroplasty.

8.2 Chapter-specific Methodology

8.2.1 Patient characteristics

For the purposes of this chapter two cohorts of patients were used. The first cohort (for the rest of the chapter will be called the University Hospital of Heraklion – UHH cohort) consisted of all hips with AVN (104 hips – 67 patients) that were used for the work presented in chapters 5 and 6. A combination of transfer learning and data augmentation were used to address the small size of the dataset as proposed by Candemir et al. [82](described below).

The second cohort was used for the external validation of the developed deep learning methodology. This was an independent anonymised cohort from a centre located in another country (Technical University of Munich – for the rest of the text will be called the TUM cohort, n=49 hips) which were retrospectively selected based on the same criteria (**Figure 22**).

8.2.2 Ground truth ARCO staging

Ground truth staging of AVN was established based on imaging, according to the ARCO classification [3]. This is currently the gold standard practice for clinical diagnosis and staging, since the diagnosis of AVN does not warrant biopsy. Ground truth staging was performed independently by two MSK radiologists (40 and 10 years of experience, respectively) and in cases of disagreement final stage was defined by consensus. To ensure accurate ground truth grading the experts had access to the whole MRI protocol including T1-w, STIR or PD/T2 fs and high-resolution 3D

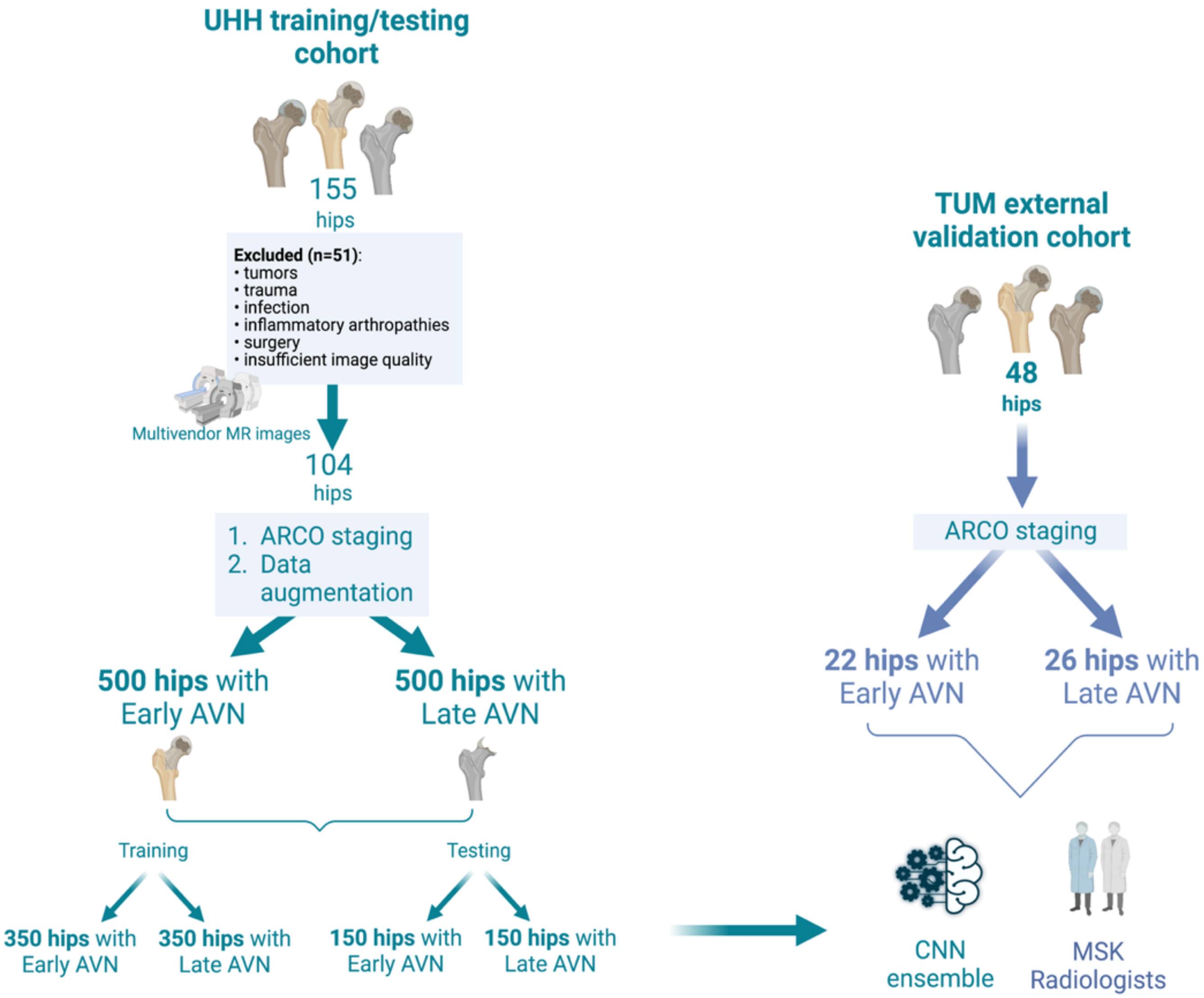


Figure 22 Flow diagram explaining the characteristics of the UHH and TUM cohorts for training/testing and external validation respectively. AVN: Avascular necrosis of the hip; CNN: Convolutional Neural Networks; MSK: Musculoskeletal; ARCO: Association Research Circulation Osseous (created with biorender.com).

gradient echo images. AVN was diagnosed by the presence of the "band-like" sign on T1-w images [16]. Subsequently, two groups of hips were defined based on T1-w, STIR and high-resolution 3D gradient echo images: (i) cases with a subchondral fracture, cases with loss of head sphericity and/or associated bone marrow edema and cases with signs of secondary osteoarthritis were classified as "late AVN" (ARCO 3-4) (ii) cases without the aforementioned findings were classified as "early AVN" (ARCO 1-2) (**Figure 23**) [1, 17, 18].

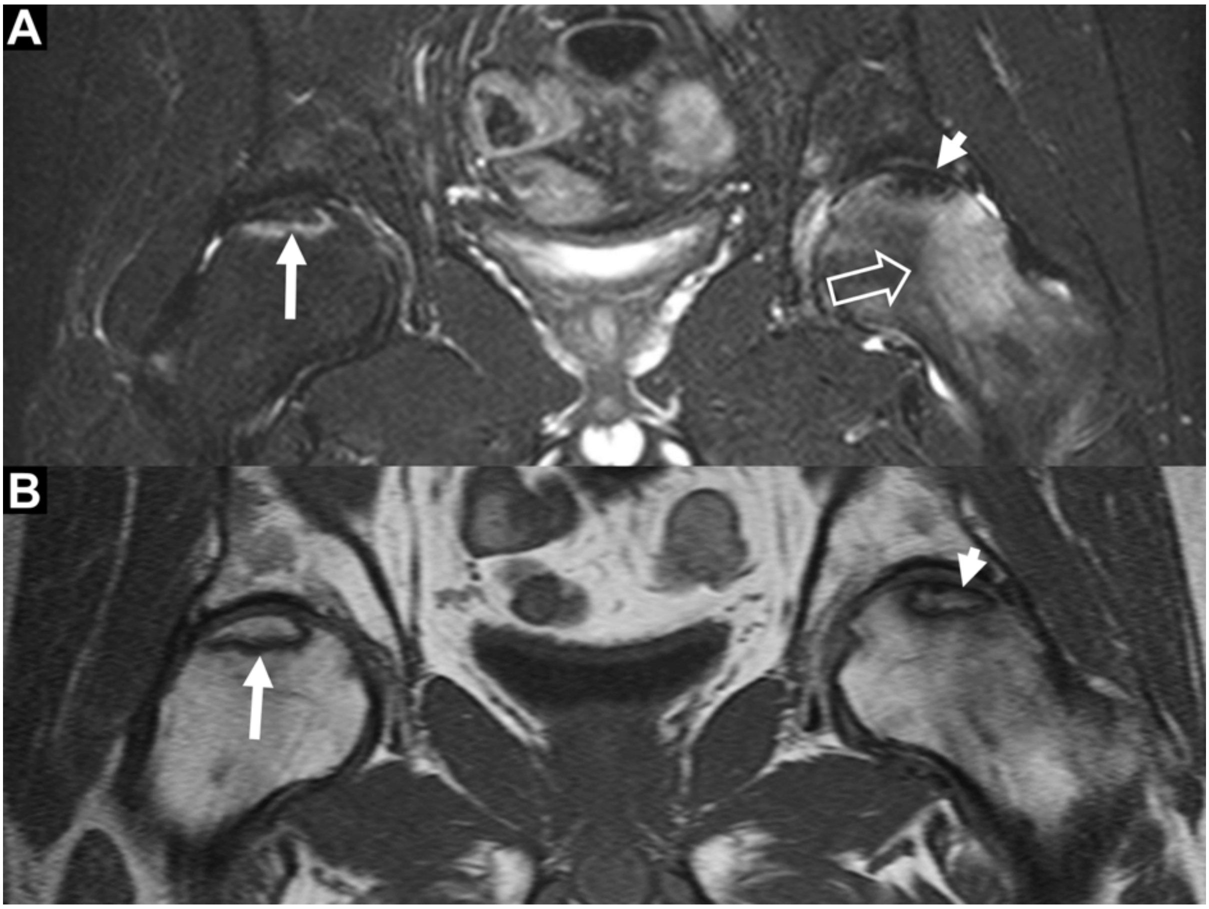


Figure 23 Coronal STIR (A) and T1-w (B) MR images, showing bilateral idiopathic avascular necrosis, in a 43-year-old male presenting with a left painful hip.

The lesion on the right hip (arrows), is asymptomatic and is occult on plain radiographs (not shown). The lesion on the left, is associated with bone marrow edema (open arrow) secondary to mild articular surface flattening (short arrows). According to ARCO classification, the right lesion is stage I and the left stage III.

8.2.3 Data pre-processing and augmentation

Mid-coronal STIR images through each femoral head were used for model training, testing and validation. Images were resized to 150 x 150 pixels and then images were randomly split 70:30 in training: testing sets. Data harmonization and bias correction was performed by matching image histograms to account for intra-scanner variability and achieve grey level normalization. In order to eliminate group imbalance bias and to expose the model to additional training/testing data, images were augmented using rotation of 10o (clockwise and anti-clockwise) as well as horizontal image flipping. The final training and testing datasets consisted of a total of 350 training and 150 testing images for each of the two groups (early vs late AVN) (**Figure 22**).

8.2.4 Convolutional Neural Network Ensemble Training and External Validation

A CNN ensemble was used as described in Chapter 4. Briefly, transfer learning was applied by obtaining the initial weights of three individual CNN architectures, VGG-16, InceptionV3 and Inception-ResNetV2, training first with the ImageNet dataset followed by weight freezing and final trainable layer finetuning with the use of our training dataset [128]. Network performance was subsequently evaluated with the use of the UHH testing dataset. A consensus ensemble decision of the three CNNs was recorded as the agreement of at least two out of three CNNs. To further benchmark the performance of the CNN ensemble, the resulting model was externally validated on a set of 49 hips from a radiology department of another country (TUM dataset). Images were resized and used without any further pre-processing. Ground truth for the TUM dataset was established with the same method as for the UHH dataset External validation images were also assessed by two experienced MSK radiologists (10 and 7 years of MSK experience) blinded to the results of the ensemble and their performance was compared to the performance of the ensemble.

8.3 Results

8.3.1 Individual and ensemble CNN performance

Each CNN architecture was initially subjected to internal testing with the UHH cohort where Inception-ResnetV2 achieved the highest individual performance with an AUC of 99.7% (95%CI 99-100%), followed by InceptionV3 and VGG-16 with AUCs of 99.3% (95%CI 98.4-100%) and 97.3% (95%CI 95.5-99.2%) respectively. VGG-16 had the highest number of misclassified cases with three early cases misclassified as late and five late cases misclassified as early.

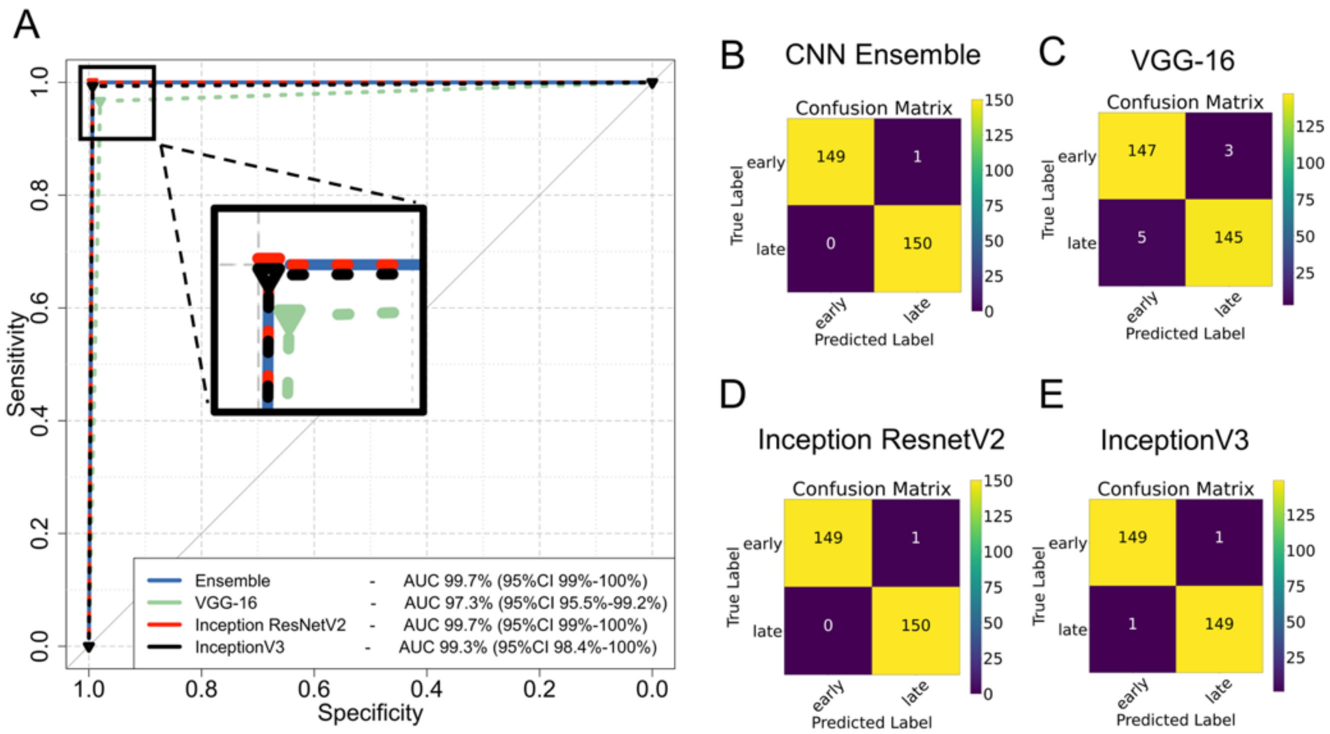


Figure 24 Performance of individual CNNs and their ensemble on the UHH cohort.

Model performance is demonstrated on Receiver Operating Characteristics (ROC) curves of the ensemble and the CNNs (A) and confusion matrices for the CNN ensemble (B) and individual CNNs (C-E). Insert on (A) magnifies the upper left corner of the ROC graph. AUC: Area under the curve

The model ensemble achieved an AUC similar to Inception ResnetV2 with only one early case misclassified as late (**Figure 24** and **Table 8**).

Table 8 Performance metrics of individual models and the neural network ensemble

	Internal Testing (UHH cohort)					External Validation (TUM cohort)			
	Group	AUC	Precision	Recall	f1-score	AUC	Precision	Recall	f1-score
Model ensemble	Early AVN	**99.7% (99-100%)**	1	0.99	1	**85.5% (72.2-93.9%)**	0.86	0.82	0.84
	Late AVN		0.99	1	1		0.85	0.88	0.87
VGG-16	Early AVN	**97.3% (95.5-99.2%)**	0.97	0.98	0.97	**78.9% (51.6-79.6%)**	0.58	1	0.73
	Late AVN		0.98	0.97	0.97		1	0.38	0.56
InceptionV3	Early AVN	**99.3% (98.4-100%)**	0.99	0.99	0.99	**74.8% (58.1-84.7%)**	0.8	0.55	0.65
	Late AVN		0.99	0.99	0.99		0.7	0.88	0.78
Inception ResNetV2	Early AVN	**99.7% (99-100%)**	1	0.99	1	**76.59% (58.1-84.7%)**	0.85	0.5	0.63
	Late AVN		0.99	1	1		0.69	0.92	0.79

AUC: Area Under the Curve; AVN: Avascular Necrosis of the hip; AUC is presented as percentage with range of 95% confidence interval

Performance of CNNs dropped when benchmarked with the TUM external validation cohort. VGG-16 achieved the highest individual AUC of 78.9% (95%CI 51.6 – 79.6%) followed by InceptionV3 and Inception ResnetV2 with AUCs of 74.8% (95%CI 58.1-84.7%) and 76.59% (95%CI 58.1-84.7%) respectively. Despite the performance drop, VGG-16 exhibited excellent precision for the diagnosis of late AVN and recall for the diagnosis of early AVN without any early cases misclassified as late. The best performance was achieved by the model ensemble which achieved an excellent AUC of 85.5% (95%CI 72.2-93.9%) with only 3 late cases misclassified as early and 4 early cases misclassified as late. Performance of the CNN ensemble was significantly higher than all individual CNNs (P-value 0.014, 0.01 and 0.028 for the comparison of the ensemble to VGG-16, Inception ResnetV2 and InceptionV3 respectively) (**Figure 25** and **Table 8**).

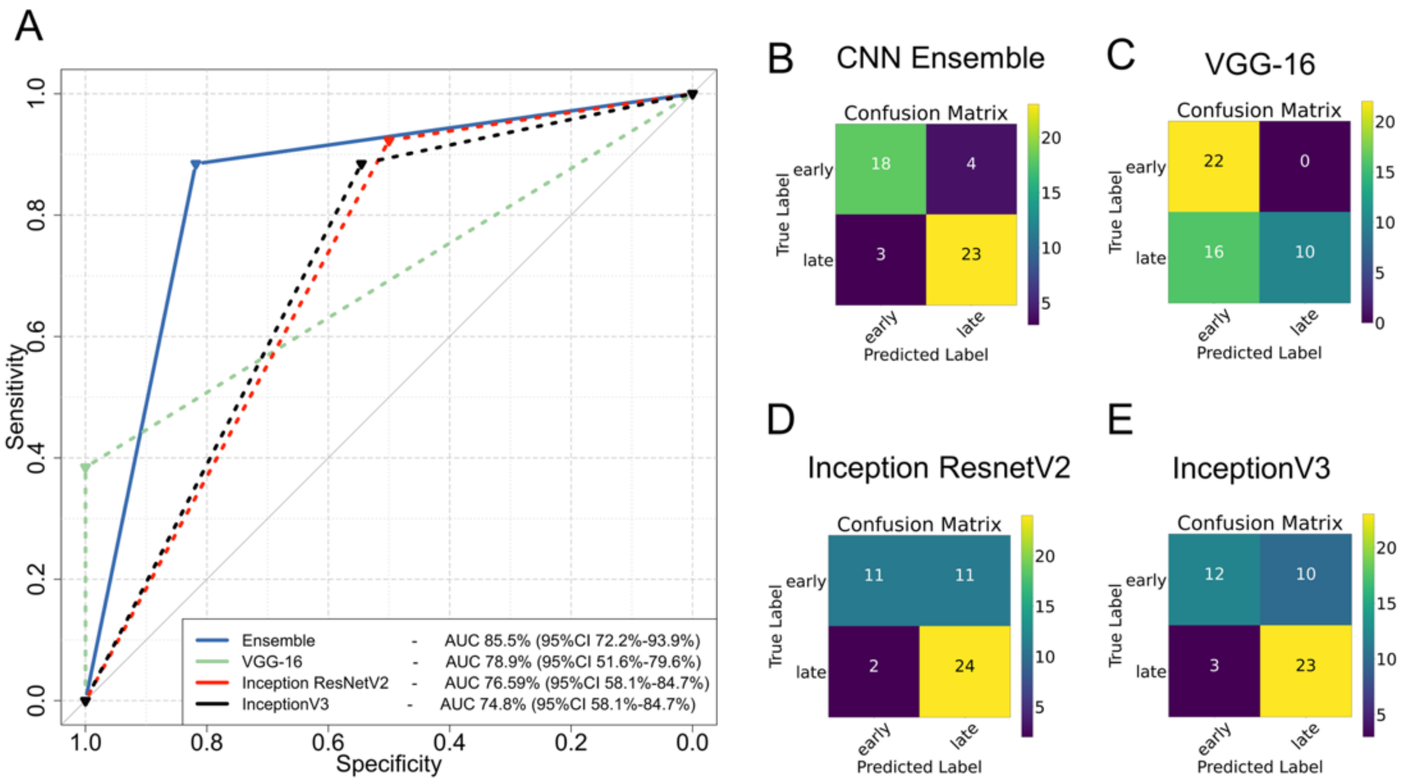


Figure 25 Performance of individual CNNs and their ensemble on the external validation TUM cohort. Model performance is demonstrated on Receiver Operating Characteristics (ROC) curves of the ensemble and the CNNs (A) and confusion matrices for the CNN ensemble (B) and individual CNNs (C-E). AUC: Area under the curve

8.3.2 Comparison between the CNN ensemble and human readers

Performance of the CNN ensemble was compared to the performance of expert readers on the TUM external validation cohort. The first MSK radiologist achieved an AUC of 75.7% (95%CI 62.7-87.9%), whereas the second achieved an AUC of 73.08% (95%CI 60.4 -86.4%). No significant difference was found between the performance of each MSK radiologist and the CNN ensemble (P-value 0.22 and 0.092 for the comparison of the CNN ensemble to the first and second MSK radiologist respectively). Both MSK radiologists achieved excellent recall values for the detection of late AVN (96.1% and 92.3% respectively) with only 1 and 2 late cases misclassified as early for the first and second MSK radiologist respectively (**Figure 26**).

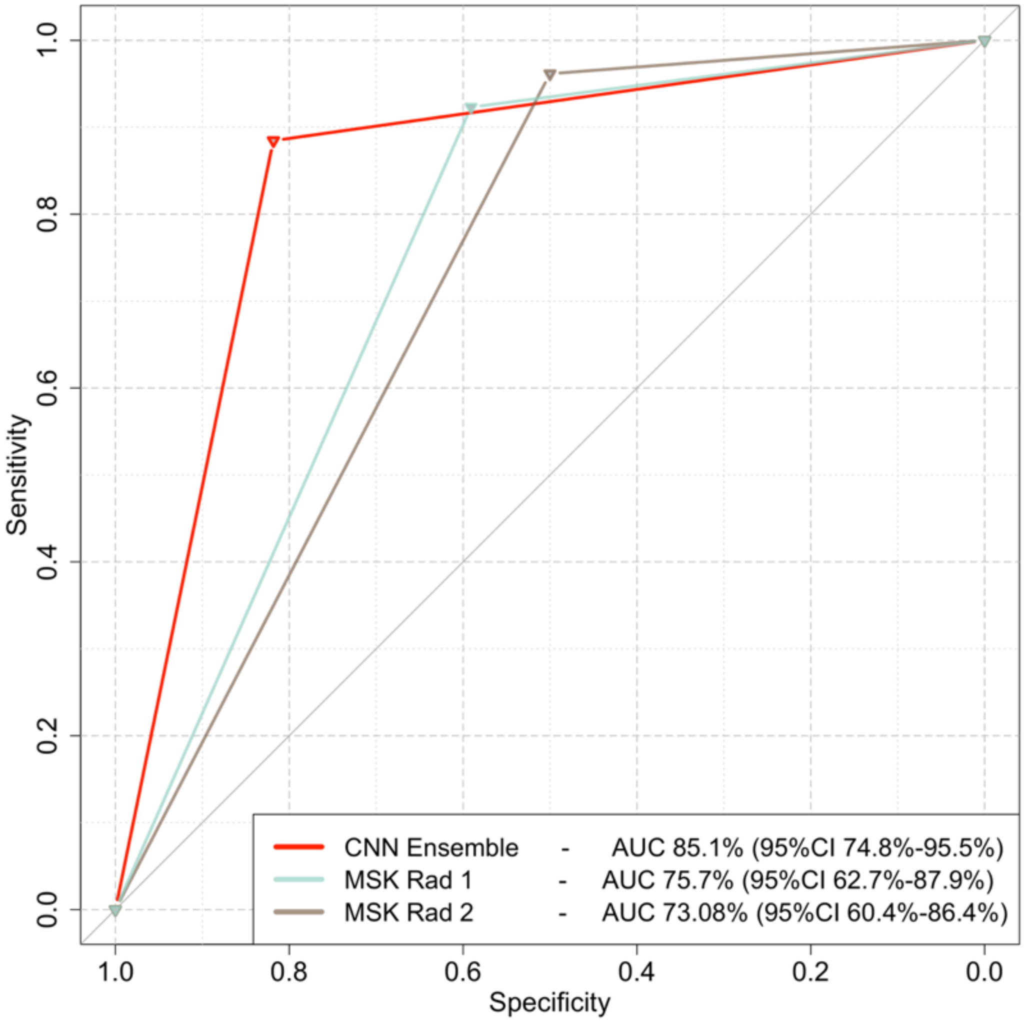


Figure 26 Receiver Operating Characteristics (ROC) curve comparing the performance of the CNN ensemble and the two expert MSK radiologists on the external validation cohort.
AUC: Area under the curve; MSK Rad: musculoskeletal radiologist

8.4 Discussion

Herein, a CNN ensemble was presented that achieved excellent performance in differentiating between early (ARCO 1-2) and late (ARCO 3-4) AVN of the hip. Three individual CNN architectures were trained and a consensus ensemble decision was derived. The excellent performance of the CNN ensemble was confirmed by external validation and was found equal to the performance of experienced MSK radiologists.

Development of a deep learning methodology to differentiate early from late AVN can be of great value in everyday clinical practice. The difficulty of distinguishing between ARCO 2 and 3A has been highlighted in several publications [39, 123, 124] and several indirect findings have been proposed as indicators of late AVN including the presence of bone marrow edema [39], joint effusion, cystic changes, bone resorption [123], and a combination of T2 signal heterogeneity, articular surface irregularity and a necrotic-viable interface with a width >3 mm [124]. Nonetheless, this remains a challenging task especially for non-experienced radiologists or in cases where high resolution sequences are not available that have been shown to be suitable for the accurate evaluation of all associated findings of AVN [129]. Given the fact that AVN can be asymptomatic in early stages [8] it can be randomly identified in pelvic MRI examinations that are not tailored to the evaluation of the proximal femur. Our CNN ensemble achieved excellent performance in distinguishing between early and late AVN only with the use of coronal STIR images. This presents a great advantage in the hands of inexperienced readers and in cases where high resolution images through the femoral head are not available to allow comprehensive ARCO staging.

Interestingly enough, all individual CNNs presented an important drop in performance when validated on the external TUM cohort. Such a performance drop has been found in the majority of externally validated deep learning studies [130, 131] . External validation is of utmost importance in establishing the "real-world" performance of deep learning algorithms but, alas, it can be found in only 6% of AI manuscripts [132]. Despite the fact that the exact reasons responsible for performance drop during external validation are still largely unknown [130], the size of the training dataset or the number of participating institutions in the training dataset have been shown to have no effect on external performance [130]. Nonetheless, being able to achieve an ensemble AUC >85% in a dataset acquired in another country provides strong evidence about the generalizability of our method.

MSK radiologists achieved AUCs in the range 70-75% which reflects the difficulty in staging the disease, especially in the absence of high-resolution images focused on the femoral head which would allow visualization of subchondral fractures equally or better than CT [13]. MSK radiologists were presented with the same coronal STIR images as the ones used for the external validation of our deep learning method. Both MSK experts achieved a high recall (sensitivity) in detecting late AVN whereas the CNN ensemble achieved high precision and recall for both late and early disease. Achieving a similar performance to MSK experts highlights the clinical value of the proposed algorithm especially in the setting of general radiology practices where highly experienced MSK radiologists are not available and protocols are not focused on the evaluation of the hip.

Our work has certain strengths and limitations. The use of a multi-institutional training dataset, the validation on an external dataset and the comparable performance to expert readers are important advantages of the proposed deep learning methodology. Limitations of the proposed work include the retrospective nature of the study and the limited training dataset. However, we have used transfer learning and data augmentation, which represent strategies suitable for deep learning with small datasets [82], alleviating this limitation as shown by the excellent performance in the internal and external cohorts. Training of the algorithms solely on coronal STIR images could be also considered as a limitation of our study. However, coronal fluid-sensitive sequences are part of most pelvic MRI protocols even when they are not focused on the hips. Such sequences can depict all the features required for ARCO staging including subchondral fractures, bone marrow edema, joint effusion, synovitis, and loss of head sphericity [14]. Therefore, being able to stage the disease based on a sequence present in most settings (even when AVN is an incidental finding), increases the clinical value of our method.

In conclusion, a CNN ensemble has been trained and validated that accurately distinguishes between early and late stages of AVN. The ensemble performs well in external data from another country and has comparable performance to expert MSK radiologists. This deep learning methodology has the potential to assist the accurate staging of AVN without the need for expertise in MSK radiology ultimately leading to the correct treatment strategy.

Chapter 9 - Conclusions & Future Perspectives

In conclusion, this PhD presented the use of radiomics and deep learning methodologies to reach important diagnostic decisions where the presence of BME could complicate the diagnosis. The algorithms presented herein performed equally or outperformed experts in all the assigned tasks, setting the basis for the introduction of such algorithms to the clinical practice. Importantly, multi-vendor datasets were used to expose the algorithms to diverse data increasing the generalizability of the methods and the AVN staging algorithm was validated on a dataset from another country demonstrating its applicability to diverse data.

The research presented in this thesis creates opportunites for further research. First of all, algorithms need to be validated on larger and more diverse datasets to ensure their generalizability. In addition, all the algorithms developed can be formulated in python packages or independent executable files with a user interface to allow immediate use in clinical practice. This will require significant computer science expertise but will potentially allow the distribution of the algorithms to collaborators and readers of our work all around the world.

Significant questions are also raised based on our results. First of all, conversion of algorithms to receive 3D inputs could significantly benefit cases with AVN in cases where the necrotic lesion is better visualized in more than one slice. However, transfer learning with 3D models is still challenging since the ImageNet dataset contains 2D images. However, the use of transfer learning with RadImageNet [83] data (once the dataset will become fully available) could significantly increase the performance and generalizability of our algorithms.

With regards to the radiomics data extracted from TOH and AVN images, these can be used in data integration projects. In such projects they can be combined with other types of omics data to derive conclusions about the underlying pathophysiological mechanisms of each of the two conditions.

Finally, the proposed algorithms could be incorporated in a federated learning framework to allow the testing from other clinical sites and the further optimization of the algorithms. Ultimately, incorporating the_algorithms into PACS systems will allow its seamless use in everyday radiological reporting.

Bibliography

1. Klontzas ME, Zibis AH, Vassalou EE, Karantanas AH (2017) MRI of the hip: Current concepts on bone marrow oedema. Hip Int 27:329–335
2. Korompilias A V, Karantanas AH, Lykissas MG, Beris AE (2009) Bone marrow edema syndrome. Skeletal Radiol 38:425–36
3. Karantanas AH (2007) Acute bone marrow edema of the hip: role of MR imaging. Eur Radiol 17:2225–2236
4. Klontzas ME, Zibis AH, Karantanas AH (2015) Osteoid osteoma of the femoral neck: use of the "half-moon" sign in MRI diagnosis. AJR Am J Roentgenol 205:353–357
5. Klontzas ME, Zibis AH, Karantanas AH (2016) Reply to "the half-moon sign of the femoral neck is nonspecific for the diagnosis of osteoid osteoma." American Journal of Roentgenology. https://doi.org/10.2214/AJR.15.15724
6. Klontzas ME, Vassalou EE, Zibis AH, Bintoudi AS, Karantanas AH (2014) MR imaging of transient osteoporosis of the hip: an update on 155 hip joints. Eur J Radiol
7. Miyanishi K, Yamamoto T, Nakashima Y, Shuto T, Jingushi S, Noguchi Y, Iwamoto Y (2001) Subchondral changes in transient osteoporosis of the hip. Skeletal Radiol 30:255–61
8. Huang G-S, Chan WP, Chang Y-C, Chang C-Y, Chen C-Y, Yu JS (2003) MR imaging of bone marrow edema and joint effusion in patients with osteonecrosis of the femoral head: relationship to pain. AJR Am J Roentgenol 181:545–9
9. Klontzas ME, Kramer J, Karantanas AH (2016) The many faces of hip bone marrow edema. In J Kramer & AH Karantanas (Eds.), MRI of the hip (ESSR Sport), Breitenseher Publisher
10. Yamamoto T, Iwamoto Y, Schneider R, Bullough PG (2008) Histopathological prevalence of subchondral insufficiency fracture of the femoral head. Ann Rheum Dis 67:150 LP – 153
11. Karantanas AH, Drakonaki E, Karachalios T, Korompilias A V, Malizos K (2008) Acute non-traumatic marrow edema syndrome in the knee: MRI findings at presentation, correlation with spinal DEXA and outcome. Eur J Radiol 67:22–33

12. Gorbachova T, Melenevsky Y, Cohen M, Cerniglia BW (2018) Osteochondral lesions of the knee: differentiating the most common entities at MRI. RadioGraphics 38:1478–1495

13. Karantanas AH (2013) Accuracy and limitations of diagnostic methods for avascular necrosis of the hip. Expert Opin Med Diagn 7:179–187

14. Karantanas AH, Drakonaki EE (2011) The role of MR imaging in avascular necrosis of the femoral head. Semin Musculoskelet Radiol 15:281–300

15. Ruckli AC, Nanavati AK, Meier MK, et al (2023) A deep learning method for quantification of femoral head necrosis based on routine hip MRI for improved surgical decision making. J Person Med. https://doi.org/10.3390/jpm13010153

16. Klontzas ME, Papadakis GZ, Marias K, Karantanas AH (2020) Musculoskeletal trauma imaging in the era of novel molecular methods and artificial intelligence. Injury 51:2748–2756

17. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: images are more than pictures, they are data. Radiology 278:563–577

18. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B (2020) Radiomics in medical imaging—"how-to" guide and critical reflection. Insights Imaging 11:91

19. Kahn CE (2019) Artificial intelligence, real radiology. Radiol Artif Intell 1:e184001

20. Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö (2019) Radiomics with artificial intelligence: A practical guide for beginners. Diagn Interv Radiol 25:485–495

21. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL (2018) Artificial intelligence in radiology. Nat Rev Cancer 18:500–510

22. Lindsey R, Daluiski A, Chopra S, et al (2018) Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci USA 115:11591–11596

23. Andrews CL (2000) From the RSNA Refresher Courses. Radiological Society of North America. Evaluation of the marrow space in the adult hip. Radiographics 20:S27-42

24. Vande Berg BC, Malghem J, Lecouvet FE, Maldague B (1998) Magnetic resonance imaging of normal bone marrow. Eur Radiol 8:1327–1334

25. Hodnett PA, Shelly MJ, MacMahon PJ, Kavanagh EC, Eustace SJ (2009) MR imaging of overuse injuries of the hip. Magn Reson Imaging Clin N Am 17:667–79

26. Klontzas ME, Zibis AH, Karantanas AH (2015) Osteoid osteoma of the femoral neck: Use of the half-moon sign in MRI diagnosis. AJR Am J Roentgenol 205:353–357

27. Curtiss Jr PH, Kincaid WE (1959) Transitory demineralization of the hip in pregnancy: a report of three cases. J Bone J Surg Am 41:1327–1333

28. Rosen RA (1970) Transitory demineralization of the femoral head. Radiology 94:509–12

29. Karantanas AH, Nikolakopoulos I, Korompilias A V, Apostolaki E, Skoulikaris N, Eracleous E (2008) Regional migratory osteoporosis in the knee: MRI findings in 22 patients and review of the literature. Eur J Radiol 67:34–41

30. Cahir JG, Toms AP (2008) Regional migratory osteoporosis. Eur J Radiol 67:2–10

31. Malizos KN, Zibis AH, Dailiana Z, Hantes M, Karachalios T, Karantanas AH (2004) MR imaging findings in transient osteoporosis of the hip. Eur J Radiol 50:238–44

32. Zhao D, Zhang F, Wang B, et al (2020) Guidelines for clinical diagnosis and treatment of osteonecrosis of the femoral head in adults (2019 version). J Orthop Transl 21:100–110

33. Mont MA, Cherian JJ, Sierra RJ, Jones LC, Lieberman JR (2015) Nontraumatic osteonecrosis of the femoral head: Where do we stand today? A ten-year update. J Bone Joint Surg Am 97:1604–27

34. Petek D, Hannouche D, Suva D (2019) Osteonecrosis of the femoral head: pathophysiology and current concepts of treatment. EFORT Open Rev 4:85–97

35. Beckmann J, Schmidt T, Schaumburger J, Rath B, Lüring C, Tingart M, Grifka J (2013) Infusion, core decompression, or infusion following core decompression in the treatment of bone edema syndrome and early avascular osteonecrosis of the femoral head. Rheumatol Int 33:1561–1565

36. Yoon B, Mont MA, Koo K, et al (2020) The 2019 revised version of Association Research Circulation Osseous staging system of osteonecrosis of the femoral head. J Arthroplasty 35:933–940

37. Jawad MU, Haleem AA, Scully SP (2012) In brief: Ficat classification: avascular necrosis of the femoral head. Clin Orthop Relat Res 470:2636–9

38. Steinberg ME, Hayken GD, Steinberg DR (1995) A quantitative system for staging avascular necrosis. J Bone Joint Surg Br 77:34–41

39. Meier R, Kraus TM, Schaeffeler C, et al (2014) Bone marrow oedema on MR imaging indicates ARCO stage 3 disease in patients with AVN of the femoral head. Eur Radiol 24:2271–2278

40. Balakrishnan A, Schemitsch EH, Pearce D, McKee MD (2003) Distinguishing transient osteoporosis of the hip from avascular necrosis. Can J Surg 46:187–92

41. Harvey EJ (2003) Osteonecrosis and transient osteoporosis of the hip : diagnostic and treatment dilemmas. J Can Chir 46:168–169

42. Yamamoto T, Kubo T, Hirasawa Y, Noguchi Y, Iwamoto Y, Sueishi K (1999) A clinicopathologic study of transient osteoporosis of the hip. Skeletal Radiol 28:621–7

43. Geith T, Stellwag AC, Müller PE, Reiser M, Baur-Melnyk A (2020) Is bone marrow edema syndrome a precursor of hip or knee osteonecrosis? Results of 49 patients and review of the literature. Diagn Interv Radiol 26:355–362

44. Geith T, Niethammer T, Milz S, Dietrich O, Reiser M, Baur-Melnyk A (2017) Transient bone marrow edema syndrome versus osteonecrosis: Perfusion patterns at dynamic contrast-enhanced MR imaging with high temporal resolution can allow differentiation. Radiology 283:478–485

45. Al-Dourobi K, Corbaz J, Bauer S, Leumessi EN (2021) Lower lumbar back pain occurring with transient hip osteoporosis: Complication of prolonged suffering and neck of femur fracture in a 24-year-old pregnant patient. BMJ Case Rep 14:e238477

46. Hong CS, Bergen MA, Watters TS (2019) Transient osteoporosis of the hip after bariatric surgery. Arthroplast Today 5:32–37

47. Yamada R, Okada D, Wang J, Basak T, Koyama S (2021) Interpretation of omics data analyses. J Human Gen 66:93–102

48. Farrell A, McLoughlin N, Milne JJ, Marison IW, Bones J (2014) Application of multi-omics techniques for bioprocess design and optimisation in chinese hamster ovary cells. J Proteome Res 140526175220007

49. Misra BB, Langefeld CD, Olivier M, Cox LA (2018) Integrated omics: tools, advances, and future approaches. J Mol Endocrinol. https://doi.org/10.1530/JME-18-0055

50. Lambin P, Leijenaar RTH, Deist TM, et al (2017) Radiomics: The bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749–762

51. Mannello F, Ligi D, Magnani M (2012) Deciphering the single-cell omic: innovative application for translational medicine. Expert Rev Proteomics 9:635–648

52. Beyer BA, Fang M, Sadrian B, Montenegro-Burke JR, Plaisted WC, Kok BPC, Saez E, Kondo T, Siuzdak G, Lairson LL (2018) Metabolomics-based discovery of a metabolite that enhances oligodendrocyte maturation. Nat Chem Biol 14:22–28

53. Fiehn O (2002) Metabolomics--the link between genotypes and phenotypes. Plant Mol Biol 48:155–71

54. Srivastava A, Evans KJ, Sexton AE, Schofield L, Creek DJ (2017) Metabolomics-based elucidation of active metabolic pathways in erythrocytes and hsc-derived reticulocytes. J Proteome Res 16:1492–1505

55. Chen H, Zhang X, Wang X, et al (2021) MRI-based radiomics signature for pretreatment prediction of pathological response to neoadjuvant chemotherapy in osteosarcoma: a multicenter study. Eur Radiol 31:7913–7924

56. Gitto S, Cuocolo R, van Langevelde K, van de Sande MAJ, Parafioriti A, Luzzati A, Imbriaco M, Sconfienza LM, Bloem JL (2022) MRI radiomics-based machine learning classification of atypical cartilaginous tumour and grade II chondrosarcoma of long bones. EBioMedicine 75:1–11

57. Gitto S, Cuocolo R, Albano D, et al (2020) MRI radiomics-based machine-learning classification of bone chondrosarcoma. Eur J Radiol 128:109043

58. Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin JC, Pieper S, Aerts HJWL (2017) Computational

radiomics system to decode the radiographic phenotype. Cancer Res 77:e104–e107

59. Leporq B, Bouhamama A, Pilleul F, Lame F, Bihane C, Sdika M, Blay JY, Beuf O (2020) MRI-based radiomics to predict lipomatous soft tissue tumors malignancy: a pilot study. Cancer Imaging 20:1–8

60. Yin P, Mao N, Wang S, Sun C, Hong N (2019) Clinical-radiomics nomograms for pre-operative differentiation of sacral chordoma and sacral giant cell tumor based on 3D computed tomography and multiparametric magnetic resonance imaging. Br J Radiol 92:20190155

61. Papanikolaou N, Matos C, Koh DM (2020) How to develop a meaningful radiomic signature for clinical use in oncologic patients. Cancer Imaging 20:33

62. Grossmann P, Stringfield O, El-Hachem N, et al (2017) Defining the biological basis of radiomic phenotypes in lung cancer. Elife. https://doi.org/10.7554/eLife.23421

63. Mallat S, Zhong S (1992) Characterization of signals from multiscale edges. IEEE Trans Pattern Anal Mach Intell 14:710–732

64. Zhang Z, Ma S, Liu H, Gong Y (2009) An edge detection approach based on directional wavelet transform. Comput Math with Appl 57:1265–1271

65. DeSouza NM, Achten E, Alberich-Bayarri A, et al (2019) Validated imaging biomarkers as decision-making tools in clinical trials and routine practice: current status and recommendations from the EIBALL* subcommittee of the European Society of Radiology (ESR). Insights Imaging. https://doi.org/10.1186/s13244-019-0764-0

66. Zwanenburg A, Vallières M, Abdalah MA, et al (2020) The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 295:328–338

67. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and application to Boosting. J Comput System Sci 55:119–139

68. Wu X, Kumar V, Ross QJ, et al (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14:1–37

69. He Z, Lin D, Lau T, Wu M (2019) Gradient boosting machine: a survey point zero one technology. ArXiv 1908.06951:1–9

70. Ruder S (2016) An overview of gradient descent optimization algorithms. ArXiv 1609.04747:1–14

71. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. ArXiv 1603.02754:1–13

72. Klontzas ME, Manikis GC, Nikiforaki K, et al (2021) Radiomics and machine learning can differentiate transient osteoporosis from avascular necrosis of the hip. Diagnostics 11:1686

73. Chen P-T, Chang D, Yen H, Liu K-L, Huang S-Y, Roth H, Wu M-S, Liao W-C, Wang W (2021) Radiomic features at CT can distinguish pancreatic cancer from noncancerous pancreas. Radiol Imaging cancer 3:e210010

74. Awe AM, vanden Heuvel MM, Yuan T, Rendell VR, Shen M, Kampani A, Liang S, Morgan DD, Winslow ER, Lubner MG (2022) Machine learning principles applied to CT radiomics to predict mucinous pancreatic cysts. Abdominal Radiol 47:221–231

75. Saba L, Biswas M, Kuppili V, et al (2019) The present and future of deep learning in radiology. Eur J Radiol 114:14–24

76. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neur Comput 1:541–551

77. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313:504–7

78. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60:84–90

79. Naveen P, Diwan B (2021) Pre-trained VGG-16 with CNN architecture to classify x-rays images into normal or pneumonia. In: 2021 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE, pp 102–105

80. Kim DH, MacKinnon T (2018) Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 73:439–445

81. Inagaki N, Nakata N, Ichimori S, Udaka J, Mandai A, Saito M (2022) Detection of sacral fractures on radiographs using artificial intelligence. JBJS Open Access. https://doi.org/10.2106/JBJS.OA.22.00030

82. Candemir S, Nguyen X V, Folio LR, Prevedello LM (2021) Training strategies for radiology deep learning models in data-limited scenarios. Radiology: Artif Intell. https://doi.org/10.1148/ryai.2021210014

83. Mei X, Liu Z, Robson PM, et al (2022) RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. Radiology Artif Intell 4:e210315

84. Hong JH, Jung JY, Jo A, Nam Y, Pak S, Lee SY, Park H, Lee SE, Kim S (2021) Development and validation of a radiomics model for differentiating bone islands and osteoblastic bone metastases at abdominal CT. Radiology 299:626–632

85. Jin Z, Zhang F, Wang Y, Tian A, Zhang J, Chen M, Yu J (2022) Single-photon emission computed tomography/computed tomography image-based radiomics for discriminating vertebral bone metastases from benign bone lesions in patients with tumors. Front Med (Lausanne). https://doi.org/10.3389/fmed.2021.792581

86. Dai Y, Yin P, Mao N, Sun C, Wu J, Cheng G, Hong N (2020) Differentiation of pelvic osteosarcoma and Ewing sarcoma using radiomic analysis based on T2-weighted images and contrast-enhanced T1-weighted images. Biomed Res Int 2020:9078603

87. Tenório APM, Ferreira-Junior JR, Dalto VF, Faleiros MC, Assad RL, Louzada-Junior P, Nogueira-Barbosa MH, Rangayyan RM, de Azevedo-Marques PM (2022) Radiomic quantification for MRI assessment of sacroiliac joints of patients with spondyloarthritis. J Digit Imaging 35:29–38

88. Nagawa K, Suzuki M, Yamamoto Y, Inoue K, Kozawa E, Mimura T, Nakamura K, Nagata M, Niitsu M (2021) Texture analysis of muscle MRI: machine learning-based classifications in idiopathic inflammatory myopathies. Sci Rep 11:9821

89. Dieckmeyer M, Inhuber S, Schlaeger S, et al (2021) Texture features of proton density fat fraction maps from chemical shift encoding-based MRI predict

paraspinal muscle strength. Diagnostics. https://doi.org/10.3390/diagnostics11020239

90. Guermazi A, Tannoury C, Kompel AJ, et al (2022) Improving radiographic fracture recognition performance and efficiency using artificial intelligence. Radiology 302:627–636

91. Fritz B, Fritz J (2022) Artificial intelligence for MRI diagnosis of joints: a scoping review of the current state-of-the-art of deep learning-based approaches. Skeletal Radiol 51:315–329

92. Liu F, Guan B, Zhou Z, et al (2019) Fully automated diagnosis of anterior cruciate ligament tears on knee MR images by using deep learning. Radiology Artif Intell 1:e180091

93. Fritz B, Yi PH, Kijowski R, Fritz J (2023) Radiomics and learning for disease detection in musculoskeletal radiology: an overview of novel MRI- and CT-based approaches. Invest Radiology 58:3–13

94. Jamaludin A, Lootus M, Kadir T, Zisserman A, Urban J, Battié MC, Fairbank J, McCall I, Genodisc Consortium (2017) ISSLS PRIZE IN BIOENGINEERING SCIENCE 2017: Automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. Eur Spine J 26:1374–1383

95. Won D, Lee H-J, Lee S-J, Park SH (2020) Spinal stenosis grading in magnetic resonance imaging using deep convolutional neural networks. Spine (Phila Pa 1976) 45:804–812

96. Shin Y, Yang J, Lee YH (2021) Deep generative adversarial networks: Applications in musculoskeletal imaging. Radiology Artif Intell. https://doi.org/10.1148/ryai.2021200157

97. Chang CY, Buckless C, Yeh KJ, Torriani M (2022) Automated detection and segmentation of sclerotic spinal lesions on body CTs using a deep convolutional neural network. Skeletal Radiol 51:391–399

98. Dutt R, Mendonca D, Phen HM, Broida S, Ghassemi M, Gichoya J, Banerjee I, Yoon T, Trivedi H (2022) Automatic localization and brand detection of cervical spine hardware on radiographs using weakly supervised machine learning. Radiology Artif Intell 4:1–10

99. Patel R, Thong EHE, Batta V, Bharath AA, Francis D, Howard J (2021) Automated identification of orthopedic implants on radiographs using deep learning. Radiology Artif Intell 3:1–8

100. Islam S, Kanavati F, Arain Z, Da Costa OF, Crum W, Aboagye EO, Rockall AG (2022) Fully automated deep-learning section-based muscle segmentation from CT images for sarcopenia assessment. Clinical Radiol Epub ahead of print

101. Rouzrokh P, Wyles CC, Kurian SJ, Ramazanian T, Cai JC, Huang Q, Zhang K, Taunton MJ, Maradit Kremers H, Erickson BJ (2022) Deep learning for radiographic measurement of femoral component subsidence following total hip arthroplasty. Radiology Artif Intell 4:e210206

102. Stotter C, Klestil T, Röder C, Reuter P, Chen K, Emprechtinger R, Hummer A, Salzlechner C, DiFranco M, Nehrer S (2023) Deep learning for fully automated radiographic measurements of the pelvis and hip. Diagnostics. https://doi.org/10.3390/diagnostics13030497

103. Roth HR, Yao J, Lu L, Stieger J, Burns JE, Summers RM (2014) Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications. arXiv: 1407.5976

104. Li Y, Zhang Y, Zhang E, Chen Y, Wang Q, Liu K, Yu HJ, Yuan H, Lang N, Su M-Y (2021) Differential diagnosis of benign and malignant vertebral fracture on CT using deep learning. Eur Radiol 31:9612–9619

105. Yeh L-R, Zhang Y, Chen J-H, Liu Y-L, Wang A-C, Yang J-Y, Yeh W-C, Cheng C-S, Chen L-K, Su M-Y (2022) A deep learning-based method for the diagnosis of vertebral fractures on spine MRI: retrospective training and validation of ResNet. Eur Spine J 31:2022–2030

106. Navarro F, Dapper H, Asadpour R, et al (2021) Development and external validation of deep-learning-based tumor grading models in soft-tissue sarcoma patients using MR imaging. Cancers (Basel). https://doi.org/10.3390/cancers13122866

107. He Y, Guo J, Ding X, van Ooijen PMA, Zhang Y, Chen A, Oudkerk M, Xie X (2019) Convolutional neural network to predict the local recurrence of giant cell tumor of bone after curettage based on pre-surgery magnetic resonance images. Eur Radiol 29:5441–5451

108. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845

109. Malizos KN, Karantanas AH, Varitimidis SE, Dailiana ZH, Bargiotas K, Maris T (2007) Osteonecrosis of the femoral head: etiology, imaging and treatment. Eur J Radiol 63:16–28

110. Klontzas ME, Stathis I, Spanakis K, Zibis AH, Marias K, Karantanas AH (2022) Deep learning for the differential diagnosis between transient osteoporosis and avascular necrosis of the hip. Diagnostics. https://doi.org/10.3390/diagnostics12081870

111. Tasci E, Uluturk C, Ugur A (2021) A voting-based ensemble deep learning method focusing on image augmentation and preprocessing variations for tuberculosis detection. Neural Comput Appl 33:15541–15555

112. Tsiknakis N, Savvidaki E, Manikis GC, Gotsiou P, Remoundou I, Marias K, Alissandrakis E, Vidakis N (2022) Pollen grain classification based on ensemble transfer learning on the cretan pollen dataset. Plants 11:1–15

113. Bento M, Fantini I, Park J, Rittner L, Frayne R (2022) Deep Learning in Large and Multi-Site Structural Brain MR Imaging Datasets. Front Neuroinform 15:1–17

114. Klontzas ME, Vassalou EE, Kakkos GA, Spanakis K, Zibis A, Marias K, Karantanas AH (2022) Differentiation between subchondral insufficiency fractures and advanced osteoarthritis of the knee using transfer learning and an ensemble of convolutional neural networks. Injury 53:2035–2040

115. Korompilias A V, Karantanas AH, Lykissas MG, Beris AE (2008) Transient osteoporosis. J Am Acad Orthop Surg 16:480–9

116. Roemer FW, Frobell R, Hunter DJ, Crema MD, Fischer W, Bohndorf K, Guermazi A (2009) MRI-detected subchondral bone marrow signal alterations of the knee joint: terminology, imaging appearance, relevance and radiological differential diagnosis. Osteoarthr Cartil 17:1115–1131

117. Sayyid S, Younan Y, Sharma G, Singer A, Morrison W, Zoga A, Gonzalez FM (2019) Subchondral insufficiency fracture of the knee: grading, risk factors, and outcome. Skeletal Radiol 48:1961–1974

118. Allam E, Boychev G, Aiyedipe S, Morrison W, Roedl JB, Singer AD, Gonzalez FM (2021) Subchondral insufficiency fracture of the knee: unicompartmental correlation to meniscal pathology and degree of chondrosis by MRI. Skeletal Radiol 50:2185–2194

119. Lim YZ, Wang Y, Wluka AE, Davies-Tuck ML, Teichtahl A, Urquhart DM, Cicuttini FM (2013) Are biomechanical factors, meniscal pathology, and physical activity risk factors for bone marrow lesions at the knee? A systematic review. Semin Arthritis Rheum 43:187–194

120. Felson DT, Mclaughlin S, Goggins J, Lavalley MP, Gale ME, Totterman S, Li W, Hill C, Gale D (2003) Bone marrow edema and its relation to progression of knee osteoarthritis. Ann Intern Med 139:330–337

121. Hunter DJ, Arden N, Conaghan PG, et al (2011) Definition of osteoarthritis on MRI: Results of a Delphi exercise. Osteoarthr Cartil 19:963–969

122. Bergman AG, Willen HK, Lindstrand AL, Pettersson HTA (1978) Osteoarthritis of the knee: correlation of subchondral MR signal abnormalities with histopathologic and radiographic features. Skeletal Radiol 193:191–193

123. Kim J, Lee SK, Kim J-Y, Kim J-H (2023) CT and MRI findings beyond the subchondral bone in osteonecrosis of the femoral head to distinguish between ARCO stages 2 and 3A. Eur Radiol Epub ahead of print

124. Shi S, Luo P, Sun L, Zhao Y, Yang X, Xie L, Yu T, Wang Z (2022) Analysis of MR signs to distinguish between ARCO stages 2 and 3A in osteonecrosis of the femoral head. J Magn Reson Imaging 55:610–617

125. Li Y, Li Y, Tian H (2021) Deep learning-based end-to-end diagnosis system for avascular necrosis of femoral head. IEEE J Biomed Health Inform 25:2093–2102

126. Shen X, Luo J, Tang X, Chen B, Qin Y, Zhou Y, Xiao J (2022) Deep learning approach for diagnosing early osteonecrosis of the femoral head based on magnetic resonance imaging. J Arthroplasty. https://doi.org/10.1016/j.arth.2022.10.003

127. Hernigou P (2023) Revisiting prediction of collapse in hip osteonecrosis with artificial intelligence and machine learning: a new approach for quantifying and ranking the contribution and association of factors for collapse. Int Orthop 47:677–689

128. Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T (2022) Transfer learning for medical image classification: a literature review. BMC Med Imaging 22:1–13

129. Shuman WP, Castagno AA, Baron RL, Richardson ML (1988) MR imaging of avascular necrosis of the femoral head: value of small-field-of-view sagittal surface-coil images. AJR Am J Roentgenol 150:1073–8

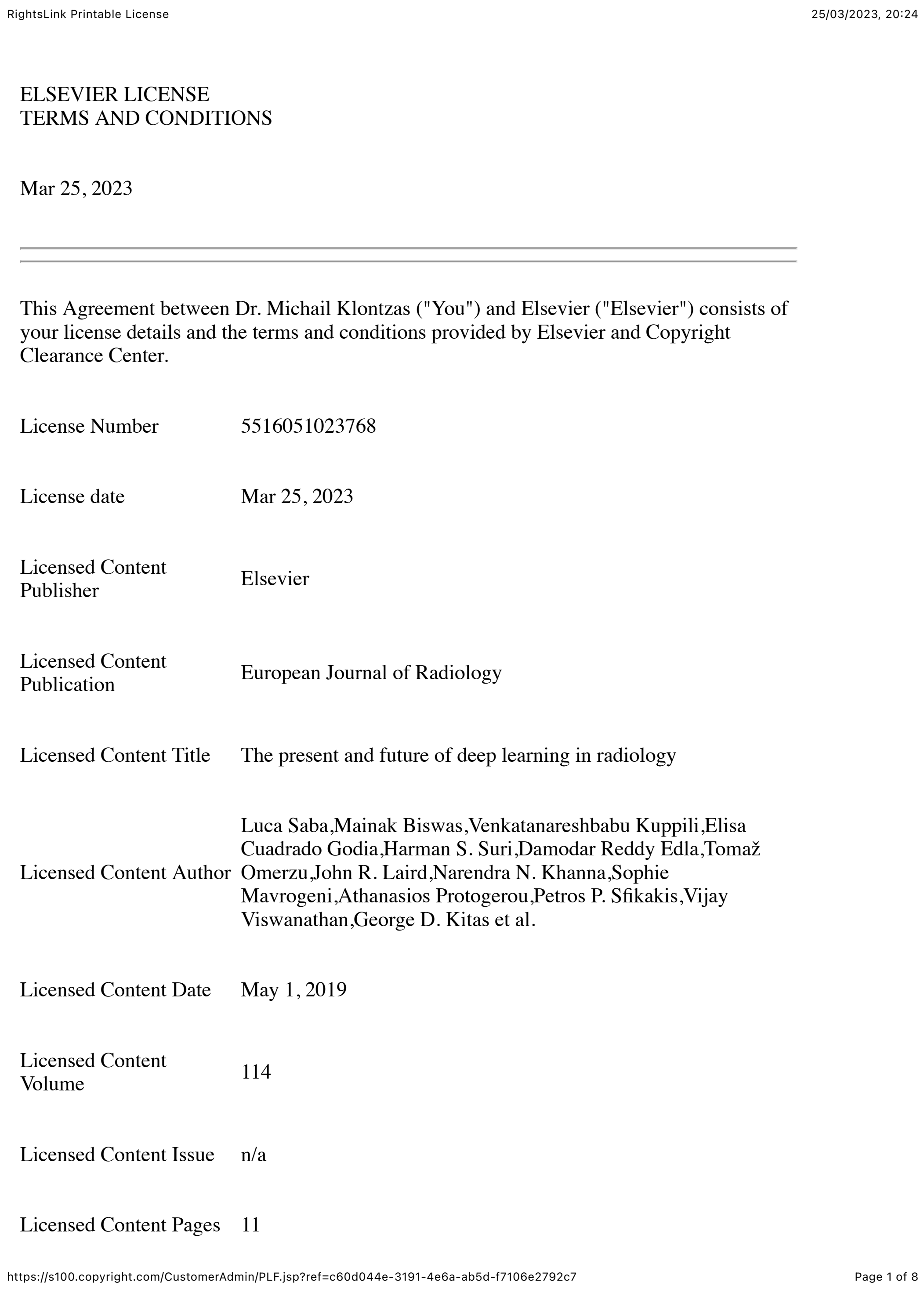
130. Yu AC, Mohajer B, Eng J (2022) External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell 4:e210064

131. Hsu W, Hippe DS, Nakhaei N, et al (2022) External validation of an ensemble model for automated mammography interpretation by artificial intelligence. JAMA Netw Open 5:e2242343

132. Kim DW, Jang HY, Kim KW, Shin Y, Park SH (2019) Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 20:405–410

Appendix

Permissions to reuse images - text

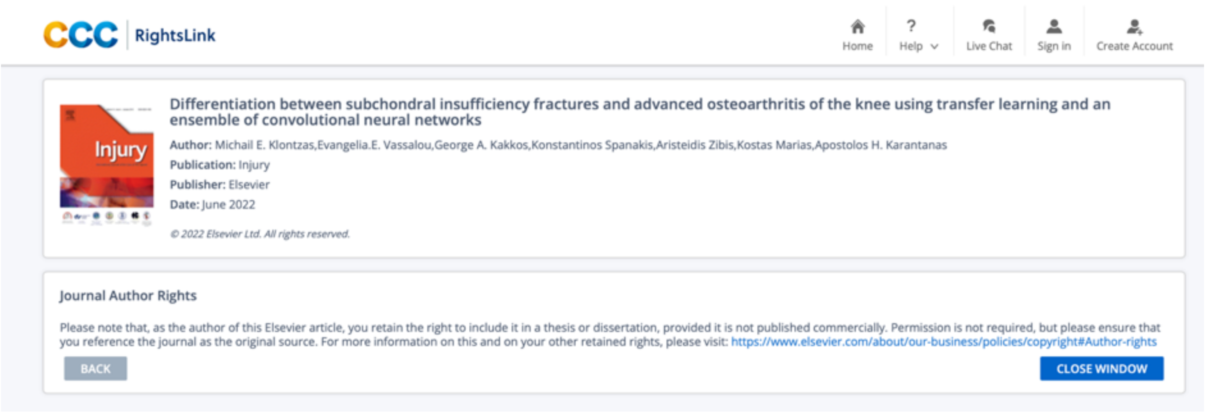


ELSEVIER LICENSE
TERMS AND CONDITIONS

Mar 25, 2023

This Agreement between Dr. Michail Klontzas ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	5516051023768
License date	Mar 25, 2023
Licensed Content Publisher	Elsevier
Licensed Content Publication	European Journal of Radiology
Licensed Content Title	The present and future of deep learning in radiology
Licensed Content Author	Luca Saba,Mainak Biswas,Venkatanareshbabu Kuppili,Elisa Cuadrado Godia,Harman S. Suri,Damodar Reddy Edla,Tomaž Omerzu,John R. Laird,Narendra N. Khanna,Sophie Mavrogeni,Athanasios Protogerou,Petros P. Sfikakis,Vijay Viswanathan,George D. Kitas et al.
Licensed Content Date	May 1, 2019
Licensed Content Volume	114
Licensed Content Issue	n/a
Licensed Content Pages	11



Differentiation between subchondral insufficiency fractures and advanced osteoarthritis of the knee using transfer learning and an ensemble of convolutional neural networks

Author: Michail E. Klontzas,Evangelia.E. Vassalou,George A. Kakkos,Konstantinos Spanakis,Aristeidis Zibis,Kostas Marias,Apostolos H. Karantanas

Publication: Injury

Publisher: Elsevier

Date: June 2022

© 2022 Elsevier Ltd. All rights reserved.

Journal Author Rights

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights